LASER BEAM SHAPING
THEORY AND TECHNIQUES

EDITED BY

FRED M. DICKEY
SCOTT C. HOLSWADE

Sandia National Laboratories
Albuquerque, New Mexico

MARCEL DEKKER, INC. NEW YORK • BASEL

ISBN: 0-8247-0398-7

This book is printed on acid-free paper.

Headquarters
Marcel Dekker, Inc.
270 Madison Avenue, New York, NY 10016
tel: 212-696-9000; fax: 212-685-4540

Eastern Hemisphere Distribution
Marcel Dekker AG
Hutgasse 4, Postfach 812, CH-4001 Basel, Switzerland
tel: 41-61-261-8482; fas: 41-61-261-8896

World Wide Web
http://www.dekker.com

The publisher offers discounts on this book when ordered in bulk quantities. For more information, write to special Sales/Professional Marketing at the headquarters address above.

Current printing (last digit):
10 9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

To our parents and children

From the Series Editor

Application of scientific advances requires the development of new technology and hence new devices, but that is not the end of the story. Devices are used in systems, and these systems are used as commercial products as instrumentation for testing, evaluation and research. These tools lead to the next generation of scientific knowledge. An important consideration in optical systems is the character of the optical beam that passes through that system to deliver energy, power, or information in an appropriate and efficient manner.

In the early days of optical systems, the main concerns were beam uniformity and brightness for imaging and interferometry. Energy delivered on target for focused systems was also a significant consideration. In image-forming systems, an additional challenge was resolution and image quality. Apodization became an important tool, e.g., to vary the pupil function of the system to achieve a particular intensity distribution in the image plane, and thus control the intensity impulse response of the system (and hence control the optical transfer function). The next parameters to be controlled were the spatial and temporal coherence of the beam as produced from incoherent primary sources.

The availability of a variety of laser sources certainly solved the coherence control issue but led to a new wave of apodization studies to control both the amplitude impulse response of coherent imaging systems and the fluctuations in high-energy beam propagation. In addition, the availability of an ever-increasing array of laser sources led to a large number of non-imaging application of optical systems such as laser material processing, cutting and welding; laser surgery; and basic optical physics research. These system applications as well as modern imaging applications require the ability to precisely control the amplitude and phase properties of laser beams.

The discussion of beam shaping in this volume provides the theoretical basis and techniques for the conditioning of laser beams required for a large

variety of applications. These are fundamental tools for optical system design and implementation.

Brian J. Thompson

Preface

Since the advent of the laser, many applications have required shaping the laser beam irradiance profile. Several of the primary applications involve material processing, including welding, cutting, and drilling, and medical procedures, such as corneal surgery and cosmetic skin treatments. Other applications include laser/material interaction studies, lithography, semiconductor manufacture, graphic arts, optical data processing, and military uses. Although the laser beam shaping literature goes back more than twenty-five years, there is not a dominant or summarizing set of papers. Furthermore, the information is widely scattered. In our experience, the knowledge of beam shaping techniques, possibilities, and limitations is thus not well disseminated throughout the engineering community. Many of the primary users of beam shaping technology are not necessarily optical experts. A book on beam shaping would thus benefit a wide audience.

The purpose of this book is to present in one volume the salient aspects of laser beam shaping. The technology involves diffraction theory, geometrical optics, optical design, and beam profile measurement methods. It is fundamentally constrained by the principles of electromagnetic theory. After the Introduction, Chapter 2 presents the underlying electromagnetic theory and mathematical techniques applicable to beam shaping. The diffraction theory approach and the geometrical optics approach to beam shaping are two facets of the same problem. Both approaches are fundamentally coupled through electromagnetic theory. Chapter 3 presents a diffraction approach to single-mode Gaussian beam shaping and includes experimental results. Geometrical optics methods, theory, and application are treated in Chapter 4. Optimization-based techniques are treated in Chapter 5. A unique technique for beam shaping using diffractive diffusers

is introduced in Chapter 6. Multiaperture beam integration systems, including experiment and design, are presented in Chapter 7. For completeness, Chapter 8 discusses the application of geometrical optics methods to classical (non-laser) shaping problems. Generally, the non-laser problem differs from the laser ones in that the source is usually highly incoherent and frequently nondirectional. Chapter 9 discusses current beam profile measurement technology, an important and necessary part of the shaping problem. Although the number of laser applications that require beam shaping is growing beyond the bounds of one book, these chapters give readers a working understanding of beam shaping techniques. They also provide insight into the potential application of laser beam profile shaping in laser system design.

The book is limited, with the exception of Chapter 8, to the treatment, theory, and practice of shaping laser beam irradiance profiles for a variety of laser processing applications. There are many application-specific techniques in the field of laser processing, and it is not practicable to try to cover all variations in one volume. However, the theory, design, measurement technology, and applications treated in this book should provide a basis for beginning work in the area or the evaluation of proposed systems. In addition, this book provides extensive references to the literature.

The book is intended primarily for optical engineers, scientists, and students who have a need to apply laser beam shaping techniques to improve laser processes. It should be a valuable asset to someone researching, designing, procuring, or assessing the need for beam shaping with respect to a given application. Due to the broad treatment of theory and practice in the book, we think it should also appeal to scientists and engineers in other disciplines.

We would like to express gratitude to the contributing authors, whose efforts made the book possible. We acknowledge the help of David L. Shealy, the first author contacted, who provided encouragement as well as serving as a confidant throughout the work. We give special thanks to our manager, James A. Wilder, Jr., for championing the project. It was a pleasure working with the staff of Marcel Dekker, Inc. We would like to acknowledge the initial evaluation and acceptance of the book proposal by the late Graham Garratt. Finally, we express our appreciation to the very helpful staff, Rita Lazazzaro, Eric Stannard, and Anja Schmidt.

Fred M. Dickey
Scott C. Holswade

Contents

Contributors

Daniel M. Brown MEMS Optical, LLC, Huntsville, Alabama

David Renick Brown Micro Optics, MEMS Optical, LLC, Huntsville, Alabama

Fred M. Dickey Sandia National Laboratories, Albuquerque, New Mexico

Neal C. Evans University of Alabama at Birmingham, Birmingham, Alabama

Scott C. Holswade Sandia National Laboratories, Albuquerque, New Mexico

Louis A. Romero Sandia National Laboratories, Albuquerque, New Mexico

Carlos B. Roundy Spiricon, Inc., Logan, Utah

David L. Shealy Physics Department, University of Alabama at Birmingham, Birmingham, Alabama

Louis S. Weichman Sandia National Laboratories, Albuquerque, New Mexico

1
Introduction

Scott C. Holswade and Fred M. Dickey
Sandia National Laboratories, Albuquerque, New Mexico

Beam shaping is the process of redistributing the irradiance and phase of a beam of optical radiation. The beam shape is defined by the irradiance distribution. The phase of the shaped beam is a major factor in determining the propagation properties of the beam profile. For example, a reasonably large beam with a uniform phase front will maintain its shape over a considerable propagation distance. Beam shaping technology can be applied to both coherent and incoherent beams.

Figure 1 illustrates the general beam shaping problem. A beam is incident upon an optical system that may consist of one or more elements. The optical system must operate upon the beam to produce the desired output. The desired output might only demand a certain irradiance distribution at a target plane, with the phase allowed to vary. This would be the entire problem for the case of incoherent beams. For coherent beams, the designer may or may not want to constrain the phase of the beam at the output plane. For example, if a collimated output beam is desired, the phase front of the beam exiting the optical system must be uniform. If the design only requires a certain irradiance distribution at the target plane, however, the optical system is usually simpler if the phase is left unconstrained.

The essence of the beam shaping problem is illustrated in Fig. 2. In the figure an input beam of rays, in this case with a Gaussian distribution, is incident on a plane representing the optical system. The output rays are bent so as to come to a uniform distribution in the output plane. This is the Gaussian to flat-top beam shaping problem. A wavefront can be computed by noting that in the geometrical optics approximation rays are normal to the wavefront. Once the wavefront is determined one can then determine a phase function that would produce the shaped beam. Implementing the

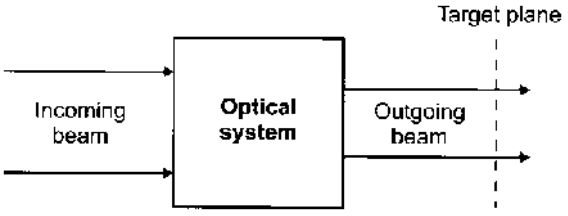


Figure 1 General beam shaping problem.

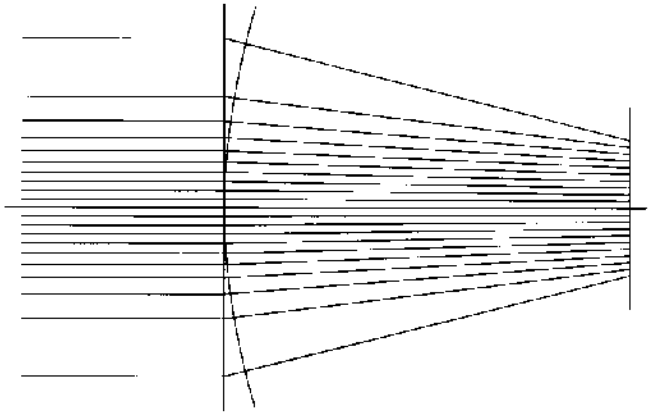


Figure 2 Shaping a Gaussian beam to a uniform beam.

phase function may be a complicated optical design process, and may be best achieved with multiple optical surfaces (elements).

In fact, the simple process outlined above can give a direct solution to a beam shaping problem if the phase function can be realized by an optically thin phase element. However, in general, determining the optical phase element is a complicated problem. Complications arise if the thin element approximation is not applicable, or if both phase and irradiance profiles are specified. The ray bending depicted in Fig. 2 is monotonic. Non-monotonic ray-bending schemes are also possible and some may offer advantage. A simple example of a non-monotonic system would be a ray-bending scheme that folds the upper half of the input into the lower half of the output and conversely. In fact, the bending might have a random component. It may be easy, if one is not careful, to arrive at bending schemes that are not realizable in that they do not, even approximately, satisfy Maxwell's equations. Further, certain ray-bending schemes may fall outside the realm of geometrical optics.

In addition to the above, what can be obtained for a given beam shaping problem is limited by physical optics (electromagnetic theory). The simplest result of this theory is an uncertainty principle that depends upon input beam size and wavelength, and output profile size and distance. This result is closely related to diffraction limits on imaging optical systems.

It is the study of various methods for arriving at the optical system that is the subject of this book. Only lossless or low-loss shaping techniques are considered. If large losses are allowed the beam shaping problem becomes trivial. For example, one can simply truncate a single-mode Gaussian beam to create a relatively uniform beam profile. As the truncating aperture becomes smaller, the remaining beam has a more uniform profile, but at the expense of large power losses in the beam. For most applications, efficiency requirements generally mandate low-loss systems. Low-loss beam shaping for laser sources can be broken down into two basic categories. The first, field mapping, includes methods where the input profile is redistributed at the target plane, such as the ray-bending scheme illustrated in Fig. 2. The second, beam integration, includes methods where the input beam is broken up into components, which assemble at the target plane to produce the desired profile. Both of these categories are developed in this book.

I. A BRIEF HISTORY

Beam shaping and its uses have a long history. Classical philosophers were familiar with burning glasses, and writings on optics survive from Empedocles (c. 490–430 BC) and Euclid (c. 300 BC) (1). The earliest reported use of beam shaping techniques is most likely a legend, but it merits some discussion. Early Greek and Roman historians reported its use by Archimedes during the siege of Syracuse from 213 to 211 BC. According to these reports, Archimedes devised a mirror system that was used to set fire to Roman ships as they entered the harbor. It is speculated that Archimedes equipped several hundred people with flat pieces of polished metal which acted as mirrors (2). Figure 3 illustrates this idea. If correct, this is the first use of multi-faceted beam shaping techniques, where an incoming beam is broken up into small areas that are directed onto a common target. This would also be a case of incoherent beam shaping.

Technically, the use of mirrors to set fire to wooden ships appears sound. In a modern experiment, 60 sailors holding 5-by-3 foot mirrors coated with bronze set fire to a wooden boat located about 160 feet away

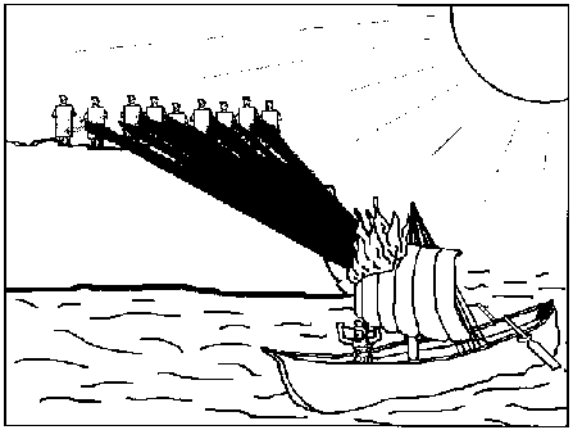


Figure 3 Use of multiple mirrors to set ships on fire during the siege of Syracuse, 213–211 BC. This is an example of using a multi-faceted, or beam integrating, technique to shape an incoherent source. (Figure courtesy of Stephen E. Yao).

(3). Under a fairly weak winter sun, the boat began to smoke in seconds, and flames appeared within two minutes. The objections to the legend of Archimedes revolve more around the historical record and the difficulty of making high-quality mirrors in the ancient world (4). First of all, a novel event such as defeating an enemy fleet by burning should have dominated contemporary accounts of the siege of Syracuse, but it is not mentioned. The story is only found in later histories. Second, it would be difficult for Archimedes to procure large, high-quality mirrors at that time in history, especially during a war. Mirrors were considered rare and precious until well into the 16^{th} century. It is likely that this debate will continue for some time.

The first practical beam shaping optic was the Fresnel lens used in lighthouses. This lens consists of concentric rings, each of which makes up part of a simple lens, as shown in Fig. 4. It concentrates the source light

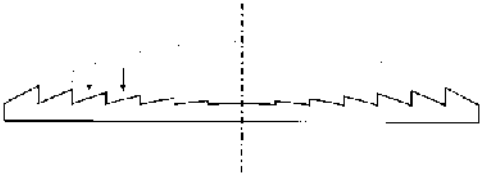


Figure 4 Cross section of a Fresnel lens. Each of the concentric rings has the shape of that part of a simple lens.

from lighthouses and searchlights into a relatively narrow beam. A simple lens sized for a lighthouse application would be prohibitively large and heavy. Instead, each of the concentric elements in the Fresnel lens is ground and polished separately. The elements are then assembled together to form the complete lens. For other applications with smaller lens sizes, the Fresnel element can be molded in one piece. It is interesting to note that Georges-Louis Leclerc de Buffon (1748) developed the idea of dividing a simple lens into concentric rings to reduce the weight. Fresnel used this idea in 1820 to construct lighthouse lenses.

Fresnel lenses for lighthouses are classified into seven classes, or orders. The distance from the light source to the lens determines the order. The first-order Fresnel lens, the largest, was installed in the larger lighthouses operating at seacoasts. Smaller, higher-order Fresnel lenses were installed in lighthouses that marked features in harbors, such as breakwaters. The Fresnel lens is an example of an incoherent beam shaping technique.

Modern laser beam shaping obviously had to wait until the arrival of the laser. The earliest known instance of lossless beam shaping was by Frieden in 1965 (5). He used geometric methods to determine the curvatures of aspheric lenses that would reshape a single-mode Gaussian beam into a beam with a uniform profile.

II. METRICS FOR BEAM SHAPING SYSTEMS

In general, there are many possible metrics for measuring the performance of a beam shaping system; in fact, the number is infinite. Only a few, however, are commonly used. Most others are not used because they are not analytically tractable. The most common metric is probably the mean (integrated) square difference between the desired and actual beam profile. This is just the L_2 norm of functional analysis (6). A linear space admits an infinity of norms designated as L_n. Other examples of metrics are the integrated absolute difference, L_1, and the maximum absolute difference, L_∞. Non-analytical criteria could include ad hoc experimental performance and "it looks good." If a metric has an established history for a given application, it is best, where practicable, to use the same metric for design and analysis.

The mean square difference is appealing because it is analytically very tractable and has an energy or power interpretation. Parseval's theorem of Fourier analysis and the Cauchy–Schwarz inequality are examples of powerful analytical tools that can be applied to the L_2 function space. In general, it is not desirable to have a large integrated energy difference between the

desired and actual beam shaper output. The integrated absolute difference is a viable criterion but it is less tractable analytically. It can readily, however, be employed with numerical techniques. Of the three mathematical criteria discussed here the absolute difference is the least useful. Two functions could have a large absolute difference over a small region, with only a small variation in their mean square difference. Since the mean square difference relates to the beam's energy distribution, it more closely follows the parameter of interest for most applications. In some situations, it may be important to smooth the beam function before applying a metric to analyze the goodness of the solution. For example, in a material removal application a fine structure in the beam would not be resolvable by the physical process of material ablation. In designing such as system, it may be desirable that the beam's irradiance averaged over a certain area be flat with respect to some metric.

The discussion so far has been concerned with the energy or amplitude of the optical field. In addition, there may be simultaneous constraints of the error in the phase of the optical field. The phase of the optical field is generally not important in energy-intensive applications such as material processing, cutting, and welding. The phase of the shaped beam is likely to be important in applications involving optical signal processing, interferometry, holography, and lithography. Any of the metrics discussed above can be applied to the phase of the shaped beam; however, the energy interpretation of the mean square difference is not associated with the phase of the optical field.

In describing a physical problem, such as beam shaping, mathematically, care must be taken that the mathematical description matches what is desired. For example, a profile with the lowest mean square difference to a uniform profile may not look the "best" to the designer who compares it to other profiles. The following example illustrates this point.

The problem of mapping a single-mode Gaussian beam into a uniform amplitude and phase beam using a pure phase function is introduced in Sec. III.B.2 of Chapter 3. For certain system geometries, one obtains a minimum mean square error solution when the input and output beam sizes are such that β (see Chapters 2 and 3) is fairly small. This solution, shown in Fig. 5, is limited by diffraction effects. If the output beam size is allowed to increase, one can get a flatter looking output beam such as shown in Fig. 6. However the mean square error between this shaped beam and the desired output is larger than that of Fig. 5. In fact, the mean square error increases to a maximum with increasing output beam size, even though the profile continues to look more uniform. In other words, larger output profiles may not be what is desired, but they suffer less from diffraction and thus look better.

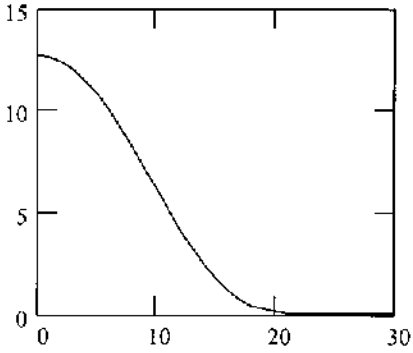


Figure 5 Solution with minimum mean square error to desired uniform profile.

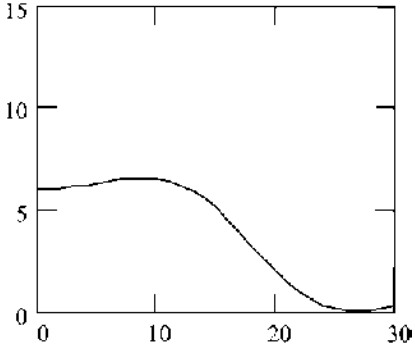


Figure 6 This solution looks more appealing than the solution of Fig. 5, but it has a larger mean square error with respect to the desired profile. This is because the profile is larger than the desired profile.

III. DISCUSSION OF CHAPTERS

The theoretical underpinnings of laser beam shaping are developed by L. A. Romero and F. M. Dickey in Chapter 2. After the introduction, the authors present a review of mathematical techniques appropriate to beam shaping. The mathematics discussed include Fourier analysis in one and two dimensions, the uncertainty principle and the associated concept of space bandwidth product, and the method of stationary phase, which is an important tool for evaluating diffraction integrals. They then give a review of physical and geometrical optics theory. The wave equation and flux flow concepts are developed starting with Maxwell's equations.

Geometrical optics is presented beginning with Fermat's principle followed by the development of the Eikonal equation. The Eikonal equation is first obtained as a high-frequency limit of the scalar wave equation and then the vector form is obtained as a high-frequency limit of Maxwell's equations. Fermat's principle is then revisited as the basis of geometrical optics. This section is illustrated with eight interesting examples.

The basics of diffraction theory and Fourier optics are reviewed as the physical optics basis for laser beam shaping. The Fresnel diffraction integral is derived in an interesting way using Fourier transform theory. The limits of the validity of the Fresnel approximation are outlined. Although the Fresnel integral is adequate for most beam shaping problems, vector diffraction theory is also introduced.

Finally, after the introduction of considerable background material, the beam shaping problem is addressed. In this chapter the laser beam shaping problem is treated as the Fourier transform of the product of the beam function and a phase element (a system consisting of a phase element and focusing lens); however, the solutions are very general. This can be seen from the fact that the phase element and focusing lens can be combined to form one element. Using Fermat's principle, the authors solve the problem of transforming a Gaussian beam (single mode) into a flat-top beam with square or circular cross section. They also outline the solution for more general distributions.

The chapter culminates in determining the solutions to the same problems using diffraction theory (Fourier optics). The relation between these solutions and the solutions obtained using geometrical optics is emphasized. The major difference is the introduction of a parameter, β. This parameter, which is related to the uncertainty principle, is a measure of the potential goodness of the solution. That is, it is a measure of how well the optics can map the input beam into the specified output irradiance. The quality of the mapping increases with increasing β. For sufficiently small β no good solution can be obtained regardless of design method. This is a major contribution to the beam shaping problem.

Although Chapter 2 may be a little too theoretical for the interest of some readers, we would like to encourage anyone interested in the book to at least read the introduction to this chapter. It is a succinct introduction to the beam shaping problem that discusses general difficulties that can arise.

In Chapter 3, F. M. Dickey and S. C. Holswade reduce the theoretical solutions of Chapter 2 for circular and square flat-top beams to engineering designs. After introducing the beam shaping problem as a minimum-mean-square error problem, they revisit the uncertainty principle and the method of stationary phase, presenting the uncertainty principle in a slightly different form than in Chapter 2.

They then give the solution to the circular and flat-top shaping problem as a stationary phase solution to the Fourier optics configuration. Two solutions, the positive and negative phase solutions, are introduced and the effects of wavefront curvature are treated in detail. Methods for producing collimated uniform irradiance beams are described.

The remainder of the chapter includes considerable detail on design considerations, alignment and scaling errors, and method of design. These sections include many numerical examples and design data. The chapter concludes with experimental data from a system designed and fabricated for use with a CO_2 laser.

In many cases, the system geometry is such that diffraction effects can be ignored, and we are in the realm of geometrical techniques. D. L. Shealy discusses these methods in Chapter 4. One benefit of the approach of Chapter 4 is that it allows the designer to take a general input shape and convert it to a general output shape. For example, a Gaussian input beam could be converted to a uniform profile on the surface of a sphere.

A general theory of the two lens refractive beam profiler is developed starting with the eikonal and invoking conservation of energy along a ray bundle between two surfaces. Special attention is given to constant optical path length designs that give collimated (minimum divergence) outputs. The result is a differential equation for the lens surfaces. Numerical evaluation and experimental results are presented. Although the theory is developed for two lens systems, the theory is basic and amenable to extensions to systems with more than two elements.

Next, the author develops parallel methods for a two lens system that uses gradient-index (GRIN) glasses. An advantage of this approach is that it allows for the replacement of aspherical surfaces obtained in the previous case with axial-GRIN elements and spherical surfaces. A design example with numerical evaluation is given. Finally, the author develops the detailed general equations for one and two element reflective (mirror) systems.

In Chapter 5, N. C. Evans and D. L. Shealy discuss the use of computer-optimization techniques for beam shaping problems, concentrating on the use of genetic algorithms (GAs). They first introduce three increasingly difficult beam shaping problems that have been solved using other techniques, in order to judge the effectiveness of optimization-based methods. All three of these problems involve transforming a Gaussian input beam into a uniform distribution of some sort. They next describe a computational method for determining irradiance profiles, which builds upon proven ray-tracing techniques.

With a computational technique identified for quantifying the efficacy of a beam shaper, in other words, a merit function, an optimization technique can now be employed. The authors discuss optimization methods, and

point out the advantages of simulated annealing and genetic algorithms for solving problems with discrete parameters, such as glasses from a glass catalog. They then address the genetic algorithm in some detail. One of the interesting aspects of optics problems is that the evaluation of the merit function takes far more time than the optimization steps. The authors discuss ways to evaluate merit functions for several options in parallel, greatly speeding up the overall process.

The authors then apply this technique to the problems described earlier. The genetic algorithm is able to find a satisfactory solution for all three cases in a significant but reasonable amount of time, since no user input is required after the problem is started. These examples demonstrate that optimization-based methods work, although the chosen problems can be solved in more general ways using analytical methods. With a general analytical solution, particular cases can be rapidly solved. In addition, analytical methods often offer insight into the beam shaping problem, including ways to change things to improve the solution. However, one does not have to go very far before the problem of obtaining an analytical solution becomes very difficult if not intractable. The most promising applications of optimization techniques lie in those areas that are analytically difficult, such as non-symmetric problems. As an example, the conversion of a Gaussian beam into a uniform beam with a triangular cross-section would fall into this category. Also, optimization techniques might be useful in obtaining improved solutions of the type discussed in Chapter 3 for the case of intermediate β values.

In some cases, a multi-mode laser beam may need to be reshaped into a complex output geometry. In Chapter 6, D. R. Brown discusses a powerful method for diffusing optical radiation over a well-controlled angular distribution. After discussing the differences of this technique from beam shapers designed for known input beam profiles, the author covers the grating theory required to understand the diffuser. He then covers a simple, illustrative mathematical description of diffuser design. To summarize, the desired output irradiance is randomized. Then, using a Fourier transform (far-field) relation between the input and output, the output is inverse transformed. This function is then binarized to give a two-state phase function that is the basis for the beam shaping element. It is interesting to note that this approach is similar, in some ways, to phase-only filtering techniques, which have been given considerable attention in the field of optical pattern recognition.

The actual method is described through a design example of a simple system that produces a ring profile. For some applications, the desired pattern will contain angles that approach the null points of the sinc function that arises from the finite pixelation of phase in the diffuser. Fabrication of

these devices involves the tradeoff of the overall diffraction envelope (caused by the technique's pixelation of phase) versus the number of phase steps in the diffuser. An example of an advanced fabrication technique is presented that eliminates these effects. The author next discusses the speckle in the output pattern that results from diffusers and methods for its reduction. He concludes the chapter with a discussion of applications where diffractive diffusers are particularly appropriate.

High power/energy lasers frequently produce highly irregular multimode beams that are non-stationary in time. A common approach to producing uniform irradiance beams in this case is to break up the input beam into discrete apertures and superimpose the resulting beamlets to get an average irradiance. This technique is sometimes referred to as beam integration. D. M. Brown, F. M. Dickey and L. S. Weichman discuss multi-aperture beam integration techniques in Chapter 7.

The authors present optical refractive, reflective, and diffractive configurations for beam integration and discuss the effects of diffraction and interference. Diffraction effects can be mitigated by increasing system complexity. Interference effects are always present, but will depend on the beam coherence and may not be resolved in some applications. The authors also discuss design and fabrication considerations. Finally, they present experimental data and discuss applications of the technique.

Although this book is primarily directed toward laser beam shaping, D. L. Shealy has contributed a chapter on classical (non-laser) irradiance mapping techniques. In Chapter 8 he treats the problem of transforming an input beam (source) profile into an irradiance distribution over a specified surface. Generally, for non-laser sources, geometrical optics is the approach of choice for designing irradiance shaping optical systems. The author, as in Chapter 4, develops geometrical optics methods for the non-laser problem. Ray tracing and the flux flow equation are combined to obtain a differential equation for the shaping surfaces. He also develops an approach to the problem using the energy balance equation presented in Chapter 4. Finally, the author discusses the application of the theory to the shaping of radiation from point and Lambertian sources. In addition to the theory developed in this chapter, the author provides references for those interested in further pursuit of the subject.

When beam shaping is used in an application, some sort of beam analysis system is usually required, if for no other reason than to verify the performance of the beam shaper. Oftentimes, the input beam must be measured to properly fabricate the shaping optics. In some cases, the analysis will indicate which form of shaping is appropriate. For example, a highly multimode beam would not work well with a Gaussian to flat-top shaping system. A beam integrator would work much better in this case.

In Chapter 9, C. Roundy provides an introduction to this subject as well as guidance on the many beam analysis methods available. He begins with laser beam applications and the need for accurate beam measurement. He then discusses manual beam measurement methods, and then moves into mechanical scanning instruments. The treatment of camera-based methods centers around the many considerations involved in choosing a camera for a particular application. Next, techniques are discussed for directing a faithful copy of the laser beam to the camera at an acceptable power or energy level. Finally, of particular importance to Gaussian beam shaping systems are an accurate measurement of the incoming beam diameter and a measurement of the degree to which the actual beam departs from the Gaussian ideal. Chapter 9 discusses many of the issues surrounding these measurements.

IV. CHOOSING A SHAPING TECHNIQUE

Laser beam shaping techniques group into two basic types: field mappers and beam integrators. Field mappers work only for beams with a known field distribution, such as single-mode beams, and they are generally highly sensitive to alignment and beam dimensions. Integrators work for both coherent and multimode beams, where the input field distribution may not be known, and they are much less sensitive to alignment and beam size. However, interference effects are a problem with integrators, especially for coherent beams. This categorization into mappers and integrators is independent of the method, such as diffraction theory or geometrical optics, used to design the beam shaping optics. Field mapping would include the diffraction-theory techniques of Chapters 2 and 3, the geometrical techniques of Chapter 4, and the optimization techniques of Chapter 5. The category of beam integration would include the diffuser approach of Chapter 6 and the direct beam integration approach of Chapter 7. The non-laser techniques of Chapter 8 are concerned with incoherent sources, but they share similarity with field mappers for laser sources.

Like many optical problems, there is no single beam shaping method that addresses all situations. The nature of the input beam, the system geometry, and the quality of the desired output beam all affect the choice of technique. In considering a beam shaping application for single-mode Gaussian beams, it is first advisable to calculate the parameter

$$\beta = \frac{2\sqrt{2\pi}r_0 y_0}{f\lambda}, \tag{1}$$

where λ is the wavelength, r_0 is the radius at the $1/e^2$ point of the input beam, y_0 is half-width of the desired output dimension, and f is the focal length of the focusing optic, or the working distance from the optical system to the target plane for systems without a defined focusing optic. For beams or output spots that don't have circular symmetry, approximate heights and widths can be used. As explained in Chapters 2 and 3, this parameter sets limits on the quality of the solution available. For simple output geometries such as circles and rectangles, some rules of thumb can be defined: If $\beta < 4$, a beam shaping system will not produce acceptable results. For $4 < \beta < 32$, diffraction effects are significant and should be included in the development of the beam shaping system. For $\beta > 32$, diffraction effects should not significantly degrade the overall shape of the output beam. However, interference effects will still be present from multifaceted or diffusing approaches. These may or may not be an issue for a particular application. For more complicated output geometries, β may have to be significantly higher to produce acceptable results. There are no good general rules for these cases.

The particular beam shaping approach depends on the nature of the input beam and the desired output shape. If the desired output shape is a complicated, non-symmetric pattern, then the diffuser approach of Chapter 6 is probably a good option. For relatively simple output shapes, such as the circles, rectangles, and squares used in material processing, there are several other approaches. Some of these depend on the quality of the input beam. If the input beam is multimode, with significant irradiance variations across its profile, then the multi-aperture approach of Chapter 7 is most appropriate. This approach reduces the effect of input beam variations, but at the expense of interference effects in the output profile. The multi-aperture approach also works well when the input beam size changes over time or between sources. The diffuser approach of Chapter 6 can also work in this situation.

For applications where the input beam has a well-characterized profile, the nature of the output beam can suggest an approach. If the output pattern must be projected onto a non-planar surface, the geometric techniques of Chapter 4 appear to work well, as long as β is sufficiently large. The geometrical approach also works well for producing a shaped beam that propagates with the desired profile, rather than just a shaped profile at a particular plane.

Oftentimes, the input beam is single-mode with a Gaussian profile, and the desired output profile is either circular or rectangular. In these cases, the diffraction-based technique of Chapter 3 works particularly well. This technique produces an output profile free of interference effects. The profile degradation due to diffraction is also predicted by the solution, and the technique suggests ways to reduce this degradation to an acceptable level. Finally, the solution is known and can be easily scaled to any situation.

As long as the input profile is known, the optimization techniques of Chapter 5 have the potential to work well for a number of desired output patterns or beams, including non-symmetric profiles. These techniques also have the advantage of a minimum of mathematical development to arrive at a solution. In addition, the basic algorithms, once implemented, would remain the same for designing additional output patterns.

V. FURTHER INFORMATION

Laser beams, particularly single-mode Gaussian beams, are often resized for particular applications. This resizing can be different in the two axes of the beam profile. Thus, circular beams can be made elliptical, or elliptical beams further modified. Cylindrical telescopes or prism pairs can be used for such transformations. These elementary techniques are sometimes termed laser beam shaping, but they are not discussed in this book. Several textbooks discuss them in more detail (7, 8). Laser beam shaping, in the sense of transforming a Gaussian beam profile to a uniform profile, is briefly reviewed by O'Shea (7). The authors do not know of another comprehensive treatment of this subject.

In this book we have tried to give a broad treatment of the beam shaping problem that would provide the reader with the background to develop solutions to specific problems, evaluate various techniques, and conduct research in the field. We have also presented solutions to problems of general interest. However, it is difficult to present or cite all relevant work that might interest a given reader with a specific application in mind. For example, iterative approaches such as the Gerchberg–Saxton algorithm (9) are not treated in this book, and techniques for super-Gaussian and flattened Gaussian beam expansions are touched upon only slightly. Also, there are many papers in the literature that address a specific implementation, such as a particular material or geometry. For this reason we have included at the end of this chapter a bibliography of beam shaping papers. Some of these papers are cited in the chapters; many are not. Three books that touch upon aspects of material processing and beam shaping are also included.

REFERENCES

1. M Born, E Wolf. Principles of Optics. Elmsford NY: Pergamon Press, 1980, p. xxi.
2. See AC Claus. On Archimedes' burning glass, and ON Stavroudis, Comments on: On Archimedes' burning glass. Applied Optics, 12(10):A14–A16, 1973.

3. MS Modiano. How Archimedes stole Sun to burn foe's fleet. New York Times. 11 November, 1973, p. 16.
4. MS Browne. Archimedes' Mirrors: some new reflections. New York Times. 11 March, 1978, p. 25.
5. BR Frieden. Lossless conversion of a plane laser wave to a plane wave of uniform irradiance. Applied Optics, 4(11):1400–1403, 1965.
6. E Kreszig. Introductory Functional Analysis with Applications. New York: John Wiley and Sons, 1978.
7. DC O'Shea. Elements of modern optical design. New York: John Wiley & Sons, 1985, pp. 230–266.
8. BEA Saleh, MC Teich. Fundamentals of Photonics. New York: John Wiley & Sons, 1991, pp. 92–100.
9. H Stark. Image recovery: theory and application. New York: Academic Press, 1987.

APPENDIX: ADDITIONAL READING

H Aagedal, M Schmid, T Beth, S Teiwes, Ff Wyrowski. Theory of speckles in diffractive optics and its application to beam shaping. Journal of Modern Optics 43(7):1409–1421, 1996.

H Aagedal, M Schmid, S Egner, J Muller-Quade, T Beth. Analytical beam shaping with application to laser-diode arrays. J Opt Soc Am A, 14(7):1549–1553, 1997.

K Ait-Ameur, F Sanchez. Gaussian beam conversion using an axicon. J Modern Optics 46(10):1537–1548, 1999.

CC Aleksoff, KK Ellis, BD Neagle. Holographic conversion of a Gaussian beam to a near-field uniform beam. Optical Engineering 30(5):537–543, 1991.

Y Arieli, N Eisenberg, A Lewis, I Glaser. Geometrical-transformation approach to optical two-dimensional beam shaping. Applied Optics, 36(35):9129–9131, 1997.

Y Arieli, N Eisenberg, A Lewis, I Glaser. Geometrical-transformation approach to optical two-dimensional beam shaping: erratum and reply to comment. Applied Optics 37(35):8226, 1998.

J Armengol, F Vega, N Lupon, F Laguarta. Two-faceted mirror for active integration of coherent high-power laser beams. Applied Optics 36(3):658–661, 1997.

K Baker, D Shealy, W Jiang. Directional light filters: three-dimensional azo dye formed micro-honeycomb images within optical resins. Proc SPIE 2404:144–158, 1995.

KM Baker. Highly corrected close-packed microlens arrays and moth-eye structuring on curved surfaces. Applied Optics 38(2):339–351, 1999.

KM Baker. Highly corrected submicrometer grid patterning on curved surfaces. Applied Optics 38(2):352–356, 1999.

I Bass. Closely spaced mirror pair for reshaping and homogenizing pump beams in laser amplifiers. Proc SPIE 1869:128–135, 1993.

J Bernges, L Unnebrink, Th Henning, EW Kreutz, R Poprawe. Mask adapted beam shaping for material processing with excimer lasers. Proc SPIE 3573:108–111, 1998.
J Bernges, L Unnebrink, Th Henning, EW Kreutz, R Poprawe. Novel designs concepts for UV beam shaping. Proc SPIE 3779:118–125, 1999.
L Bincheng, E Welsch. Probe-beam diffraction in a pulsed top-hat beam thermal lens with a mode-mismatched configuration. Applied Optics 38(24):5241–5249, 1999.
R Borghi, F. Gori, M Santarsiero. Optimization of Laguerre–Gauss truncated series. Optics Comm 125:197–203, 1996.
R Borghi, M Santarsiero. Modal decomposition of partially coherent flat-topped beams produced by multimode lasers. Optics Letters 23(5):313–315, 1998.
DR Brown, A Kathman. Multi-element diffractive optical designs using evolutionary programming. Proc SPIE 2383:1995.
D Burkhard, DL Shealy. A different approach to lighting and imaging: formulas for flux density, exact lens and mirror equations and caustic surfaces in terms of the differential geometry of surfaces. Proc SPIE 692:248–272, 1986.
Q Cao. Geometrical-transformation approach to optical two-dimensional beam shaping: comment. Applied Optics 37(35):8224–8225, 1998.
PS Carney, G Gbur. Optimal apodizations for finite apertures. J Opt Soc Am A 16(7):1638–1640, 1999.
EG Churin. Diffraction-limited laser beam shaping by use of computer-generated holograms with dislocations. Optics Letters 24(9):620–622, 1999.
W-X Cong, N-X Chen, B-Y Gu. Beam shaping and its solution with the use of an optimized method. Applied Optics 37(20):4500–4503, 1998.
A Cutolo, G Calafiore, L Zeni. Supergaussian mirrors based on thermo-optical effect. Proc SPIE 1869:149–157, 1993.
D Dagenais, J Woodroffe, I Itzkan. Optical beam shaping of a high power laser for uniform target illumination. Applied Optics 24(5):671–675, 1985.
X Deng, S Liang, Z Chen, W Yu, R Ma. Uniform illumination of large targets using a lens array. Applied Optics 25(3):377–381, 1986.
F M Dickey, SC Holswade. Gaussian laser beam profile shaping. Optical Engineering 35(11):3285–3295, 1996.
FM Dickey, BD O'Neil. Multifaceted laser beam integrators, general formulation and design concepts. Opt Eng 27(11):999–1007, 1988.
S Dixit, J Lawson, K Manes, H Powell, K Nugent. Kinoform phase plates for focal plane irradiance profile control. Optics Letters 19(6):417–419, 1994.
VJ Doherty. Design of mirrors with segmented conical surfaces tangent to a discontinuous aspheric base. Proc SPIE 399:263–271, 1983.
T Dresel, M Beyerlein, J Schwider. Design and fabrication of computer-generated beam-shaping holograms. Applied Optics 35(23):4615–4621, 1996.
T Dresel, M Beyerlein, J Schwider. Design of computer generated beam-shaping holograms by iterative finite-element mesh adaption. Applied Optics 35(35):6865–6874, 1996.
MT Eismann, AM Taj, JN Cederquist. Iterative design of a holographic beamformer. Applied Optics 28(13):2641–2650, 1989.

NC Evans, DL Shealy. Design and optimization of an irradiance profile-shaping system with a genetic algorithm method. Applied Optics 37(22):5216–5221, 1998.
R Facklam, K Guenther, S Velsko. Intracavity and extracavity control of laser beam properties. Proc SPIE 1869:1993.
MA Golub, M Duparre, EB Kley, R Dowarschik, B Ludge, W Rockstroh, HJ Fuchs. New diffractive beam shaper generated with the aid of e-beam lithography. Opt Eng 35(5):1400–1406, 1996.
J Gordon, A Rabl. Reflectors for uniform far-field irradiance: Fundamental limits and example of an axisymmetric solution. Applied Optics 37(1):44–47, 1998.
F Gori. Flattened gaussian beams. Optics Communications, 107:335–341, 1994.
F Gori, M Santarsiero, R Borghi, G Guattari. Intensity-based modal analysis of partially coherent beams with Hermite–Gaussian modes. Optics Letters 23(13):3989–3991, 1998.
DW Griffin. Gaussian beam intensity flattener. Optical Engineering 37(7):2185–2187, 1998.
V Hedin, R Martinez, B O'Neil, D Hodgkiss. Production of a line image from a multi-mode high energy laser. Proc SPIE 1625:32–43, 1992.
S Heinmann. Computer generated beam shaping and focusing optical elements for laser material processing. Optics Communications 119:613–622, 1995.
Th Henning, M Scholl, L Unnebrink. UV laser beam shaping by multi-faceted beam integrators: fundamental principles and advanced design concepts. Proc SPIE 2703:62–73, 1996.
Th Henning, M Scholl, L Unnebrink, U Habich, R Lebert, G Herziger. Beam shaping for laser material processing with non-rotational symmetric optical elements. Proc SPIE 2703:126–129, 1996.
J Hobfeld, E Jager, T Tschudi, E Churin, V Koronkevich. Rectangular focus spots with uniform intensity profile formed by computer generated holograms. Proc SPIE 1574:159–166, 1991.
SC Holswade, FM Dickey. A system design for Gaussian laser beam shaping. OSA Annual Meeting, paper ThQQ3, 14 September 1995.
SC Holswade, FM Dickey. Gaussian laser beam shaping: test and evaluation. Proc SPIE 2863:237–245, 1996.
XG Huang, MR Wang, C Yu. High-efficiency flat-top beam shaper fabricated by a nonlithographic technique. Optical Engineering 38(2):208–213, 1999.
S Jahan, M Karim. Refracting systems for Gaussian-to-uniform beam transformations. Optics & Laser Technology 21(1):27–30, 1989.
W Jiang, D Shealy, J Martin. Design and testing of a refractive reshaping system. Proc SPIE, 2000:64–75, 1993.
H-J Kahlert, U Sarbach, B Burghardt, B Klimt. Excimer laser illumination and imaging optics for controlled microstructure generation. Proc SPIE 1835:110–118, 1992.
M Karim, A Cherri, A Awwal, A Basit. Refracting system for annular laser beam transformation. Applied Optics 26(12):2446–2449, 1987.
T Kosoburd, J Kedmi. Beam shaping with diffractive optical elements. Proc SPIE 1971: 390–399, 1993.

V Kotlyar, S Khonina, V Soifer. Calculation of phase formers of non-diffracting images and a set of concentric rings. Optik 102(2):45–50, 1996.
M Kuittinen, P Vahimaa, M Honkanen, J Turunen. Beam shaping in the nonparaxial domain of diffractive optics. Appl Opt 36:2034–2041, 1997.
W Lee. Method for converting a Gaussian laser beam into a uniform beam. Opt Comm 36(6):469–471, 1981.
Q Li, H Gao, Y Dong, Z Shen, Q Wang. Investigation of diffractive optical element for shaping a Gaussian beam into a ring-shaped pattern. Optics and Laser Technology, 30:511–514, 1999.
B Li, S Martin, E Welsh. Pulsed top-hat thermal-lens measurement for ultraviolet dielectric coatings. Optics Letters, 24(20):1398–1400, 1999.
B Lu, B Zhang, S Luo. Far-field intensity distribution, M^2 Factor, and propagation of flattened Gaussian Beams. Applied Optics 38(21):4581–4584, 1999.
B Lu, B Zhang, X Wang. Three-dimensional intensity distribution of focused super-Gaussian beams. Optics Comm 126:1–6, 1996.
PH Malyak. Two-mirror unobscured optical system for reshaping the irradiance redistribution of a laser beam. Appl Opt 31(22):4377–4383, 1992.
SM Metev, VP Veiko, Laser-Assisted Microtechnology. New York: Springer-Verlag, 1998.
L Migliroe (editor). Laser Materials Processing. New York: Marcel Dekker, 1996.
JC Miller, RF Haglund, Jr. Laser Ablation and Desorption. New York: Academic Press, 1998.
M Mrochen, V Semchishen, V Seminogov, T Seiler. Beam homogenization using a defined statistical surface relief for medical applications. Laser und Optoelektronik 30(4):71–76, 1998.
JL Ocana, A Garcia-Beltran, F Laguarta, J Armengol, N Lupon, F Vega. Laser heat treatments driven by integrated beams: role of irradiation nonuniformities. Applied Optics 38(21):4570–4576, 1999.
V Oliker, L Prussner, DL Shealy, S Mirov. Optical design of a two-mirror asymmetrical reshaping system and its application in superbroadband color center lasers. Proc SPIE 2263:11–18, 1994.
J Pu, S Nemoto, H Zhang. Reshaping Gaussian Schell-model beams to uniform profiles by lenses with spherical aberration. J Modern Optics 46(11):1611–1620, 1999.
PW Rhodes, DL Shealy. Refractrive optical systems for irradiance redistribution of collimated radiation: their design and analysis. Applied Optics 19(20):3545–3553, 1980.
NC Roberts. Beam shaping by holographic filters. Appl Opt 28(1):31–32, 1989.
LA Romero, FM Dickey. Lossless laser beam shaping. J Opt Soc Am A 13(4):751–760, 1996.
M Rossi, T Hessler. Stray-light effects of diffractive beam-shaping elements in optical microsystems. Applied Optics 38(14):3068–3076, 1999.
M Santarsiero and R Borghi. Correspondence between super-Gaussian and flattened Gaussian beams. J Opt Soc Am. A 16(1):188–190, 1999.
DJ Schertler, N George. Uniform scattering patterns from grating-diffuser cascades for display applictions. Applied Optics 38(2):291–303, 1999.

D Shafer. Gaussian to flat-top in diffraction far-field. Applied Optics 36(34):9092–9093, 1997.
C Sheppard, S Saghafi. Flattened light beams. Optics Comm. 132:144–152, 1996.
GJ Swanson, WB Veldkamp. Diffractive optical elements for use in infrared systems. Optical Engineering 28(6):605–608, 1989.
W Sweatt. Transforming a circular laser beam into a square or trapezoid—almost. Optical Engineering 31(2):245–250, 1992.
W Sweatt, R Setchell, M Warren, M Farn. Injecting a pulsed yag laser beam into a fiber. Proc SPIE 3010:266–271, 1997.
X Tan, B Gu, G Yang, B Dong. Diffractive phase elements for beam shaping: a new design method. Applied Optics 34(8):1314–1320, 1995.
L Unnebrink, Th Henning, EW Kreutz, R Poprawe. Excimer beam homogenization for materials processing. Proc SPIE 3573:126–129, 1998.
L Unnebrink, Th Henning, EW Kreutz, R Poprawe. Optical system design for excimer laser materials processing. Proc SPIE 3779:413–422, 1999.
L Unnebrink, Th Henning, M Scholl, EW Kreutz. Advanced concepts for UV laser beam shaping with non-rotationally symmetric optical elements. Proc SPIE 3097:328–334, 1997.
B Veldkamp. Laser beam profile shaping with binary diffraction gratings. Opt Comm 38(5,6):381–386, 1981.
WB Veldkamp. Laser beam profile shaping with interlaced binary diffraction gratings. Applied Optics 21(17):3209–3212, 1982.
L Videau, C Rouyer, J Garnier Amigus. Motion of hot spots in smoothed beams. J Opt Soc Am. A 16(7):1672–1681, 1999.
FJ Villarreal, HJ Baker, RH Abram, DR Jones, DR Hall. Beam reformatting of one- and two-dimensional arrays of CO_2 waveguide lasers. IEEE J Quantum Electronics 35(3):267–272, 1999.
C Wang and D Shealy. Differential equation design of finite-conjugate reflective systems. Applied Optics 32(7):1179–1188, 1993.
W Yen, C Huang, H Liu, L Lee. A Nd:YAG laser with a flat-top beam profile and constant divergence. Optics & Laser Technology 29(2):57–61, 1997.
Z Zeng, N Ling, W Jiang. The investigation of controlling laser focal profile by deformable mirror and wave-front sensor. J Modern Optics 46(2):341–348, 1999.
J Zhai, Y Yan, G Jin, M Wu. Design of kinoform phase plate for spatial beam smoothing. J Modern Optics 46(4):667–674, 1999.
Y Zhang, B Dong, B Gu, G Yang. Beam shaping in the fractional Fourier transform domain. J Opt Soc Am A 15(5):1114–1120, 1998.

2

The Mathematical and Physical Theory of Lossless Beam Shaping

Louis A. Romero and Fred M. Dickey
Sandia National Laboratories, Albuquerque, New Mexico

I. INTRODUCTION

In this chapter we will present the basic mathematics and physics that are required to understand the theory of lossless beam shaping. Figure 1 is a diagram of the physical situation that we are concerned with. We assume that a parallel beam of coherent light enters an aperture at the plane $z = 0$. At the aperture the light gets refracted by a combination of a Fourier transform lens with focal length f, and a beam shaping lens. We are interested in the irradiance of the beam at the focal plane $z = f$. The separation of the refractive elements at the aperture into a Fourier transform lens and a beam shaping lens is convenient for our analysis, and sometimes convenient in practice, but it should be emphasized that these two lenses could in fact be combined into a single lens.

The beam shaping problem is concerned with how to choose the beam shaping lens so that we can transform a beam with an initial irradiance distribution at the plane $z = 0$ into a beam with a desired irradiance distribution at the focal plane $z = f$. We assume that the beam shaping lens is lossless. This means that it does not absorb or block out any of the energy of the incoming beam. If we assume that the laws of geometrical optics apply, it is possible to transform any initial distribution into any desired output distribution, provided only that the total energy of the incoming and outgoing beams are the same. When we include the effects of diffraction it is in general not possible to accomplish our goal exactly.

One of the major themes of this chapter is to determine the scaling

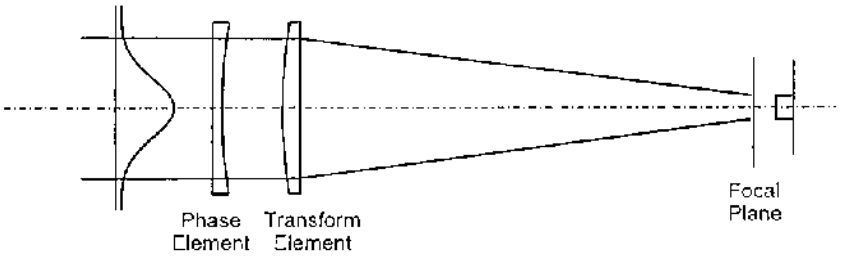


Figure 1 This is a schematic of the basic beam shaping system. A parallel beam of light enters the aperture where it encounters a Fourier transform lens, and then the beam shaping element. We would like to choose the beam shaping element so that the output at the focal plane has the desired intensity distribution.

properties of beam shaping systems (what happens when we make our system bigger or smaller, or change the wavelength?). In particular we want to know when the laws of geometrical optics can successfully be applied to designing our system. Due to our emphasis on scaling, we choose to write many of our functions in terms of dimensionless coordinates. For example, if the incoming beam has a radially symmetric Gaussian irradiance distribution many authors would write the irradiance distribution as

$$I(r) = g(r), \tag{1a}$$

where

$$g(r) = e^{-r^2/R^2}. \tag{1b}$$

Here the parameter R determines the basic scale of the irradiance distribution. In this chapter, we would prefer to write this irradiance distribution as

$$I(r) = g(r/R), \tag{2a}$$

where

$$g(\xi) = e^{-\xi^2}. \tag{2b}$$

It might appear simpler to say that the initial irradiance is given by $g(r)$, rather than saying it is given by $g(r/R)$. However, when we consider the scaling properties, the second form is much more powerful. In particular, if we say that the initial distribution is given by $g(r/R)$, then it will be much clearer how to apply the analysis of a system with distribution $g(r/R_1)$ to a system with distribution $g(r/R_2)$.

This approach is motivated by the practice commonly used in fluid mechanics of writing equations in dimensionless form (1, 2). This approach

in fluid mechanics allows one to show that different physical systems will have the same behavior provided only that certain "dimensionless parameters" are the same. For example, when fluid flows past a sphere, the behavior of the flow depends on the Reynolds number

$$\mathrm{Re} = \frac{RU_0}{\nu}, \tag{3}$$

where R is the radius of the sphere, U_0 is the velocity far from the sphere, and ν is the kinematic viscosity. If two flows have the same Reynolds number, then the patterns of fluid flow will be identical, after rescaling our coordinates. However, if the Reynolds numbers are different, then the flow patterns can look dramatically different. For example, in one case the flow may be turbulent, and in the other case not.

Ideas similar to these can be applied to the theory of beam shaping. Suppose our initial irradiance distribution is given by $g(x/R, y/R)$, and that our desired output irradiance distribution is given by $Q(x/D, y/D)$. The parameter R gives the characteristic length of the incoming beam, and D is the characteristic length of the output beam. If the wavelength of the light is λ, and we are imaging our output at a distance f from the aperture, then the dimensionless parameter

$$\beta = \frac{2\pi RD}{\lambda f} \tag{4}$$

is very important to understanding beam shaping. In particular, suppose that we design a lens that solves the beam shaping problem in the geometrical optics limit, and now we analyze how this lens works when the wavelength is finite. We will see that the irradiance distributions of two beam shaping systems will be geometrically similar, provided only that they have the same shape functions $g(s, t)$, and $Q(s, t)$, and provided the parameters β for the two systems are the same. This means that we can transform the irradiance distribution of one system into the irradiance distribution of the other system by merely rescaling our axes. In particular, one system will suffer from diffraction effects if and only if the other system (with identical β) also does.

Geometrical optics is a short wavelength approximation, so it is clear that we would like β to be large in order for geometrical optics to hold. We will see that if β is large it is relatively simple to do beam shaping, but if it is small, the uncertainty principle of signal analysis shows that it is essentially impossible.

Another important feature in determining the difficulty of a beam shaping problem is the continuity of the beam shaping lens. If the surface of the element designed using geometrical optics is infinitely differentiable,

then we will not need a very high value of β in order to achieve good results. To be more precise, the effects of diffraction will die down like $1/\beta^2$ as β gets to be large. However, if the lens has a discontinuity in its third derivative, the effects of diffraction will die down like $1/\sqrt{\beta}$ in parts of the image plane, and hence we will need a much larger value of β in order to approach the geometrical optics limit. If the lens has discontinutities in the first or second derivatives, we will need to use even larger values of β before we can ignore the effects of diffraction.

If the input beam is smooth (such as Gaussian), then the continuity properties of the lens designed using geometrical optics are controlled by the continuity of the desired output beam. If one has a good understanding of geometrical beam shaping, it is not too difficult too see how the continuity of the desired output beam will affect the continuity of the lens. However, if one is not familiar with this theory, the results can be somewhat surprising. For example, Fig. 2 shows examples of three desired output beams. One might naively think that all of these beams have abrupt discontinuities in them, so they may all lead to equally difficult beam shaping problems. It turns out, however, that the output in Fig. 2a will lead to an infinitely differentiable lens, the beam in Fig. 2b leads to a lens with a discontinuity in the second derivative, and the lens required to produce Fig. 2c will have a discontinuity in the first derivative. The outputs 2a, 2b, and 2c get progressively harder to achieve.

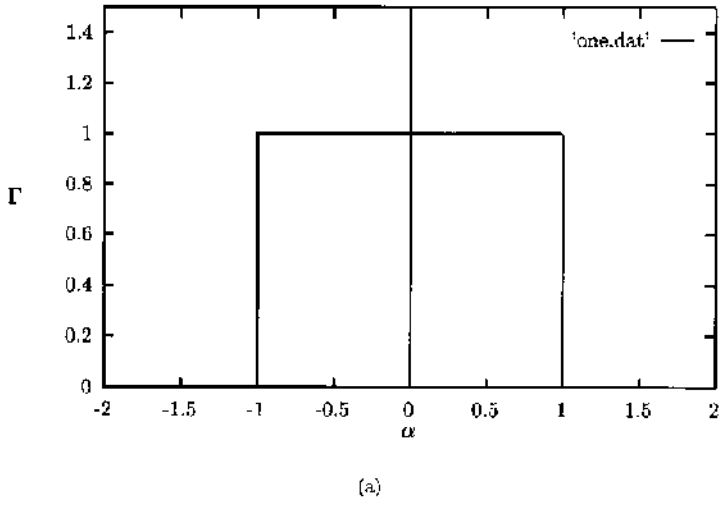
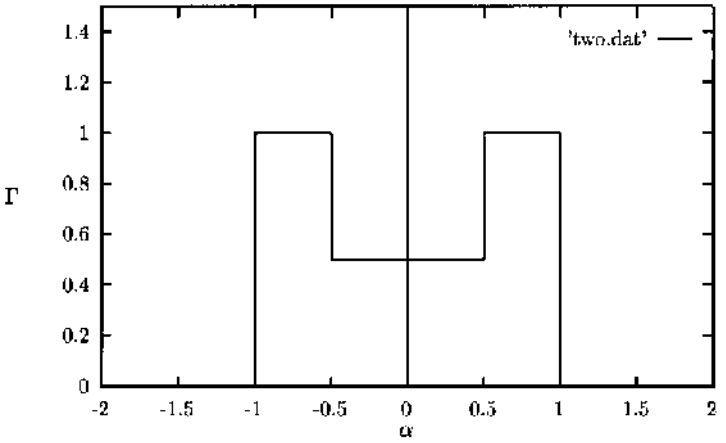
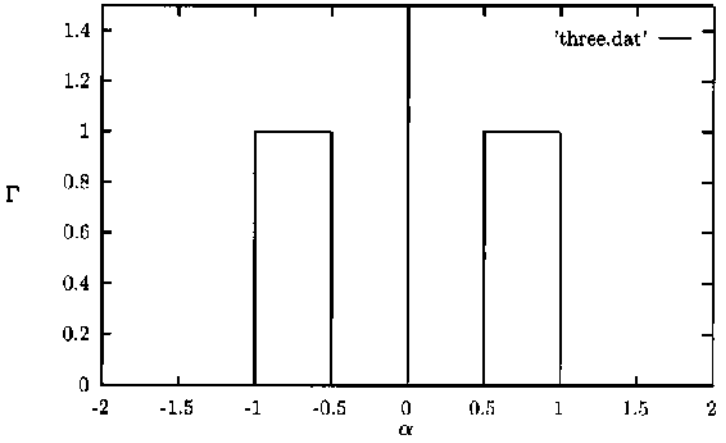


Figure 2 Three examples of desired output distributions. The outputs (a), (b), and (c) get progressively harder to achieve when diffraction effects are taken into account.



(b)



(c)

This whole chapter is devoted to understanding the points we have just discussed. We feel that it is worth writing them down as succintly as possible.

- In the geometrical optics approximation it is possible to turn a beam with a given initial distribution into a beam with any desired output distribution, provided only that the total energy of the input and output beams are the same.

- Diffraction effects make it impossible to do beam shaping exactly when we take into account the finite wavelength of light. For given shapes of the input and output beams, the parameter $\beta = 2\pi RD/\lambda f$ determines the difficulty of the beam shaping problem. If β is large, then the laws of geometrical optics will be a good approximation.
- If the surface of the element designed using geometrical optics has discontinuities in its first, second, or third derivatives, then we will need higher values of β in order for geometrical optics to be a good approximation.

In Sec. II we discuss some mathematical prerequisites for understanding the theory of beam shaping. After a brief summary of the basics of Fourier transforms, we prove the uncertainty theorem from signal analysis. In Sec. VII) this theorem will be used to show why it is impossible to do a good job of beam shaping when β is small. Section II also includes a discussion of how to use the Hankel transform in order to obtain radially symmetric Fourier transforms. This is important when analyzing the effects of diffraction on radially symmetric problems.

In Sec. III we outline the theory of stationary phase, with an emphasis on how discontinuities in the higher derivatives of the phase function can slow down the convergence. In Sec. VIII we use the method of stationary phase in order to obtain the large β approximation to the diffractive theory of beam shaping. We will see that the first term in the stationary phase approximation is equivalent to the geometrical optics approximation. We also use the method of stationary phase in Sec. V in order to analyze the errors introduced by making the Fresnel approximation.

Sections IV–VI discuss the electromagnetic theory necessary to understand beam shaping. Section IV presents a review of Maxwell's equations, Sec. V discusses the geometrical optics limit with an emphasis on Fermat's principle, and Sec. VI discusses the theory of Fresnel diffraction. Fresnel diffraction theory allows us to turn the physical problem of beam shaping into a mathematical problem involving Fourier transforms.

In Secs. VII and VIII we bring all of our tools together and discuss the theory of beam shaping. Section VII gives the theory of beam shaping in the geometrical optics limit, and Sec. VIII discusses the theory of beam shaping with diffraction effects taken into account. When the diffractive equations for beam shaping are written in dimensionless form, the importance of the parameter β will become evident. We will use the method of stationary phase to analyze the large β limit of the equations. The fact that our geometrical optics solution is based on a stationarity condition (Fermat's principle), and our large β approximation is also based on a stationarity

condition (stationary phase), causes these two analyses to look almost identical. We end Sec. VIII by giving some examples that illustrate the principles concerning the importance of β, and the smoothness of the shape of the lens.

II. MATHEMATICAL PRELIMINARIES

A. Basic Fourier Analysis

The theory of Fresnel diffraction will allow us to write our beam shaping problem as a problem in Fourier analysis. For this reason it is impossible to understand our theoretical treatment of beam shaping if one is not familiar with some of the basic concepts from Fourier analysis. In later sections we will use both one- and two-dimensional Fourier analysis.

There are several definitions of the Fourier transform used in the literature. The differences are very minor, concerning only the sign of the complex exponential, and the constant in front of the integral. However, these differences can be annoying when one is using a table of Fourier transforms, or applying theorems such as Parseval's equalilty. The definition we use here is probably the most commonly used (3, 4).

Definition 1 The Fourier transform of a function $f(x)$ is defined as

$$F(\omega) = Tf(x) = \int_{-\infty}^{\infty} f(x)e^{-i\omega x}\,dx. \tag{5}$$

An almost identical definition holds for two-dimensional functions.

Definition 2 The Fourier transform of a function $f(x, y)$ is defined as

$$F(\omega_x, \omega_y) = Tf(x, y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)e^{-i(\omega_x x+\omega_y y)}\,dx\,dy. \tag{6}$$

The following are some well-known theorems in Fourier analysis that we will use throughout this chapter.

Theorem 1 *One-dimensional Fourier Inversion Theorem—If $F(\omega)$ is the Fourier transform of $f(x)$, then*

$$f(x) = T^{-1}(F(\omega)) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{i\omega x}\,d\omega. \tag{7}$$

Theorem 2 *Two-dimensional Fourier Inversion Theorem—If $F(\omega_x, \omega_y)$ is the Fourier transform of $f(x, y)$, then*

$$f(x, y) = T^{-1}(F(\omega_x, \omega_y)) = \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(\omega_x, \omega_y) e^{i(\omega_x x + \omega_y y)} \, d\omega_x \, d\omega_y. \tag{8}$$

Theorem 3 *One-dimensional Parseval's Equality—A function $f(x)$ and its Fourier transform $F(\omega)$ satisfy*

$$\int_{-\infty}^{\infty} |F(\omega)|^2 \, d\omega = 2\pi \int_{-\infty}^{\infty} |f(x)|^2 \, dx. \tag{9}$$

Theorem 4 *Two-dimensional Parseval's Equality—A function $f(x, y)$ and its Fourier transform $F(\omega_x, \omega_y)$ satisfy*

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |F(\omega_x, \omega_y)|^2 \, d\omega_x \, d\omega_y = 4\pi^2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |f(x, y)|^2 \, dx \, dy. \tag{10}$$

Theorem 5 *The One-dimensional Fourier Convolution Theorem—Suppose $F(\omega)$ and $G(\omega)$ are the Fourier transforms of the functions $f(x)$ and $g(x)$. The inverse Fourier transform of $F(\omega)G(\omega)$ is given by*

$$T^{-1}(F(\omega)G(\omega)) = \int_{-\infty}^{\infty} f(\xi) g(x - \xi) \, d\xi. \tag{11}$$

Theorem 6 *The Two-dimensional Fourier Convolution Theorem—Suppose $F(\omega_x, \omega_y)$ and $G(\omega_x, \omega_y)$ are the Fourier transforms of the functions $f(x, y)$ and $g(x, y)$. The inverse Fourier transform of $F(\omega_x, \omega_y)G(\omega_x, \omega_y)$ is given by*

$$T^{-1}(F(\omega_x, \omega_y)G(\omega_x, \omega_y)) = \int_{-\infty}^{\infty} f(\xi, \eta) g(x - \xi, y - \eta) \, d\xi \, d\eta. \tag{12}$$

Theorem 7 *Transforms of derivatives—The Fourier transform of the derivative is given by*

$$T\left(\frac{df}{dx}\right) = i\omega F(\omega). \tag{13}$$

Theorem 8 *Transforms of partial derivatives—The Fourier transform of the partial derivatives are given by*

$$T\left(\frac{\partial f}{\partial x}\right) = i\omega_x F(\omega_x, \omega_y) \tag{14}$$

and

$$T\left(\frac{\partial f}{\partial y}\right) = i\omega_y F(\omega_x, \omega_y). \tag{15}$$

Although the Cauchy–Schwartz inequality is not really a theorem in Fourier analysis, we will need it in our proof of the uncertainty principle, and hence now state it.

Theorem 9 *The Cauchy–Schwartz Inequality (for infinite integrals)—For any function $f(x)$ and $g(x)$ we must have*

$$\left|\int_{-\infty}^{\infty} f(x)\bar{g}(x)\,dx\right|^2 \leq \int_{-\infty}^{\infty} |f(x)|^2\,dx \int_{-\infty}^{\infty} |g(x)|^2\,dx. \tag{16}$$

The two sides are equal if and only if there is a constant λ such that $f(x) = \lambda g(x)$.

B. The Uncertainty Principle and the Space Bandwidth Product

In this section we discuss the space bandwidth product, and the uncertainty principle of signal analysis (5). This discussion is crucial to understanding the theory of beam shaping. As we shall see in later sections, in a beam shaping system, the space bandwidth product is related to the parameter β discussed in the introduction. In Sec. VIII we will use the the uncertainty principle to show that it is impossible to do a good job of beam shaping if β is small.

The Heisenberg uncertainty principle of quantum mechanics (6) states that the product of the uncertainty in position times the uncertainty in momentum must be greater than $h/2\pi$:

$$\Delta p\,\Delta x > \frac{h}{2\pi}. \tag{17}$$

In order to make this precise we must define precisely what we mean by Δx and Δp. This principle was one of Heisenberg's basic assumptions in his development of matrix mechanics. However, it can also be derived by assuming the wave mechanics of Schroedinger. The derivation of the result depends on the fact that the wave function for momentum is the Fourier transform of the wave function for position, and on the subject of this section, the uncertainty principle from Fourier analysis.

All of our derivations will be limited to one-dimensional functions and their transforms, but almost identical derivations apply for two-dimensional

transforms. Once we have derived the one-dimensional results, we will state the two-dimensional results without proof. We now define the uncertainty in $f(x)$ and $F(\omega)$.

Definition 3 The uncertainty in $f(x)$ and its transform $F(\omega)$ are given by

$$\Delta_f = \sqrt{\frac{\int_{-\infty}^{\infty} x^2 |f(x)|^2 \, dx}{\int_{-\infty}^{\infty} |f(x)|^2 \, dx}}, \tag{18}$$

and

$$\Delta_F = \sqrt{\frac{\int_{-\infty}^{\infty} \omega^2 |F(\omega)|^2 \, d\omega}{\int_{-\infty}^{\infty} |F(\omega)|^2 \, d\omega}}. \tag{19}$$

The uncertainty principle concerns the product of these two quantities, and is simply related to the space bandwidth product.

Definition 4 The space bandwith product of a function $f(x)$ is defined as

$$\text{space bandwith product} = \Delta_f \, \Delta_F. \tag{20}$$

It should be noted that the space bandwidth product of a function does not depend on the scaling of the function.

Lemma 1 *For any nonzero constant a, and nonzero real number b, the space bandwidth product of* $af(bx)$ *is the same as the space bandwidth product of* $f(x)$.

We are now ready to state the uncertainty principle of signal analysis.

Theorem 10 *The One-dimensional Uncertainty Principle—For any square integrable function* $f(x)$ *the space bandwidth product must be greater than* $1/2$. *In other words,*

$$\Delta_f \, \Delta_F \geq 1/2. \tag{21}$$

Proof: The Cauchy–Schwartz inequality implies that

$$\left| \int_{-\infty}^{\infty} (x\bar{f}) \frac{df}{dx}\, dx \right|^2 \leq \int_{-\infty}^{\infty} x^2 |f(x)|^2\, dx \int_{-\infty}^{\infty} \left| \frac{df}{dx} \right|^2 dx. \tag{22}$$

Clearly

$$\left| \int_{-\infty}^{\infty} (x\bar{f}) \frac{df}{dx}\, dx \right|^2 \geq \left| \mathrm{Re} \int_{-\infty}^{\infty} (x\bar{f}) \frac{df}{dx}\, dx \right|^2. \tag{23}$$

We can write

$$\mathrm{Re} \int_{-\infty}^{\infty} x\bar{f} \frac{df}{dx}\, dx = \frac{1}{2} \int_{-\infty}^{\infty} x \left(f \frac{d\bar{f}}{dx} + \bar{f} \frac{df}{dx} \right) dx = -\frac{1}{2} \int_{-\infty}^{\infty} |f(x)|^2\, dx. \tag{24}$$

The inequalities (22) and (23) now imply

$$\frac{1}{4} \left| \int_{-\infty}^{\infty} |f(x)|^2\, dx \right|^2 \leq \int_{-\infty}^{\infty} x^2 |f(x)|^2\, dx \int_{-\infty}^{\infty} \left| \frac{df}{dx} \right|^2 dx. \tag{25}$$

Since the Fourier transform of df/dx is $i\omega F(\omega)$, Parseval's equality implies that

$$\int_{-\infty}^{\infty} \left| \frac{df}{dx} \right|^2 dx = \frac{1}{2\pi} \int_{-\infty}^{\infty} \omega^2 |F(\omega)|^2\, d\omega. \tag{26}$$

The inequality (25) can now be written as

$$\frac{1}{4} \left| \int_{-\infty}^{\infty} |f(x)|^2\, dx \right|^2 \leq \frac{1}{2\pi} \int_{-\infty}^{\infty} x^2 |f(x)|^2\, dx \int_{-\infty}^{\infty} \omega^2 |F(\omega)|^2\, d\omega. \tag{27}$$

Using Parseval's equality we can write this as

$$\frac{1}{4} \frac{1}{2\pi} \int_{-\infty}^{\infty} |F(\omega)|^2\, d\omega \int_{-\infty}^{\infty} |f(x)|^2\, dx \leq \frac{1}{2\pi} \int_{-\infty}^{\infty} x^2 |f(x)|^2\, dx \int_{-\infty}^{\infty} \omega^2 |F(\omega)|^2\, d\omega. \tag{28}$$

If we now divide both sides of this inequality by the left-hand side we arrive at the desired result. QED

Lemma 2 *We have* $\Delta_f \Delta_F = 1/2$ *iff the function* $f(x)$ *is a real Gaussian,* $f(x) = Ae^{-\alpha x^2}$ *where* α *is a real number.*

Proof: In order to get an equality in the uncertainty relation we must have an equality in the Cauchy Schwartz inequality in (22). This implies that

$df/dx = -2x\lambda f$, and hence $f(x) = Ae^{-\lambda x^2}$. It is also necessary that we get an equality in (23). This will be the case iff $\bar{f}(df/dx)$ is real, which will be true iff λ is real. QED

Although the space bandwidth product can never be less than 1/2, there is no limitation to how big it can be. For example, the function $f(x) = e^{ix^2}$ has an infinite space bandwidth product.

Suppose we change the phase of the function $f(x)$ by multiplying it by the phase function $e^{iq(x)}$. How should we choose the phase q so that the function $f(x)e^{iq(x)}$ has a minimum space bandwidth product? Note that the phase function does not change the uncertainty in x, but it does change the uncertainty in ω. This question has implications for the depth of field of a laser beam shaping system. The following theorem gives a very simple answer to this question.

Theorem 11 *The function $q(x)$ that minimizes the space bandwidth product of $f(x)e^{iq(x)}$ is the one that makes the phase of $f(x)e^{iq(x)}$ constant.*

Proof: The only integral in the space bandwidth product that changes with the function $q(x)$ is the integral

$$\int_{-\infty}^{\infty} \omega^2 |G(\omega)|^2 \, d\omega, \tag{29}$$

where $G(\omega)$ is the Fourier transform of $f(x)e^{iq(x)}$. Let $f(x)e^{iq(x)} = A(x)e^{i\psi(x)}$, where $A(x)$ is a positive real function. Parseval's equality, and the formula for the Fourier transform of a derivative, show that

$$\int_{-\infty}^{\infty} \omega^2 |G(\omega)|^2 \, d\omega = 2\pi \int_{-\infty}^{\infty} \left| \frac{d}{dx} \left(A(x)e^{i\psi(x)} \right) \right|^2 dx. \tag{30}$$

This last integral can be written as

$$2\pi \int_{-\infty}^{\infty} \left(\left(\frac{dA}{dx} \right)^2 + A^2(x) \left(\frac{d\psi}{dx} \right)^2 \right) dx \geq 2\pi \int_{-\infty}^{\infty} \left(\frac{dA}{dx} \right)^2 dx. \tag{31}$$

This clearly implies that this integral, and hence the space bandwidth product is minimized by choosing the function ψ so that it is constant. QED

This theorem will be used in Chapter 3 (Sec. III) of this book when discussing the collimation of beams.

We now summarize how these results apply for two-dimensional functions. In two dimensions, the uncertainty will be defined as

$$(\Delta_f)^2 = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x^2+y^2)|f(x,y)|^2\,dx\,dy}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|f(x,y)|^2\,dx\,dy}, \tag{32}$$

$$(\Delta_F)^2 = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(\omega_x^2+\omega_y^2)|F(\omega_x,\omega_y)|^2\,d\omega_x\,d\omega_y}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|F(\omega_x,\omega_y)|^2\,d\omega_x\,d\omega_y}. \tag{33}$$

The space bandwidth product is once again defined as $\Delta_f\,\Delta_F$. The two-dimensional uncertainty principle gives

Theorem 12 *The Two-dimensional Uncertainty Principle—For any square integrable function $f(x,y)$ the space bandwidth product must be greater than* 1. *In other words,*

$$\Delta_f\,\Delta_F \geq 1. \tag{34}$$

C. Separation of Variables in Cylindrical Coordinates

When you take the Fourier transform of a function $f(x,y)$ that has radial symmetry, you end up with a Fourier transform $F(\omega_x,\omega_y)$ that has radial symmetry in the Fourier domain. That is, if we can write

$$f(x,y) = g(r), \tag{35}$$

where $r = \sqrt{x^2+y^2}$, then we can write

$$F(\omega_x,\omega_y) = G(\alpha), \tag{36}$$

where $\alpha = \sqrt{\omega_x^2+\omega_y^2}$. The transformation that takes the function $g(r)$ into the function $G(\alpha)$ is known as a Hankel transform (4). This transform allows us to find the two dimensional Fourier transform of a radially symmetric function by performing a one-dimensional integral. The Hankel transform can be very useful when analyzing diffraction effects in beam shaping problems with radial symmetry.

In order to understand Hankel transforms it is necessary to be familiar with an identity in the theory of Bessel functions. In order to understand this identity we begin by considering the reduced wave equation in polar coordinates:

$$\nabla^2 p + k^2 p = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial p}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 p}{\partial \phi^2} + k^2 p = 0. \tag{37}$$

If we assume solutions of the form

$$p(r, \phi) = f(r)e^{im\phi}, \tag{38}$$

we find that the function $f(r)$ must satisfy

$$\frac{1}{r}\frac{d}{dr}\left(r\frac{df}{dr}\right) - \frac{m^2 f}{r^2} + k^2 f = 0. \tag{39}$$

If $g(r)$ is a solution to

$$\frac{1}{r}\frac{d}{dr}\left(r\frac{dg}{dr}\right) - \frac{m^2 g}{r^2} + g = 0, \tag{40}$$

then $f(r) = g(kr)$ is a solution to Eq. (39).

The equation (40) is known as Bessel's equation. The solutions that are regular at $r = 0$ are called Bessel functions. They are written as $J_m(r)$. If we were interested in the waves emitted from a circular cylinder, we would not require that the solution was finite at $r = 0$, but that as $r \to \infty$ the solution represented only outgoing waves. In this case we would use the solution to Bessel's equation $H_m^1(kr)$. This is known as the Hankel function of the first kind. Our goal is to understand the Hankel transform as a circularly symmetric Fourier transform. For this purpose we only need the regular solutions to Bessel's equation, which means we only need to consider the function $J_n(kr)$ where n is an integer.

One of the most elegant ways of approaching the theory of Bessel functions (7) is through the use of an integral identity, which we will now derive. This identity allows us to derive almost all of the most commonly known properties of Bessel functions such as their asymptotic behavior for large indices, asymptotic behavior for large argument, recursion formulas, and the behavior near the origin. This identity is almost the only property of Bessel functions that we will need in order to understand the Hankel transform.

The integral identity can be derived by considering the function

$$F(x, y) = e^{ix}. \tag{41}$$

This clearly satisfies the two-dimensional reduced wave equation

$$\nabla^2 F + F = 0. \tag{42}$$

We can express this in terms of polar coordinates, and then expand the function in a Fourier series. If we do this we find that

$$e^{ir\cos(\theta)} = \sum_{k=-\infty}^{\infty} a_k(r)e^{ik\theta}. \tag{43}$$

From our discussion at the begining of the section we know that this last infinite sum will satisfy the reduced wave equation if the functions $a_k(r)$ satisfy Bessel's equation. Due to the rotational symmetry of the reduced wave equation, it can be shown that in order for this infinite sum to satisfy the reduced wave equation it is necessary that each individual term satisfy the reduced wave equation. This means that it is necessary (not just sufficient) that the functions $a_k(r)$ satisfy Bessel's equation. It is also clear that they must be bounded at $r = 0$. It follows that they are multiples of the Bessel functions $J_k(r)$. We will in fact define the Bessel functions so that the multiplicative factor is unity. This gives us the result

$$e^{ir\cos(\theta)} = \sum_{k=-\infty}^{\infty} J_k(r)e^{ik\theta}. \tag{44}$$

Using the fact that the right-hand side is the Fourier expansion of the function $e^{ir\cos(\theta)}$ we arrive at the identity

$$J_k(r) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i(r\cos(\theta)-k\theta)}\,d\theta. \tag{45}$$

D. Hankel Transforms

The Fourier transform of $f(x, y)$ can be written as

$$F(\omega_x, \omega_y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-i(\omega_x x+\omega_y y)} f(x, y)\,dx\,dy. \tag{46}$$

Suppose we write both the original function $f(x, y)$ and the Fourier transform in terms of polar coordinates:

$$(x, y) = r(\cos(\theta), \sin(\theta)), \tag{47}$$

$$(\omega_x, \omega_y) = \alpha(\cos(\phi), \sin\phi)). \tag{48}$$

The Fourier transform can be written as

$$F(\alpha, \phi) = \int_0^{\infty}\int_{-\pi}^{\pi} e^{-ir\alpha\cos(\theta-\phi)} f(r, \theta)r\,dr\,d\theta. \tag{49}$$

If the function $f(x, y)$ is independent of θ, then the transform $F(\alpha, \phi)$ will be independent of ϕ. It follows that we can write

$$F(\alpha) = \int_0^{\infty}\int_{-\pi}^{\pi} e^{-ir\alpha\cos(\theta)} f(r)r\,dr\,d\theta. \tag{50}$$

If we perform the integral with respect to θ first, and use the integral representation of J_0, we get

$$F(\alpha) = 2\pi \int_0^\infty J_0(\alpha r) f(r) r \, dr. \tag{51}$$

The function $F(\alpha)$ is known as the Hankel transform of the function $f(r)$. We can apply the same steps to show that the inverse Hankel transform is given by

$$f(r) = \int_0^\infty J_0(kr) F(k) k \, dk. \tag{52}$$

III. THE METHOD OF STATIONARY PHASE

A. The Basic Idea of Stationary Phase

The method of stationary phase (8) is an asymptotic method, first used by Stokes and Kelvin, for evaluating integrals whose integrands have a very rapidly varying phase. The method is very important in the theory of dispersive wave propagation where it motivates the concept of group velocity (9, 10). In the theory of beam shaping it can be used to derive the geometrical optics limit from the theory of Fresnel diffraction, and more importantly, it gives us bounds on when the geometrical theory is applicable.

We will now give a brief heuristic derivation of the lowest-order term in the approximation. Suppose we have an integral of the form

$$H(\gamma) = \int_{-\infty}^{\infty} e^{i\gamma q(\xi)} f(\xi) \, d\xi, \tag{53}$$

and we are interested in evaluating this integral for large values of γ. Intuitively we expect that intervals where the function $\gamma q(\xi)$ is changing rapidly will give neglibible contributions to this integral. If the derivative of q vanishes at $\xi = \xi_0$, we expect the main contribution to come from the region very near ξ_0. To a first approximation we can write

$$H(\gamma) \approx f(\xi_0) e^{i\gamma q(\xi_0)} \int_{-\infty}^{\infty} e^{i\gamma q''(\xi_0)(\xi-\xi_0)^2/2} \, d\xi. \tag{54}$$

We have arrived at this expression by assuming that the major contribution comes from a small region around ξ_0, and hence we have approximated the function $f(\xi)$ as being constant, and equal to $f(\xi_0)$. We have also expanded the function $q(\xi)$ in a Taylor series about ξ_0, keeping only the

terms up to the quadratic. The integral can now be evaluated analytically to give

$$H(\gamma) \approx f(\xi_0) e^{i\mu\pi/4} e^{i\gamma q(\xi_0)} \sqrt{\frac{2\pi}{\gamma |q''(\xi_0)|}}, \tag{55}$$

where

$$\mu = \operatorname{sgn}\left(\left. \frac{d^q(\xi)}{d\xi^2} \right|_{\xi_0} \right). \tag{56}$$

Here we have assumed that there is exactly one point where the phase is stationary. If there is more than one point, then we must sum over all points that are stationary in order to get our asymptotic expansion. If there are no stationary points, then the integral will die down exponentially fast with γ provided the functions $q(\xi)$ and $f(\xi)$ are infinitely differentiable, and the function $f(\xi)$ and all of its derivatives decay as $|\xi| \to \infty$. If there is no stationary point, but the function $f(\xi)$ has a discontinuity in it, we are at least guaranteed that the integral dies down like $1/\gamma$ as $\gamma \to \infty$. It is not too difficult to make this heuristic derivation more rigorous.

B. The Rate of Convergence of the Method of Stationary Phase

In our discussion of beam shaping we will see that the lowest-order term in the stationary phase approximation to the diffraction integral gives us the geometrical optics approximation. In this case the parameter β discussed in the introduction will serve as our large parameter in the phase of the integrand. In order to understand what sorts of errors are produced when we use the geometrical optics approximation, we need to understand the higher-order terms in the method of stationary phase. It is not important for us to have exact expressions for the higher-order terms, but we need to know how fast they die down with γ.

The subject of how to correct the lowest-order term in the method of stationary phase gets somewhat technical, so we feel that it is best if we begin by summarizing the main results. In our analysis of beam shaping we will have another parameter in our phase function, so our integrals will be of the form

$$H(x, \gamma) = \int_{-\infty}^{\infty} e^{i\gamma q(\xi, x)} f(\xi)\, d\xi. \tag{57}$$

Here the parameter ξ represents a point on the aperture, and x represents a point at the focal plane. The function $q(\xi, x)$ will be proportional to the

travel time required to get from a point ξ on the aperture to a point x in physical space. In practice ξ and x will be two-dimensional vectors, but we assume they are scalars here in order to simplify the presentation. This one-dimensional case will be directly relevent for the case where our input and output beams can be written as a direct product of two one-dimensional distributions.

Let $\xi_0(x)$ be the point at which the phase is stationary. We will show that if the functions $q(\xi, x)$ and $f(\xi)$ are infinitely differentiable at the stationary point $\xi_0(x)$, and

$$\left.\frac{\partial^2 q(\xi, x)}{\partial \xi^2}\right|_{\xi_0} \neq 0, \tag{58}$$

then the next order correction dies down like $1/\gamma^{3/2}$. This gives us an expression of the form

$$H(x, \gamma) = \frac{A(x)}{\gamma^{1/2}} + \frac{iB(x)}{\gamma^{3/2}} + \cdots. \tag{59}$$

In this case the relative error between the first-order term and the exact solution will die down like $1/\gamma$. If the function $f(\xi)$ is real, then the functions $A(x)$ and $B(x)$ will have the same phase. This implies that the relative error between $|H(x, \gamma)|^2$ and the value predicted by the first term in the method of stationary phase will be $O(1/\gamma^2)$.

This expression will remain valid provided the functions $f(\xi)$ and $\partial^2 q/\partial \xi^2$ are differentiable at $\xi_0(x)$. If these functions are continuous, but not differentiable at some point x^*, then at the point x^*, the next order term in the method will be of the form

$$H(x^*, \gamma) = \frac{A(x^*)}{\gamma^{1/2}} + \frac{B(x^*)}{\gamma} + \cdots. \tag{60}$$

We see that in this case the relative error between the first order-term and the exact solution will die down like $1/\gamma^{1/2}$. Furthermore, in this case the functions $A(x)$ and $B(x)$ are not phase shifted by 90° when $f(\xi)$ is real. This means that the relative error between $|H(x, \gamma)|^2$ and the term predicted by the first term in the method of stationary phase will be $O(1/\sqrt{\gamma})$. This means that we will need a much larger value of γ before the first-order term is a good approximation. In terms of our beam shaping problem this will imply that if the surface of the beam shaping lens designed using geometrical optics has a discontinuity in the third derivative, then we will require much larger values of β in order for the results of geometrical optics to be a good approximation.

Suppose that at some point x_0, the function $f(\xi)$ or $\partial^2 q/\partial\xi^2$ is discontinuous at $\xi_0(x_0)$. Since the first-order term in the method of stationary phase requires us to know $f(\xi)$ and $\partial^2 q(\xi, x)/\partial\xi^2$ at $\xi_0(x)$, it is clear that we need to modify the results of the lowest-order term in our stationary phase approximation when $x = x_0$. More importantly, the method of stationary phase will hold for values of x near x_0, but the convergence near these points will be dramatically affected. The analysis of this situation is based on the Fresnel integral (11), and we see that this situation is related to the diffraction by an semi-infinite half plane. As with that case, we end up getting oscillations near the point x_0. For the beam shaping problem, this implies that if the surface of the lens designed using geometrical optics has a discontinuity in the second derivative, then we will get even worse convergence, and this will be accompanied by oscillations in the amplitude. When the surface of the lens has a discontinuity in the first derivative, the convergence towards the geometrical optics limit is affected even more dramatically.

Clearly, if the discontinuities are small enough, they will have little effect on the convergence towards the geometrical optics limit. For example, the elements are often manufactured by approximating the element by a piecewise constant element. This should not be any problem if the steps are small enough.

So far we have assumed that the second derivative of $q(\xi, x)$ does not vanish at the stationary point $\xi_0(x)$. In optics, points where this condition is violated are said to lie on a caustic surface. Suppose we have a point source of light whose rays get refracted by an inhomogeneous medium. It is possible that at certain points in the medium we might have more than one ray arriving from this point source, or possibly none at all. The surfaces separating regions where there are different numbers of rays are known as the caustic surfaces. When we analyze the diffraction integral using the method of stationary phase we find that on the caustic surface the second derivative of $q(\xi, x)$ vanishes.

We are not so much interested in computing the integral for $H(x, \gamma)$ right at a point where the stationary point ξ_0 has a vanishing second derivative. Instead, we are interested in analysing the integral $H(x, \gamma)$, for x near x_0 where

$$\left.\frac{\partial^2 q(\xi, x_0)}{\partial\xi^2}\right|_{\xi=\xi_0(x_0)} = 0. \tag{61}$$

In optics, the point x_0 would be a point on the caustic surface. We would find that for points on one side of x_0, there are no stationary points, and on the other side there are two stationary points. In order to understand the behavior of $H(x, \gamma)$ near such points we need to include cubic terms in the

Taylor series expansion of the phase near the stationary point, and this analysis is based on the Airy integral.

We will not give any further discussion of the Airy integral or caustics since when discussing beam shaping we do not present any examples where caustics occur in the classical sense of the word. All of the problems we analyze lead to lenses whose phase functions do not have inflection points. However, in some of the lenses, the phase function grows linearly as we move far away from the center of aperture. This results in a situation where the caustic occurs at a value of $\xi_0 = \infty$.

C. A Preliminary Transformation

In our analysis of the higher-order terms in the method of stationary phase we will begin by analyzing the situation where $q(\xi) = \xi^2$. This leads us to consider integrals of the form

$$P(\gamma) = \int_{-\infty}^{\infty} f(\xi) e^{i\gamma\xi^2}\, d\xi. \tag{62}$$

By making a preliminary transformation, we can transform the analysis of the integral in Eq. (53):

$$H(\gamma) = \int_{-\infty}^{\infty} e^{i\gamma q(\xi)} f(\xi)\, d\xi, \tag{63}$$

into the analysis of this simpler problem. In order to do this we assume that $q(\xi)$ has a single stationary point at ξ_0. In this case we can introduce a new variable s such that

$$s^2 = \mu(q(\xi) - q(\xi_0)), \tag{64}$$

$$\mu = \operatorname{sgn}\left(\frac{d^2 q(\xi_0)}{d\xi}\right). \tag{65}$$

in the neighborhood of ξ_0. Making the change of variables $s(\xi) = \sqrt{\mu(q(\xi) - q(\xi_0))}$, we end up with an integral of the form

$$H(\gamma) = e^{i\gamma q(\xi_0)} \int_{-\infty}^{\infty} e^{i\mu\gamma s^2} f(\xi(s)) \frac{d\xi}{ds}\, ds. \tag{66}$$

This gives us an integral of the same form as (62), but with the function

$$g(s) = f(\xi(s)) \frac{d\xi}{ds} \tag{67}$$

replacing the function $f(\xi)$.

When we apply the method of stationary phase to the integral (62) we see that there is a stationary point at $\xi = 0$. The continuity properties of $f(\xi)$ are important in determining how quickly the first-order term in the method of stationary phase converges towards the exact answer. For the general case it is important to know the continuity properties of the function $g(s)$. A discontinuity in the *kth* derivative of $g(s)$ can arise by the *kth* derivative of $f(\xi)$ being discontinuous at $\xi = \xi_0$, or by the *kth* derivative of $d\xi/ds$ being discontinuous at $s = 0$. The derivatives of $\xi(s)$ depend on the derivatives with respect to ξ of $q(\xi)$ at ξ_0. These derivatives can be calculated using implicit differentiation. In particular, note that

$$2s \frac{ds}{d\xi} = \mu \frac{dq}{d\xi}. \tag{68}$$

If we evaluate this at $\xi = \xi_0$ we find that both sides of this equation vanish, and we have not determined any derivatives. However, if we differentiate once more with respect to ξ we get

$$2\left(\frac{ds}{d\xi}\right)^2 + 2s \frac{d^2s}{d\xi^2} = \mu \frac{d^2q}{d\xi^2}, \tag{69}$$

and when we evaluate this at $\xi = \xi_0$ we get

$$2\left(\frac{ds(\xi_0)}{d\xi}\right)^2 = \mu \frac{d^2q(\xi_0)}{d\xi^2}. \tag{70}$$

This gives us two possible values of $ds(\xi_0)/d\xi$. We can choose either sign we want to. When we take further derivatives we find that the $(d^k/d\xi^k)s(\xi_0)$ is determined by the derivatives of $q(\xi)$ up to $k+1$. This means that the $(d^k/ds^k)\xi(0)$ is also determined by these same derivatives. Finally we see that the derivatives of $g(s)$ up to k will be continuous only if the derivatives of q up to $k+2$ are continuous. In particular, we see that the function $g(s)$ will have a continuous first derivative if and only if the derivative of $f(\xi)$ and the third derivative of $q(\xi)$ are both continuous.

D. Generalized Functions

Before discussing the higher-order terms in the method of stationary phase we consider some relevant concepts from the theory of generalized functions (12). The Dirac delta function and its derivatives are examples of generalized functions. The definition of these functions often arise as infinite integrals whose integrands do not decay at infinity. For example the delta function is

the inverse Fourier transform of a constant, and hence is defined by a divergent integral.

$$\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i\omega x}\, d\omega. \tag{71}$$

One way of thinking of this function is to imagine that it is defined by taking the inverse Fourier transform of $e^{-\alpha\omega^2}$, and then letting $\alpha \to 0$. The function that we get by doing this is the Dirac delta function. Even though it is a rather unusual function, it is extremely useful in practice.

In finding the higher-order terms for the method of stationary phase it will be useful to consider the integrals

$$R_k(\gamma) = \int_{-\infty}^{\infty} e^{i\gamma\xi^2}\xi^k\, d\xi. \tag{72}$$

These integrals can be confusing since if γ is real then the integrands of these integrals do not approach zero as $\xi \to \infty$. However, if we evaluate these integrals over a finite interval, and let the region of integration go to infinity, we find that these are in fact convergent integrals. Furthermore, if we give γ a very small positive imaginary part, then the integrands approach zero. After evaluating these integrals we could then let the imaginary part go to zero. When we let the imaginary part go to zero we find that all of the integrals R_k are well defined. We could also get the integrals by taking the derivatives of the integral $R_0(\gamma)$ with respect to γ. If we do this we find that

$$R_{k+2} = -i\,\frac{dR_k}{d\gamma}. \tag{73}$$

Due to the asymmetry of the integrand we get

$$R_k = 0 \text{ for } k \text{ odd}. \tag{74}$$

Carrying out this process we find that the first few of these integrals are given by

$$R_0(\gamma) = \frac{C_0}{\gamma^{1/2}}, \tag{75}$$

$$R_1(\gamma) = 0, \tag{76}$$

$$R_2(\gamma) = \frac{iC_0}{2\gamma^{3/2}}, \tag{77}$$

where

$$C_0 = e^{i\pi/4}\sqrt{\pi}. \tag{78}$$

We will also be concerned with the integrals

$$S_k(\gamma) = \int_0^\infty \xi^k e^{i\gamma\xi^2}\, d\xi. \tag{79}$$

We can use the same sorts of reasoning on these integrals. If k is even then

$$S_k(\gamma) = \frac{1}{2} R_k(\gamma) \text{ for } k \text{ even}. \tag{80}$$

However, unlike R_k these integrals do not vanish when k is odd. In particular, when $k = 1$

$$S_1(\gamma) = \frac{i}{2\gamma}. \tag{81}$$

The rest of the integrals can be evaluated using

$$S_{k+2}(\gamma) = -i \frac{d}{d\gamma} S_k(\gamma). \tag{82}$$

E. Higher Order Terms in the Method of Stationary Phase

We begin our analysis of the higher-order terms in the method of stationary phase by considering the special case

$$H(\gamma) = \int_{-\infty}^\infty e^{i\gamma\xi^2} f(\xi)\, d\xi. \tag{83}$$

This has the stationary point at $\xi = 0$. To obtain the first term in the method of stationary phase we argued that the major contribution to this integral came from the region around $\xi = 0$. For this reason we expanded $f(\xi)$ in a Taylor series about $\xi = 0$, and then kept only the first term in the series. It makes sense that we should get more accurate answers if we keep more terms in the Taylor series. For example, if we kept three terms in the Taylor series this would lead to an approximation of the form

$$H(\gamma) \approx f(0) R_0(\gamma) + \frac{df(0)}{d\xi} R_1(\gamma) + \frac{1}{2} \frac{d^2 f(0)}{d\xi^2} R_2(\gamma) + \cdots \tag{84}$$

where $R_k(\gamma)$ are the integrals that we discussed in the previous section. We conclude that the higher-order approximations for $H(\gamma)$ can be written as

$$H(\gamma) \approx e^{i\pi/4} \sqrt{\frac{\pi}{\gamma}} \left(f(0) + \frac{i}{4\gamma} \frac{d^2 f(0)}{d\xi} \right) + \cdots \tag{85}$$

In this special case, this shows that the next order term in the method of stationary phase dies down like $1/\gamma^{3/2}$ provided $f(\xi)$ is sufficiently differentiable.

If the derivative of $f(\xi)$ has a discontinuity at $\xi = 0$, then if we keep two terms in our Taylor series about $\xi = 0$, we end up with an expression

$$H(\gamma) \approx f(0)R_0(\gamma) + \frac{df(0_+)}{d\xi}\int_0^\infty e^{i\gamma\xi^2}\xi\, d\xi + \frac{df(0_-)}{d\xi}\int_{-\infty}^0 e^{i\gamma\xi^2}\xi\, d\xi + \cdots \tag{86}$$

We can write this as

$$H(x,\gamma) \approx f(0)R_0(\gamma) + S_1(\gamma)\left(\frac{df(0_+)}{d\xi} - \frac{df(0_-)}{d\xi}\right) + \cdots \tag{87}$$

Using our values of R_0 and S_1 we get

$$H(\gamma) \approx f(0)e^{i\pi/4}\sqrt{\pi/\gamma} + \frac{i}{2\gamma}\left(\frac{df(0_+)}{d\xi} - \frac{df(0_-)}{d\xi}\right) + \cdots \tag{88}$$

We see that if $f(\xi)$ has a discontinuous derivative at $\xi = 0$, then the relative error between the first-order term and the exact answer will die down like $1/\gamma^{1/2}$. This is much slower than when the derivative of $f(\xi)$ is continuous.

As we noted earlier, the general case where $q(\xi)$ is not quadratic can be transformed into the quadratic case, but replacing the function $f(\xi)$ by the function $g(s)$ in Eq. (67). We saw that the function $g(s)$ will have a discontinuous derivative if the function $f(\xi)$ has a discontinous derivative, or if $q(\xi)$ has a discontinuous third derivative. It follows that as long as the first derivative of $f(\xi)$ and the third derivative of $q(\xi)$ are continuous, then the next order term in the method of stationary phase will die down like $1/\gamma^{3/2}$. If either of these derivatives are discontinuous, then the next order term will die down like $1/\gamma$. This can be used to justify our earlier statement concerning the effect of a discontinuity in the third derivative of the lens surface on the rate of convergence towards the geometrical optics limit.

F. Lower-Order Discontinuities in the Phase Functions

When the functions $f(\xi)$ or $(d^2/d\xi^2)q(\xi, x)$ are discontinuous, we can get very slow convergence from the first term in the stationary phase approximation. We will begin with a simple example illustrating this point. By suitably changing coordinates, more general problems can in fact be related to this simple example.

Consider the integral

$$V(x,\gamma) = \frac{1}{\sqrt{\pi}} \int_0^\infty e^{i\gamma(\xi - x)^2}\, d\xi. \tag{89}$$

This is a special case of our general problem where $q(\xi, x)$ has a quadratic dependence on ξ and where $f(\xi) = 1$ for $\xi > 0$ and 0 for $\xi < 0$. This is an example where the function $f(\xi)$ is discontinuous.

If we apply the method of stationary phase to this integral we see that there is no stationary point for $x < 0$, and that the method predicts that the integral is independent of x for $x > 0$. More specifically the method predicts

$$V(x,\gamma) \approx 0 \text{ for } x < 0, \tag{90}$$

$$V(x,\gamma) \approx \sqrt{\frac{1}{\gamma}}\, e^{i\pi/4} \text{ for } x > 0. \tag{91}$$

In the stationary phase approximation, the magnitude of $V(x,\gamma)$ is a multiple of the Heavyside function.

$$|V(x,\gamma)|^2 \approx 0 \text{ for } x < 0, \tag{92}$$

$$|V(x,\gamma)|^2 \approx \frac{1}{\gamma} \text{ for } x > 0. \tag{93}$$

We now consider this integral in more detail. A simple change of variables allows us to write

$$V(x,\gamma) = \frac{1}{\gamma^{1/2}} \operatorname{Fr}(-x\gamma^{1/2}), \tag{94}$$

where

$$\operatorname{Fr}(s) = \frac{1}{\sqrt{\pi}} \int_s^\infty e^{it^2}\, dt. \tag{95}$$

The function $\operatorname{Fr}(s)$ is known as a complex Fresnel integral. Figure 3 shows a plot of the intensity $|\operatorname{Fr}(s)|^2$ of the function $\operatorname{Fr}(s)$. This graph shows that for $x < 0$ the function $V(x,\gamma)$ has a monotonic decay towards zero while for $x > 0$ we get an oscillatory approach towards the constant value of 1. We see that if $|x\sqrt{\gamma}| \gg 1$, then $V(x,\gamma)$ will agree very well with the stationary phase solution. The difference between the exact solution and the stationary phase solution is that the stationary phase solution approximates the lower limit of the integrand as being equal to $-\infty$. Since $\gamma \gg 1$, this is usually a good approximation, but when x is close to zero, this is not so good. This is the root of the slow convergence of the method of stationary phase for all problems that have a discontinuity in $f(\xi)$ or $d^2q/d\xi^2$.

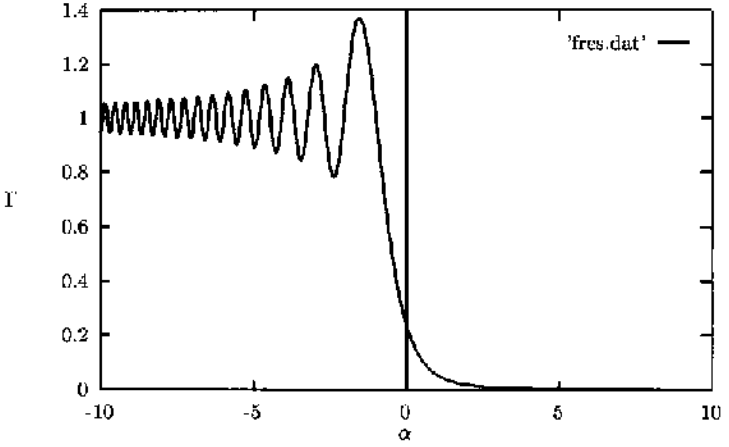


Figure 3 A plot of the function $|\mathrm{Fr}(x)|^2$ that models the intensity when light gets diffracted by a plane.

We can understand the behavior of $\mathrm{Fr}(s)$ by considering the behavior of $\mathrm{Fr}(s)$ for large values of $|s|$. For $s \gg 0$ we can integrate by parts to show that

$$\mathrm{Fr}(s) = \frac{1}{\sqrt{\pi}}\, e^{is^2}\, \frac{i}{2s} + \frac{1}{\sqrt{\pi}} \int_s^\infty \frac{e^{it^2}}{2it}\, dt. \tag{96}$$

This shows that

$$\mathrm{Fr}(s) = \frac{1}{\sqrt{\pi}}\, e^{is^2}\, \frac{i}{2s} + O(1/s^2). \tag{97}$$

Similarly for $s \ll 0$ we can write

$$\mathrm{Fr}(s) = e^{i\pi/4} - \frac{1}{\sqrt{\pi}} \int_{-\infty}^s e^{it^2}\, dt. \tag{98}$$

An integration by parts now shows that

$$\mathrm{Fr}(s) = e^{i\pi/4} - \frac{1}{\sqrt{\pi}}\, e^{is^2}\, \frac{i}{2s} + O(1/z^2). \tag{99}$$

When we compute the amplitude $|\mathrm{Fr}(s)|^2$ we see that

$$|\mathrm{Fr}(s)|^2 = \frac{1}{4\pi s^2} + \cdots \text{ for } s \gg 0, \tag{100}$$

and

$$|\mathrm{Fr}(s)|^2 = \left(1 - \frac{\sin(s^2 - \pi/4)}{\sqrt{\pi}s}\right) \text{ for } s \ll 0. \tag{101}$$

We see that the solution approaches its asymptotic value much slower for $s \ll 0$ than for $s \gg 0$, and that it approaches it in an oscillatory fashion. This is only the asymptotic behavior, but it gives quite an accurate picture of the function $\mathrm{Fr}(s)$.

This shows that the function $V(x, \gamma)$ will have oscillations for $x > 0$, and be smooth for $x < 0$. Note that the convergence for large values of γ will be very slow if x is near 0, since $x\sqrt{\gamma}$ will be relatively small in this case as $\gamma \to \infty$.

A very slight generalization of this problem is to consider the function

$$V(x, \gamma) = \int_0^\infty e^{i\gamma(s-x)^2}\, ds + \alpha \int_\infty^0 e^{i\gamma(s-x)^2}\, ds. \tag{102}$$

This reduces to the previous problem when $\alpha = 0$. When α is non-zero we can analyze this problem in a similar fashion. In this case we see that we will get oscillations on both sides of $x = 0$, but the oscillations will be bigger on the side where the stationary phase solutions predicts that V is bigger. Generalizing further we see that when we have an integral of the form

$$H(x, \gamma) = \int_{-\infty}^{\infty} e^{i\gamma q(x,\xi)} f(\xi)\, d\xi, \tag{103}$$

and $f(\xi)$ has a discontinuity at ξ^*, then we need to break the integral up into two parts.

$$H(x, \gamma) = \int_{-\infty}^{\xi^*} e^{i\gamma q(x,\xi)} f(\xi)\, d\xi + \int_{\xi^*}^{\infty} e^{i\gamma q(x,\xi)} f(\xi)\, d\xi. \tag{104}$$

Suppose that the function $q(x, \xi)$ has a stationary point at $\xi(x)$, and that $\xi(x^*) = \xi^*$. When we apply the reasoning behind the method of stationary phase to the integral from ξ^* to ∞ we get

$$\int_{\xi^*}^{\infty} e^{i\gamma q(x,\xi)} f(\xi)\, d\xi \approx e^{i\gamma q(x,\xi(x))} f(\xi(x)) \int_{\xi^*}^{\infty} e^{i\gamma 1/2(d^2 q(x,\xi(x))/d\xi^2)(\xi-\xi(x))^2}\, d\xi. \tag{105}$$

We now make the substitution

$$s^2 = \gamma\, \frac{1}{2}\, \frac{d^2 q(x, \xi(x))}{d\xi^2}\, (\xi - \xi(x))^2. \tag{106}$$

This gives us the approximation

$$\int_{\xi^*}^{\infty} e^{i\gamma q(x,\xi)} f(\xi)\, d\xi \approx e^{i\gamma q(x,\xi(x))} f(\xi(x)) \frac{\sqrt{\pi}}{\sqrt{\gamma/2 \left| \dfrac{d^2 q(x,\xi(x))}{d\xi^2} \right|}} \operatorname{Fr}(s^*), \tag{107}$$

where

$$s^* = (\xi^* - \xi(x)) \sqrt{\frac{d^2 q(x,\xi(x))}{d\xi^2}\, \gamma/2} \tag{108}$$

If $s^* \ll 0$, we can make the approximation

$$F(s^*) \approx \operatorname{Fr}(-\infty) = e^{i\pi/4}, \tag{109}$$

and we get back the first term in the method of stationary phase. However, if $\xi(x)$ is too close to ξ^*, this will not be a very good approximation unless γ is extremely large. However, our results should be good if we keep s^* in our expression rather replacing it by $-\infty$. This is exactly what we did in our analysis of the integral in Eq. (89).

As in our analysis of Eq. (89), we can patch up our approximation for points x such that $\xi(x) \approx \xi^*$ by using the Fresnel integral. If we were to do this in the general case, the formulas would get extremely cumbersome. However, it is clear that in this general case we will get the same qualitative behavior as in Eq. (89).

G. The Method of Stationary Phase in Higher Dimensions

The method of stationary phase carries over to higher-dimensional integrals. In particular, suppose we have an integral of the form

$$H(\gamma) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\xi, \eta) e^{i\gamma q(\xi,\eta)}\, d\xi\, d\eta. \tag{110}$$

Once again if the function $q(\xi, \eta)$ has a stationary point where

$$\nabla q(\xi_0, \eta_0) = 0, \tag{111}$$

then the major contribution to the integral will come from points right around this stationary point, and we can approximate the integral by

$$H(\gamma) \approx f(\xi_0, \eta_0) e^{i\gamma q(\xi_0,\eta_0)} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{i1/2\gamma Q(\xi,\eta)} d\xi\, d\eta, \tag{112}$$

where

$$Q(\xi,\eta) = \frac{\partial^2 q(\xi_0,\eta_0)}{\partial \xi^2}(\xi-\xi_0)^2 + 2\frac{\partial^2 q(\xi_0,\eta_0)}{\partial \xi\,\partial \eta}(\xi-\xi_0)(\eta-\eta_0) + \frac{\partial^2 q(\xi_0,\eta_0)}{\partial \eta^2}(\eta-\eta_0)^2. \quad (113)$$

This integral can be evaluated to give

$$H(\gamma) \approx \frac{2\pi i}{\gamma\sqrt{J(\xi_0,\eta_0)}}\, e^{i\gamma q(\xi_0,\eta_0)}, \quad (114)$$

where

$$J(\xi_0,\eta_0) = \frac{\partial^2 q(\xi_0,\eta_0)}{\partial \xi^2}\frac{\partial^2 q(\xi_0,\eta_0)}{\partial \eta^2} - \frac{\partial^2 q(\xi_0,\eta_0)}{\partial \xi\,\partial \eta}\frac{\partial^2 q(\xi_0,\eta_0)}{\partial \xi\,\partial \eta}. \quad (115)$$

IV. MAXWELL'S EQUATIONS

A. Maxwell's Equations

The theory of beam shaping is based on diffraction theory, which is itself based on electromagnetic field theory. For this reason we will now review basic eletromagnetic theory (13). The governing equations of eletromagnetic field theory are

$$\nabla \times \mathbf{E} + \frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} = 0, \quad (116a)$$

$$\nabla \times \mathbf{B} - \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} = \frac{4\pi}{c}\mathbf{L} \quad (116b)$$

and

$$\nabla \cdot \mathbf{E} = 4\pi\rho, \quad (117a)$$

$$\nabla \cdot \mathbf{B} = 0. \quad (117b)$$

In these equations $\mathbf{E}$ and $\mathbf{B}$ are the electric and magnetic fields, and ρ and $\mathbf{J}$ are the charge and current densities. The densities ρ and $\mathbf{J}$ are related to each other through the law of conservation of charge.

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \mathbf{J} = 0. \quad (118)$$

The first of Eqs. (116) is the differential form of Faraday's principle of electromagnetic induction. The second equation describes both Ampere's law and Maxwell's displacement current. The first of equations (117) is

known as Gauss's law, and the second equation describes the fact that there is no such thing as a magnetic monopole. It should be noted that the second set of Maxwell's equations almost follows from the first set. If we take the divergence of each of the equations in (116), use the fact that the divergence of a curl is zero, and use the law of conservation of charge, we get

$$\frac{\partial}{\partial t} \nabla \cdot \mathbf{B} = 0, \tag{119}$$

and

$$\frac{\partial}{\partial t} (\nabla \cdot \mathbf{E} - 4\pi\rho) = 0. \tag{120}$$

We see that if the second set of Maxwell's equations is true initially, then the first set requires that they be true for all time.

We are frequently concerned with wave propagation through some medium such as air, water, or glass. In this case there is an interaction between the charge distributions and the electromagnetic field. This interaction is usually taken into account by assuming that the electric field induces a polarization charge $\mathbf{P}$ such that the charge density is given by

$$\rho = -\nabla \cdot \mathbf{P}. \tag{121}$$

Assuming that there are no other charges other than those induced by the electric field this gives us the equation

$$\nabla \cdot (\mathbf{E} + 4\pi\mathbf{P}) = 0. \tag{122}$$

The assumption is typically made that the polarization is given by

$$\mathbf{P} = \chi\mathbf{E}. \tag{123}$$

Gauss's law can now be written as

$$\nabla \cdot \mathbf{D} = 0, \tag{124a}$$

where

$$\mathbf{D} = \epsilon\mathbf{E}, \tag{124b}$$

and

$$\epsilon = 1 + 4\pi\chi. \tag{124c}$$

The linearity between the polarization and the electric field is usually valid unless the electric field gets to be very large. Here we have also assumed that the medium is isotropic, so there are no preferred directions. In a non-isotropic medium, the polarization is related to the applied electric field by a symmetric second rank tensor. In order to describe the phenomonon of birefringence in crystals it is necessary to use the general tensor form

for the polarization (this tensor is diagonal for the special case of an isotropic medium). This relation between the electric field and the polarization also assumes that the polarization depends only on the local value of the electric field. Using such a theory it is not possible to describe the rotation of the polarization field by an optically active material.

A similar approximation is used to take into account the effect of currents that are produced by the magnetic field. In this case the currents in the material induce a polarization current **M** such that the current is given by

$$\mathbf{J}_M = c\nabla \times \mathbf{M}. \tag{125}$$

It follows that the magnetic field must satisfy

$$\nabla \times \mathbf{B} - \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} = 4\pi\nabla \times \mathbf{M}. \tag{126}$$

If we introduce the quantity **H** defined by

$$\mathbf{H} = \mathbf{B} - 4\pi\mathbf{M}, \tag{127}$$

then assuming the only currents are those arising from the induced current $\mathbf{J}_M$, we can write

$$\nabla \times \mathrm{H} - \frac{1}{c}\frac{\partial \mathbf{E}}{\partial t} = 0. \tag{128}$$

In the simplest case it is assumed that the fields **B** and **H** are linearly proportional to each other

$$\mathbf{B} = \mu\mathbf{H}. \tag{129}$$

For most materials that are used in optics the linear relation between the **B** and **H** fields is totally satisfactory. In fact, the constant μ is very nearly equal to unity for most materials of optical interest.

We now collect the macroscopic form of Maxwell's equations in a linear isotropic material:

$$\frac{1}{c}\frac{\partial \mathbf{D}}{\partial t} - \nabla \times \mathbf{H} = 0, \tag{130a}$$

$$\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t} + \nabla \times \mathbf{E} = 0. \tag{130b}$$

$$\nabla \cdot \mathbf{D} = 0, \tag{131a}$$

$$\nabla \cdot \mathbf{B} = 0. \tag{131b}$$

where

$$\mathbf{D} = \epsilon \mathbf{E}, \tag{132a}$$

$$\mathbf{B} = \mu \mathbf{H}. \tag{132b}$$

We have omitted any sources of charges and currents other than those produced by the interaction of the fields with the materials.

B. The Wave Equation

Our analysis of diffraction effects in Sec. VI is based on the fact that in a linear, homogeneous, and isotropic medium, each component of the electric and magnetic fields satisfies the wave equation. We now give a derivation of this fact. We begin by deriving an equation for $\mathbf{E}$ that does not assume that ϵ and μ are constants.

To begin with we write the second of Eqs. (130) as

$$\frac{1}{c}\frac{\partial \mathbf{H}}{\partial t} + \frac{1}{\mu}\nabla \times \mathbf{E} = 0. \tag{133}$$

We now take the curl of this equation, and use the first of Eqs. (130) to arrive at the result

$$\frac{\partial^2 \mathbf{E}}{\partial t^2} = -c^2 \frac{1}{\epsilon}\nabla \times \left(\frac{1}{\mu}\nabla \times \mathbf{E}\right). \tag{134}$$

This is the form of the wave equation for $\mathbf{E}$ in a medium where ϵ and μ are not assumed to be constant. If we assume that μ is constant, we can write this equation as

$$\frac{\partial^2 \mathbf{E}}{\partial t} = -c^2 \frac{1}{\epsilon\mu}\nabla \times \nabla \times \mathbf{E}. \tag{135}$$

We can simplify this equation by using the identity $\nabla \times \nabla \times \mathbf{A} = -\nabla^2 \mathbf{A} + \nabla\nabla \cdot \mathbf{A}$, along with the fact that $\nabla \cdot \mathbf{E} = 0$ (assuming that ϵ is constant). For a homogeneous medium this gives us the equation

$$\frac{\partial^2 \mathbf{E}}{\partial t^2} = \frac{c^2}{\epsilon\mu}\nabla^2 \mathbf{E}. \tag{136}$$

This shows that each component of the electric field satisfies the wave equation. If the fields are time harmonic, with frequency ω, the spatial dependence of the electric field must satisfy

$$k^2 \mathbf{E} + \nabla^2 \mathbf{E} = 0, \tag{137}$$

where

$$k = \frac{\omega\epsilon\mu}{c} \tag{138}$$

We refer to this equation as the reduced wave equation, or the Helmholtz equation.

Similar arguements show that the field **H** satisfies

$$\frac{\partial^2 \mathbf{H}}{\partial t^3} = -c^2 \frac{1}{\mu} \nabla \times \left(\frac{1}{\epsilon} \nabla \times \mathbf{H} \right). \tag{139}$$

Note that this is not quite the same as Eq. (134) for **E** since we have put ϵ inside the curl and μ outside the curl. However, if μ and ϵ are constant, we once again arrive at the conclusion that each component of **H** (and hence **B**) will satisfy the scalar wave equation.

C. The Energy Flux

We will now derive an expression for the flux of energy in an electromagnetic field. If we dot the first of Eqs. (130) with respect to **E**, and the second equation with respect to **H**, and add the results, we get the equation

$$\frac{1}{2c} \frac{\partial}{\partial t} (\mathbf{E} \cdot \mathbf{D} + \mathbf{B} \cdot \mathbf{H}) - \mathbf{E} \cdot \nabla \times \mathbf{H} + \mathbf{H} \cdot \nabla \times \mathbf{E} = 0. \tag{140}$$

If we use the identity

$$\nabla \cdot (\mathbf{A} \times \mathbf{B}) = \mathbf{B} \cdot \nabla \times \mathbf{A} - \mathbf{A} \cdot \nabla \times \mathbf{B}. \tag{141}$$

We see that

$$\frac{1}{2c} \frac{\partial}{\partial t} (\mathbf{E} \cdot \mathbf{D} + \mathbf{B} \cdot \mathbf{H}) + \nabla \cdot (\mathbf{E} \times \mathbf{H}) = 0. \tag{142}$$

When put in integral form this equation can be written as

$$\frac{1}{2c} \int_V (\mathbf{E} \cdot \mathbf{D} + \mathbf{B} \cdot \mathbf{H})\, dV = -\int_S (\mathbf{E} \times \mathbf{H}) \cdot \mathbf{n}\, dS, \tag{143}$$

where **n** is the outward facing normal to the surface. This equation can be interpreted as the fact that the quantity $1/2c(\mathbf{E} \cdot \mathbf{D} + \mathbf{B} \cdot \mathbf{H})$ is the energy density, and $\mathbf{E} \times \mathbf{H}$ is the flux of energy. The interpretation of $1/2c(\mathbf{E} \cdot \mathbf{D} + \mathbf{B} \cdot \mathbf{H})$ as the energy density of the field is actually clearer when we include charges in Maxwell's equations. In this case we would have to add a term to these equations that would represent the change in kinetic energy of the particles in the system. The vector

$$\mathbf{S} = c\mathbf{E} \times \mathbf{H}, \tag{144}$$

is referred to as the Poynting vector.

We will use the Poynting vector to justify the fact that the rays in geometrical optics are in fact the direction that energy is begin transported.

V. GEOMETRICAL OPTICS

A. Fermat's Principle

In order to understand our discussion of beam shaping it is essential to know how to use the laws of geometrical optics. Although understanding the derivation of the basic principles clearly leads to a deeper understanding, this is not essential for our presentation. For this reason we begin by stating the basic principles, and showing how we will use them. Once this is done, we will discuss the derivation of the principles.

Our treatment of geometrical optics is based on Fermat's principle (11). Fermat's principle is often stated as saying that the ray that gets from a point **a** to a point **b** will take the path that minimizes the travel time. This is a very concise statement of the principle, but it is not technically correct. Rather than saying that the true path minimizes the travel time, we need to say that the true path is stationary with respect to travel time. In many situations, the travel time is in fact minimized, but it is not always the case.

Before discussing what stationarity means in geometrical optics, we will clarify what we mean by stationarity in a simpler setting. The function $F(x,y) = (x-x_0)^2 + (y-y_0)^2$ has a minimum at $(x,y) = (x_0,y_0)$. A necessary condition that it has a minimum at (x_0,y_0) is that the partial derivatives of F vanish at (x_0,y_0). The vanishing of the partial derivatives is equivalent to saying that the function $F(x,y)$ is stationary at (x_0,y_0). Another way of putting this is to say that if we take any numbers $(\hat{x},\hat{y})$, then

$$F(x_0+\epsilon\hat{x}, y_0+\epsilon\hat{y}) = F(x_0,y_0) + O(\epsilon^2) \quad \text{as} \quad \epsilon \to 0. \tag{145}$$

At a point that is not stationary we would have $F(x_0+\epsilon\hat{x}, y_0+\epsilon\hat{y}) = F(x_0,y_0) + O(\epsilon)$ as $\epsilon \to 0$. In order for a function to have a minimum at (x_0,y_0) it must be stationary, but stationarity does not imply that the function is a minimum. For example, the function

$$F(x,y) = -x^2 - y^2 \tag{146}$$

is stationary at $(0,0)$, but it has a maximum rather than a minimum. The function

$$F(x,y) = x^2 - y^2 \tag{147}$$

is stationary at $(0,0)$, but it has neither a minimum nor a maximum.

Returning to geometrical optics, we will parametrize curves going from point **a** to **b** by a parameter s such that $0 \leq s \leq 1$. Let $\mathbf{x}(s) = (x(s), y(s), z(s))$ be a curve such that $\mathbf{x}(0) = \mathbf{a}$ and $\mathbf{x}(1) = \mathbf{b}$, then we will denote the travel time along this curve as

$$T(\mathbf{x}(s)) = \text{travel time}. \tag{148}$$

Suppose that $\mathbf{x}_0(s)$ is the true path that a light ray takes to get from **a** to **b**. The stationarity condition implies that for any functions $\hat{\mathbf{x}}(s)$ such that $\hat{\mathbf{x}}(0) = \hat{\mathbf{x}}(1) = 0$,

$$T(\mathbf{x}_0(s) + \epsilon\hat{\mathbf{x}}(s)) = T(\mathbf{x}_0(s)) + O(\epsilon^2). \tag{149}$$

Fermat's principle applies in an enormous variety of situations. Many times we put constraints on the travel paths. For example we can use Fermat's principle to show that the angle of incidence equals the angles of reflection for a light ray bouncing off of a mirror. In this case we use the constraint that a ray goes from point **a** to **b** after first touching a surface.

If a light ray gets from point **a** to point **b** by first passing through an intermediate point **c**, it can be shown that the paths from **a** to **c** and from **c** to **b** must each be stationary.

We will now give some concrete examples illustrating Fermat's principle. The first few examples that we give are not directly relevant to the beam shaping problem, but the last example is absolutely essential to understanding our discussion of beam shaping.

Example 1 Suppose we have a medium that has a constant speed of light. For a ray to get from a point **a** to a point **b** in the least amount of time it is clear that it must travel in a straight line. Since the travel path to get from **a** to **b** is a minimum, it is clear that it is also stationary. It can be shown that in this case, straight lines are the only stationary paths.

Example 2 Suppose the plane $z = 0$ separates medium I with a velocity of c_I from medium II with velocity c_{II}. What path does a light ray take to get from a point **a** in medium I to a point **b** in medium II. From our last example we already know that the path must be a straight line in each medium. For simplicity we will assume that the light ray travels in the plane $y = 0$. Suppose that $\mathbf{a} = (x_1, 0, z_1)$, and $\mathbf{b} = (x_2, 0, z_2)$. Suppose that in going from **a** to **b** the ray goes through $\mathbf{c} = (\xi, 0, 0)$ on the interface between the two media. We do not know the value of ξ ahead of time, but it can be determined using Fermat's principle. The total travel time to get from **a** to **b** by going through **c** is

$$T(\xi) = \frac{1}{c_I}\sqrt{(x_1 - \xi)^2 + z_1^2} + \frac{1}{c_{II}}\sqrt{(x_2 - \xi)^2 + z_2^2}. \tag{150}$$

In order for the travel time to be stationary, we must have

$$\frac{dT}{d\xi} = 0. \tag{151}$$

This implies that

$$\frac{\sin(\theta_I)}{c_I} = \frac{\sin(\theta_{II})}{c_{II}}, \tag{152}$$

where

$$\sin(\theta_I) = \frac{\xi - x_1}{\sqrt{(x_1 - \xi)^2 + z_1^2}}, \tag{153}$$

$$\sin(\theta_{II}) = \frac{x_2 - \xi}{\sqrt{(x_2 - \xi)^2 + z_2^2}}. \tag{154}$$

This is equivalent to Snell's law of refraction.

Example 3 Suppose we would like to design a mirror that focuses all of the light rays coming from a point **a** to a point **b**. For simplicity we will consider this problem to take place in two dimensions. We also assume that the speed of light is constant throughout our medium. Suppose a ray comes from **a** at an angle of θ with the horizontal. Suppose that this ray bounces off the mirror at a point $\mathbf{p}(\theta)$, and then goes to the point **b**. Let $T(\theta)$ be the travel time to get from **a** to $\mathbf{p}(\theta)$ and then to **b**. In order for Fermat's principle to hold we must have

$$\frac{dT}{d\theta} = 0. \tag{155}$$

This means that if $\mathbf{q}_1$ is any point on the mirror, then the distance from **a** to $\mathbf{q}_1$ plus the distance from $\mathbf{q}_1$ to **b** must be the same as for any other point $\mathbf{q}_2$ on the mirror. This implies that the mirror must in fact have the shape of an ellipse, with foci at **a** and **b**.

It should be noted that as we move the point **a** off to ∞, this ellipse ends up turning into a parabola. This gives us the solution of how to focus rays coming in from ∞ to a single point **b**.

Example 4 When using Fermat's principle for parallel beams of light it is necessary to be familiar with the following argument. Suppose we have a parallel beam of rays coming in from ∞. We can think of such rays as coming from a very distant point source. Suppose the point source is at

$\mathbf{p} = (-L, 0, 0)$. Assuming a homogeneous medium, the time to get from $\mathbf{p}$ to a point (x, y, z) is given by

$$T(x, y, z) = \frac{1}{c}\sqrt{(x+L)^2 + y^2 + z^2}. \tag{156}$$

As $L \to \infty$ we can make the approximation

$$T(x, y, z) = \frac{1}{c}(L + x + O(1/L)). \tag{157}$$

This shows that the travel time to get to any point in space (x, y, z) is independent of y and z. When applying Fermat's principle, the travel time L/c will not matter since it is the same for all paths. We will use this fact whenever we are applying Fermat's principle to rays that are coming in parallel.

Example 5 We now give an example that shows that the ray paths do not always minimize the travel time, but they are still stationary with respect to travel time. Suppose we have a cylindrical mirror (see Fig. 4) whose surface is given by

$$(x, y) = R(\cos(\theta), \sin(\theta)) - \pi/2 \le \theta \le \pi/2. \tag{158}$$

We are interested in finding the paths of rays that are coming in parallel from $x = -\infty$. As we have already mentioned, this can be thought of as rays coming from a distant source at $\mathbf{p} = (-L, 0)$ where L is very large. The travel time to get from $\mathbf{p}$ to a point on the surface of the mirror $R(\cos(\theta), \sin(\theta))$, and then to a point (x, y) is (assuming $L \gg 1$)

$$T(\theta) = \frac{1}{c}(L + R\cos(\theta) + \sqrt{(x - R\cos(\theta))^2 + (y - R\sin(\theta))^2}). \tag{159}$$

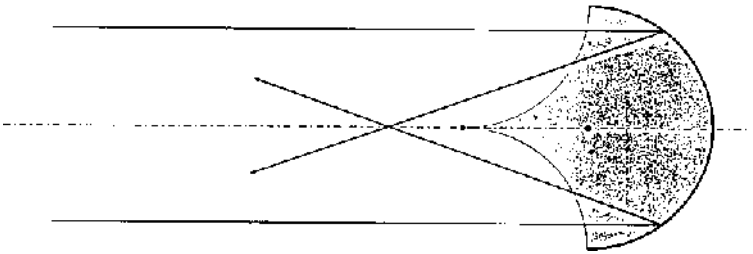


Figure 4 A schematic of light being reflected by a circular mirror. In the shaded region, there are two rays that reach each point by a single reflection off of the mirror. Outside of this region, there is only one ray that reaches each point by a single reflection.

Given a point (x, y) the equation $dT/d\theta = 0$ will determine where the ray that reaches (x, y) reflects off the mirror. For simplicity we will limit ourselves to points such that $y = 0$. In this case we can write the stationarity condition as

$$\frac{dT}{d\theta} = \frac{R\sin(\theta)}{c}\left(-1 + \frac{x}{\sqrt{(x - R\cos(\theta))^2 + R^2\sin^2(\theta)}}\right). \tag{160}$$

This equation has the solutions

$$\sin(\theta) = 0, \tag{161}$$

or

$$x^2 = (x - R\cos(\theta))^2 + R^2\sin^2(\theta). \tag{162}$$

This last equation can be written as

$$x = \frac{R}{2\cos(\theta)}. \tag{163}$$

For any value of x the first of these equations gives us the solution $\theta = 0$. However, the second equation will have no solutions if $x < R/2$, and will have two solutions if $x > R/2$. When we look throughout the xy plane we find that there will be a region that has three reflected rays reaching each point, and another region with only one reflected ray reaching each point. The curve separating the two regions is an example of a caustic surface. These caustic surfaces are easily observed since the irradiance at the surface becomes much larger than at a typical point in the plane. This particular situation can be observed when looking into a cup of tea, or a bowl of sugar under the light from a concentrated source such as an incandescent light bulb.

Note that not all of the rays have minimum travel time. If we compute $d^2T/d\theta^2$ we get

$$\begin{aligned}\frac{d^2T}{d\theta^2} = {} & \frac{R\cos(\theta)}{c}\left(-1 + \frac{x}{\sqrt{(x - R\cos(\theta))^2 + R^2\sin^2(\theta)}}\right) \\ & - \frac{R^2\sin^2(\theta)}{c}\frac{x^2}{((x - R\cos(\theta))^2 + R^2\sin^2(\theta))^{3/2}}.\end{aligned} \tag{164}$$

If we restrict our attention to the ray that hits the mirror at $\theta = 0$, we have

$$\frac{d^2T(0)}{d\theta^2} = \frac{R}{c}\left(-1 + \frac{x}{|x - R|}\right). \tag{165}$$

This shows that if $x > R/2$, then the travel time is a minimum, but if $x < R/2$ the travel time is in fact a local maximum.

Example 6 This last example plays a large role in our theory of beam shaping since it shows that our Fourier transform lens has a quadratic time delay, and hence a quadratic phase function. Suppose we would like to place a lens at $x = 0$ that focuses all of the rays coming in from $x = -\infty$ to a single point at $(x, y) = (f, 0)$. We will make several approximations. To begin with, we assume that the lens is thin. This means that the rays that enter the lens at $(0, y)$ emerge at very nearly the same point. We can model the effect of the lens by saying that it introduces a time delay of $t_L(y)$. This means that a ray that enters the lens at $(0, y)$ takes a time $t_L(0, y)$ to emerge from the lens. In practice this time delay can be introduced by making the lens out of a material that has a different index of refraction than the medium that the rays are traveling in, and by varying the thickness of the lens.

We would like to determine the function $t_L(y)$ such that all of the rays from ∞ get focused to the point $(f, 0)$. The time for a ray to go from a distant point $(-L, 0)$ to a point $(0, y)$ and then to the point $(f, 0)$ is approximately

$$t(y) = \frac{1}{c}(L + \sqrt{f^2 + y^2}) + t_L(y). \tag{166}$$

Here we have made the approximation that $L \gg 1$, and used the simplified expression for the distance from $(-L, 0)$ to $(0, y)$. We will now make the paraxial approximation. This assumes that all points on the lens satisfy $y^2/f^2 \ll 1$. In this case we can approximate the square root using

$$\sqrt{f^2 + y^2} \approx f + \frac{y^2}{2f}. \tag{167}$$

Using this approximation, we can write

$$t(y) \approx \frac{1}{c}\left(L + f + \frac{y^2}{2f}\right) + t_L(y). \tag{168}$$

Fermat's principle requires that the path that gets from $(-L, 0)$ to $(f, 0)$ must be stationary with respect to all nearby paths. This implies that

$$\frac{dt}{dy} = 0 \text{ for all } y. \tag{169}$$

This means that we must have $t(y) = \text{constant}$, and hence

$$t_L(y) = -\frac{y^2}{2fc}. \tag{170}$$

This shows that in the paraxial approximation, we must use a quadratic lens in order to focus an initially parallel beam of rays to a point.

B. The Eikonol Equation

The laws of geometrical optics can be derived as a high-frequency approximation to the solutions of Maxwell's equations. The rays of light are related to the direction of energy propagation. There is a very strong connection between Fermat's principle and the method of stationary phase. Both the method of stationary phase and the laws of geometrical optics are high frequency limits, they both are centered about the phase of the wave field, and they both use a stationarity condition.

Before considering the high-frequency limit of Maxwell's equations, we will begin by considering the high-frequency limit of the scalar wave equation. Suppose we have a solution $p(\mathbf{x}, \omega)$ to the equation

$$\nabla^2 p + \frac{\omega^2}{c^2(\mathbf{x})}\, p = 0. \tag{171}$$

This is the time harmonic wave equation, also known as the reduced wave equation. We are interested in determining the behavior of these solutions for large values of ω. In particular we ask what solutions that are coming from a single point source look like. The theory of Green's functions shows that the general high-frequency limit (not necessarily from a point source) can be built up by integrating over many such point sources. If the velocity is constant we know that the point source solutions can be written as

$$p(\mathbf{x}, \omega) = A\, \frac{e^{i\omega r/c}}{r}. \tag{172}$$

Here $r^2 = x^2 + y^2 + z^2$. This solution has a very rapidly varying phase (it varies more rapidly the bigger ω is), and a slowly varying amplitude (that is independent of ω). In the case of variable $c(\mathbf{x})$ we assume that even though the amplitude may not be completely independent of ω, it depends very weakly on ω. Generalizing to the case of non-homogeneous media, in the high-frequency limit we will assume that $p(\mathbf{x}, \omega)$ can be written as

$$p(\mathbf{x}, \omega) = A(\mathbf{x}) e^{i\omega\phi(\mathbf{x})}. \tag{173}$$

This is only the first term in an asymptotic expansion. The general solution needs to include corrections to the amplitude that depend on ω.

We can write

$$\nabla p = (\nabla A + Ai\omega\nabla\phi)e^{i\omega\phi}. \tag{174}$$

Using this we can now write

$$\nabla^2 p = \nabla \cdot \nabla p = (\nabla^2 A + 2i\omega\nabla A \cdot \nabla\phi + i\omega A\nabla^2\phi - \omega^2 A\nabla\phi \cdot \nabla\phi)e^{i\omega\phi}. \tag{175}$$

If we substitute this expression into Eq. (171) and keep only the highest-order term in ω, we find that

$$|\nabla\phi|^2 = \frac{1}{c(\mathbf{x})^2}. \tag{176}$$

This equation is usually referred to as the eikonol equation. The next higher-order term gives us

$$2\nabla A \cdot \nabla\phi + A\nabla^2\phi = 0. \tag{177}$$

This last equation can be written as

$$\nabla \cdot (A^2\nabla\phi) = 0. \tag{178}$$

The fact that this equation can be written in divergence form suggests that the quantity $A^2\nabla\phi$ is the flux of some quantity that is conserved. When we apply these arguments to optical systems we will see that this quantity is in fact proportional to the flux of energy.

We mentioned that the general high-frequency approximation can be built up by integrating or summing over a family of point sources. As a simple example, if our wave field comes from two point sources, the high-frequency limit of the wave field will look like

$$p(\mathbf{x}, \omega) = A_1(\mathbf{x})e^{i\omega\phi_1(\mathbf{x})} + A_2(\mathbf{x})e^{i\omega\phi_2(\mathbf{x})}. \tag{179}$$

It should be mentioned that even for a single point source there may be points in space where the high-frequency limit consists of a sum of terms as in the previous equation. This will be the case if the rays are bent so that more than one ray from the same source reaches the same point. The surfaces separating the regions where there are different numbers of rays are the caustic surfaces.

C. The Eikonol Equation and Maxwell's Equations

In the previous section we derived the eikonol equation from the scalar wave equation. Each component of the eletromagnetic wave field satisfies this equation, so it is not surprising that the eikonol equation also arises when considering the high-frequency limit of Maxwell's equations. In this section we will derive the eikonol equation using Maxwell's equations, and we will see that Poynting's theorem shows that the energy of the electromagnetic field is in fact being propagated normal to the surfaces of constant phase. The derivation of the eikonol equation from Maxwell's equations is almost identical to the analysis of plane monochromatic plane waves given in most textbooks on electrodynamics (12).

The time-harmonic Maxwell's equations are

$$i\omega \frac{1}{c_0} \mathbf{D} - \nabla \times \mathbf{H} = 0, \tag{180a}$$

$$i\omega \frac{1}{c_0} \mathbf{B} + \nabla \times \mathbf{E} = 0, \tag{180b}$$

where

$$\mathbf{D} = \epsilon(\mathbf{x})\mathbf{E}, \tag{181a}$$

$$\mathbf{B} = \mu(\mathbf{x})\mathbf{H}. \tag{181b}$$

Similar to our derivation of the eikonol equation for the scalar wave equation, we assume a solution of the form.

$$\mathbf{E}(\mathbf{x}, \omega) = \mathbf{E}_0(\mathbf{x}) e^{i\omega\phi(\mathbf{x})}, \tag{182}$$

$$\mathbf{H}(\mathbf{x}, \omega) = \mathbf{H}_0(\mathbf{x}) e^{i\omega\phi(\mathbf{x})}. \tag{183}$$

Substituting this expression into Maxwell's equations and using the vector identity

$$\nabla \times (f(\mathbf{x})\mathbf{A}(\mathbf{x})) = f(\mathbf{x})\nabla \times \mathbf{A}(\mathbf{x}) + \mathbf{A}(\mathbf{x}) \times \nabla f(\mathbf{x}), \tag{184}$$

we get

$$i\omega \frac{\epsilon}{c_0} \mathbf{E}_0 - (\nabla \times \mathbf{H}_0 + i\omega \mathbf{H}_0 \times \nabla\phi) = 0, \tag{185a}$$

$$i\omega \frac{\mu}{c_0} \mathbf{H}_0 + (\nabla \times \mathbf{E}_0 + i\omega \mathbf{E}_0 \times \nabla\phi) = 0. \tag{185b}$$

If we only keep the highest-order terms in ω in this equation we get

$$\frac{\epsilon}{c_0}\,\mathbf{E}_0 - \mathbf{H}_0 \times \nabla\phi = 0, \tag{186a}$$

$$\frac{\mu}{c_0}\,\mathbf{H}_0 + \mathbf{E}_0 \times \nabla\phi = 0. \tag{186b}$$

If we dot each of these equations with $\nabla\phi$ we find that

$$\mathbf{E}_0 \cdot \nabla\phi = 0, \tag{187}$$

$$\mathbf{H}_0 \cdot \nabla\phi = 0. \tag{188}$$

If we dot the first of Eqs. (186) with respect to $\mathbf{H}_0$, or the second with respect to $\mathbf{E}_0$, we find that

$$\mathbf{E}_0 \cdot \mathbf{H}_0 = 0. \tag{189}$$

If we eliminate $\mathbf{H}_0$ from Eqs. (186) we find that

$$\frac{\epsilon\mu}{c_0}\,\mathbf{E}_0 + c_0(\mathbf{E}_0 \times \nabla\phi) \times \nabla\phi = 0. \tag{190}$$

Using the identity

$$(\mathbf{a} \times \mathbf{b}) \times \mathbf{c} = \mathbf{b}(\mathbf{a} \cdot \mathbf{c}) - \mathbf{a}(\mathbf{b} \cdot \mathbf{c}), \tag{191}$$

we find that this can be written as

$$\frac{\epsilon\mu}{c_0^2}\,\mathbf{E}_0 - (\mathbf{E}_0 \nabla\phi \cdot \nabla\phi - \nabla\phi \mathbf{E}_0 \cdot \nabla\phi) = 0. \tag{192}$$

Using the fact that $\mathbf{E}_0 \cdot \nabla\phi = 0$ this can be written as

$$\mathbf{E}_0 \left(\frac{1}{c^2(\mathbf{x})} - \nabla\phi \cdot \nabla\phi \right) = 0, \tag{193}$$

where

$$c^2(\mathbf{x}) = \frac{c_0^2}{\epsilon\mu}. \tag{194}$$

We see that we have once again arrived at the eikonol equation.

The direction of energy flux is given by the Poynting vector

$$\mathbf{S} = c\mathbf{E} \times \mathbf{H}. \tag{195}$$

Using the fact that $\mathbf{E}$ and $\mathbf{H}$ are orthogonal to each other, and also to $\nabla\phi$, it follows that this vector is in the direction of $\nabla\phi$. We see that the direction of energy flux is in fact normal to the surfaces of constant phase.

D. First-Order Non-linear Partial Differential Equations

The theory of ray tracing from the eikonal equation is a special case of the solution of non-linear first-order partial differential equations (14, 15). In this section we will give a brief outline of this theory. Suppose we have an equation of the form

$$F(x, y, z, p, q, r) = 0, \tag{196a}$$

where

$$p = \frac{\partial \phi}{\partial x}, \tag{196b}$$

$$q = \frac{\partial \phi}{\partial y}, \tag{196c}$$

$$r = \frac{\partial \phi}{\partial z}. \tag{196d}$$

For optical applications we are especially concerned with the case where

$$F(x, y, z, p, q, r) = \frac{1}{2}\left(p^2 + q^2 + r^2 - \frac{1}{c^2(\mathbf{x})}\right), \tag{197}$$

which is just the eikonol equation. In this section we will consider the case for a general function F rather than limiting ourselves to the eikonal equation. Our analysis could be extended to the case where the function F also depends on the function ϕ, but that case is just a bit more complicated, and it never arises in optical applications, so we will not consider it here.

Suppose we know the function ϕ and all of its first derivatives at some point (x_0, y_0, z_0). Is it possible to determine the solution ϕ in the neigborhood of the point (x_0, y_0, z_0)? In particular, is it possible to determine the second derivatives of the function ϕ at the point (x_0, y_0, z_0)? If we take the derivatives of our equation with respect to x, y, and z we end up with the equations

$$\frac{\partial F}{\partial x} + \frac{\partial F}{\partial p}\frac{\partial p}{\partial x} + \frac{\partial F}{\partial q}\frac{\partial q}{\partial x} + \frac{\partial F}{\partial r}\frac{\partial r}{\partial x} = 0, \tag{198a}$$

$$\frac{\partial F}{\partial y} + \frac{\partial F}{\partial p}\frac{\partial p}{\partial y} + \frac{\partial F}{\partial q}\frac{\partial q}{\partial y} + \frac{\partial F}{\partial r}\frac{\partial r}{\partial y} = 0, \tag{198b}$$

$$\frac{\partial F}{\partial z} + \frac{\partial F}{\partial p}\frac{\partial p}{\partial z} + \frac{\partial F}{\partial q}\frac{\partial q}{\partial z} + \frac{\partial F}{\partial r}\frac{\partial r}{\partial z} = 0. \tag{198c}$$

The partial derivatives of p, q and r can all be expressed in terms of the six second-order partial derivatives of ϕ:

$$\frac{\partial^2\phi}{\partial x^2}, \frac{\partial^2\phi}{\partial y^2}, \frac{\partial^2\phi}{\partial z^2}, \frac{\partial^2\phi}{\partial x\,\partial y}, \frac{\partial^2\phi}{\partial x\,\partial z}, \frac{\partial^2\phi}{\partial y\,\partial z}. \tag{199}$$

By differentiating our equation $F(x, y, p, q, r) = 0$ we have arrived at three equations for the six second-order partial derivatives of ϕ. Clearly we do not have enough equations to determine the second-order partial derivatives. The question now arises, is it possible to determine the derivatives of p, q and r in a particular direction? It turns out that this is in fact possible. To do this we use the fact that

$$\frac{\partial p}{\partial y} = \frac{\partial^2\phi}{\partial x\,\partial y} = \frac{\partial^2\phi}{\partial y\,\partial x} = \frac{\partial q}{\partial x}. \tag{200a}$$

Similarly

$$\frac{\partial r}{\partial x} = \frac{\partial p}{\partial z}. \tag{200b}$$

It follows that the first of Eqs. (198) can be written as

$$\frac{\partial F}{\partial x} + \frac{\partial F}{\partial p}\frac{\partial p}{\partial x} + \frac{\partial F}{\partial q}\frac{\partial p}{\partial y} + \frac{\partial F}{\partial r}\frac{\partial p}{\partial z} = 0. \tag{201a}$$

Simlarly, by switching the order of the other mixed partial derivatives we can get

$$\frac{\partial F}{\partial y} + \frac{\partial F}{\partial p}\frac{\partial q}{\partial x} + \frac{\partial F}{\partial q}\frac{\partial q}{\partial y} + \frac{\partial F}{\partial r}\frac{\partial q}{\partial z} = 0, \tag{201b}$$

and

$$\frac{\partial F}{\partial z} + \frac{\partial F}{\partial p}\frac{\partial r}{\partial x} + \frac{\partial F}{\partial q}\frac{\partial r}{\partial y} + \frac{\partial F}{\partial r}\frac{\partial r}{\partial z} = 0. \tag{201c}$$

These equations can be written as

$$\frac{\partial F}{\partial x} + \nabla p \cdot \mathbf{a} = 0, \tag{202a}$$

$$\frac{\partial F}{\partial y} + \nabla q \cdot \mathbf{a} = 0, \tag{202b}$$

and

$$\frac{\partial F}{\partial z} + \nabla \mathbf{r} \cdot \mathbf{a} = 0, \tag{202c}$$

where

$$\mathbf{a} = \left(\frac{\partial F}{\partial p}, \frac{\partial F}{\partial q}, \frac{\partial F}{\partial r}\right). \tag{202d}$$

These equations show that although we do not know the derivatives of p, q and r in any arbitrary direction,, we do know the derivatives in the direction $\mathbf{a} = (\partial F/\partial p, \partial F/\partial q, \partial F/\partial r)$. This suggests that there may be special curves $(x(s), y(s), z(s))$ such that we can determine $(p(s), q(s), r(s))$. In particular, if

$$\dot{x} = \frac{\partial F}{\partial p}, \tag{203}$$

$$\dot{y} = \frac{\partial F}{\partial q}, \tag{204}$$

$$\dot{z} = \frac{\partial F}{\partial r}, \tag{205}$$

then we have

$$\dot{p} = -\frac{\partial F}{\partial x}, \tag{206}$$

$$\dot{q} = -\frac{\partial F}{\partial y}, \tag{207}$$

$$\dot{r} = -\frac{\partial F}{\partial z}. \tag{208}$$

The function ϕ changes according to the equation

$$\dot{\phi} = \frac{\partial \phi}{\partial x}\,\dot{x} + \frac{\partial \phi}{\partial y}\,\dot{y} + \frac{\partial \phi}{\partial z}\,\dot{z} = p\dot{x} + q\dot{y} + r\dot{z} = p\,\frac{\partial F}{\partial p} + q\,\frac{\partial F}{\partial q} + r\,\frac{\partial F}{\partial r}. \tag{209}$$

This is a seventh-order ordinary differential equation for the unknowns (x, y, z, p, q, r, ϕ). We can solve this system of equations provided we specify initial values of (x, y, z, p, q, r, ϕ). It should be noted that we cannot specify these values arbitrarily, but must require that they satisfy the equation $F(x, y, z, p, q, r) = 0$.

In optics the function ϕ is the phase of our wave field. The curves $(x(s), y(s), z(s))$ along which we propagate our solution are know as the rays. In optics they are what we intuitively think of as being the rays of light. If we know the phase $\phi(x, y, z)$ on some plane $z = z_0$, then we can parametrically map out the phase in all of space by tracing all the rays from the plane $z = z_0$.

This process of tracing out the phase field assumes that one and only one ray passes from a given point (x, y, z) to the plane $z = z_0$. In practice it is possible that no rays pass through some points, and multiple rays pass through other points. The surfaces separating regions with different numbers of rays are once again the caustic surfaces.

Example 7 We will now apply this theory to the eikonol equation where

$$F(x, y, z, p, q, r) = \frac{1}{2}\left(p^2 + q^2 + r^2 - \frac{1}{c^2(\mathbf{x})}\right). \tag{210}$$

We will use the shorthand notation $\mathbf{p} = (p, q, r)$ and $\mathbf{x} = (x, y, z)$. The theory we have just derived shows that

$$\frac{d\mathbf{x}}{ds} = \mathbf{p}, \tag{211a}$$

$$\frac{d\mathbf{p}}{ds} = -\frac{\nabla c}{c^3}, \tag{211b}$$

and

$$\frac{d\phi}{ds} = \frac{1}{c^2(\mathbf{x})}, \tag{211c}$$

where $\mathbf{p}$ is required to satisfy the initial condition

$$\mathbf{p} \cdot \mathbf{p} = 1/c^2. \tag{212}$$

We can eliminate $\mathbf{p}$ from this equation to get

$$\frac{d^2\mathbf{x}}{ds^2} = -\frac{\nabla c}{c^3}. \tag{213}$$

Note that the equation for ϕ can be written as

$$\frac{d\phi}{ds} = \frac{|\dot{\mathbf{x}}|}{c(\mathbf{x})}. \tag{214}$$

This equation can be interpreted as saying that the change is ϕ in going from $\mathbf{x}(s_0)$ to $\mathbf{x}(s_1)$ is the travel time to get from $\mathbf{x}(s_0)$ to $\mathbf{x}(s_1)$ along the curve $\mathbf{x}(s)$.

Example 8 The system of equations (213) is in some ways the simplest set of equations we could write down for the paths of the light rays. However, it suffers from one problem. The equations are not invariant under a change of parametrization. If we parameterize our curves by $\xi = \xi(s)$, we will end up getting a different differential equation for $\mathbf{x}(\xi)$ than we got for $\mathbf{x}(s)$. The solutions will result in the same curve in physical space, but the differential

equations will be different. Unless one is extremely concerned with the aesthetic properties of their equations this is not a real problem. However, it turns out that the equations that we derive using Fermat's principle will be invariant under a change of parametrization, and hence it will be difficult to compare the two sets of equations unless we write equations (213) so that they are also invariant.

In order to do this we note that the first of Eqs. (213) requires that

$$|\dot{\mathbf{x}}| = |\mathbf{p}| = \frac{1}{c}. \tag{215}$$

It follows that along our solution curve we have

$$|\dot{\mathbf{x}}| c(\mathbf{x}) = 1. \tag{216}$$

It follows that we will not change the solutions to Eqs. (211) if we divide the left-hand sides by $|\dot{\mathbf{x}}| c(\mathbf{x})$. In this case we get the equations

$$\frac{1}{|\dot{\mathbf{x}}| c(\mathbf{x})} \dot{\mathbf{x}} = \mathbf{p}, \tag{217}$$

$$\frac{1}{|\dot{\mathbf{x}}| c(\mathbf{x})} \dot{\mathbf{p}} = -\nabla c \frac{1}{c^3}. \tag{218}$$

Now if we eliminate $\mathbf{p}$ from these equations we end up with the system of equations

$$\frac{1}{|\dot{\mathbf{x}}| c(\mathbf{x})} \frac{d}{ds} \left(\frac{\dot{\mathbf{x}}}{|\dot{\mathbf{x}}| c(\mathbf{x})} \right) = -\frac{\nabla c}{c^3}. \tag{219}$$

This is the final system of equations that we will use to compare to the curves obtained by using Fermat's principle. Note that if we make a change of variables $\xi = \xi(s)$, the differential equation in terms of ξ will be identical to the differential equation in terms of s.

E. Fermat's Principle without Reflections

In the last example we derived the equations for the path $\mathbf{x}(s)$ that a light ray follows in an inhomogeneous medium. We will now show that the path that gets from point $\mathbf{x}_0$ to point $\mathbf{x}_1$ is stationary with respect to the travel time between these two points. Suppose we have a curve $\mathbf{x}(s)$ such that $\mathbf{x}(s_0) = \mathbf{x}_0$ and $\mathbf{x}(s_1) = \mathbf{x}_1$. The time to get from the point $\mathbf{x}_0$ to the point $\mathbf{x}_1$ may be written as

$$T = \int_{s_0}^{s_1} \sqrt{\dot{\mathbf{x}} \cdot \dot{\mathbf{x}}} \frac{1}{c(\mathbf{x})} \, ds. \tag{220}$$

The first variation of this integral may be written as

$$\delta T = \int_{s_0}^{s_1} \frac{\delta\dot{\mathbf{x}} \cdot \dot{\mathbf{x}}}{c(\mathbf{x})\sqrt{\dot{\mathbf{x}} \cdot \dot{\mathbf{x}}}} - |\dot{\mathbf{x}}| \frac{\nabla c \cdot \delta\mathbf{x}}{c^2(\mathbf{x})}\, ds. \tag{221}$$

If we integrate by parts to get rid of the derivative with respect to $\delta\mathbf{x}$, and if we require that $\delta\mathbf{x}$ vanish at the end points, we find that

$$\delta T = -\int_{s_0}^{s_1} \left(\frac{d}{ds}\left(\frac{\dot{\mathbf{x}}}{|\dot{\mathbf{x}}|c(\mathbf{x})}\right) + |\dot{\mathbf{x}}| \frac{\nabla c}{c^2(\mathbf{x})} \right) \cdot \delta\mathbf{x}\, ds. \tag{222}$$

If the path is stationary, then this integral must vanish for all functions $\delta\mathbf{x}$, and hence $\mathbf{x}$ must satisfy the equation

$$\frac{1}{c(\mathbf{x})|\dot{\mathbf{x}}|} \frac{d}{ds}\left(\frac{\dot{\mathbf{x}}}{|\dot{\mathbf{x}}|c(\mathbf{x})}\right) = -\frac{\nabla c}{c^3}. \tag{223}$$

This is identical to the Eq. (219) which we derived using the eikonol equation, and requiring that the equation be invariant under a change of parameterization.

F. Fermat's Principle for Reflecting Surfaces

In our analysis of beam shaping systems we will not consider any cases where the rays reflect off of mirrors. However, since it may sometimes be desirable to use reflecting surfaces in beam shaping systems, we now consider Fermat's principle for reflecting surfaces. Suppose we have a surface S defined parametrically by $\mathbf{x} = \mathbf{f}(\xi_1, \xi_2)$. Suppose that a ray goes from the point $\mathbf{x}_0$ to the point $\mathbf{x}_1$, but first bounces off the surface S. The theory of waves shows that at the point where the rays get reflected by the surface, the following conditions hold:

- The normal to the surface, the incident ray, and the reflected ray all lie in the same plane.
- The incident and reflected rays make the same angle with respect to the normal to the surface.

We now show that these conditions can be derived by assuming that the path from $\mathbf{x}_0$ to $\mathbf{x}_1$ that touches the surface S is stationary with respect to travel time. Suppose we have a path that goes from $\mathbf{x}_0$ to $\mathbf{x}_1$ after first touching some point $\mathbf{q} = \mathbf{f}(\xi_1, \xi_2)$ on the surface S. Clearly the paths from $\mathbf{x}_0$ to $\mathbf{q}$ and from $\mathbf{q}$ to $\mathbf{x}_1$ must themselves be stationary. It follows that in order to determine the true path we need only determine the point $\mathbf{q}$ on the surface S. In particular, suppose $\phi(\mathbf{x}, \mathbf{z})$ gives the travel time to get from the point $\mathbf{x}$ to the point $\mathbf{z}$. We have shown that the travel time $\phi(\mathbf{x}, \mathbf{z})$ is in fact

a solution to the eikonol equation. The total travel time from $\mathbf{x}_0$ to $\mathbf{x}_1$ is given by

$$T = \boldsymbol{\phi}(\mathbf{x}_0, \mathbf{f}(\xi_1, \xi_2)) + \boldsymbol{\phi}(\mathbf{x}_1, \mathbf{f}(\xi_1, \xi_2)). \tag{224}$$

If this travel time is stationary, then we must have

$$\frac{\partial \mathbf{f}}{\partial \xi_k} \cdot (\mathbf{p}_i + \mathbf{p}_r) = 0 \text{ for } k = 1, 2, \tag{225}$$

where

$$\mathbf{p}_i = \frac{\partial \phi(\mathbf{x}_0, \mathbf{q})}{\partial \mathbf{q}} \tag{226a}$$

and

$$\mathbf{p}_r = \frac{\partial \phi(\mathbf{x}_1, \mathbf{q})}{\partial \mathbf{q}} \tag{226b}$$

are the incident and the reflected ray vectors.

Note that the vectors $\mathbf{p}_i$ and $\mathbf{p}_r$ must satisfy $|\mathbf{p}| = 1/c(\mathbf{q})$, and hence we must have $|\mathbf{p}_i| = |\mathbf{p}_r|$. Let $\mathbf{n}$ be the normal to the surface S at $\mathbf{f}(\xi_1, \xi_2)$, and let $\mathbf{t}_1$ and $\mathbf{t}_2$ be two independent tangent vectors to the surface. These vectors can be written as linear combinations of the vectors $\partial \mathbf{f}/\partial \xi_1$ and $\partial \mathbf{f}/\partial \xi_2$, and hence Eq. (225) shows that the tangential components of $\mathbf{p}_i$ and $\mathbf{p}_r$ must be negatives of each other. That is,

$$\mathbf{p}_i = a_i \mathbf{n} + b_i \mathbf{t}_1 + c_i \mathbf{t}_2, \tag{227}$$

then

$$\mathbf{p}_r = a_r \mathbf{n} - b_i \mathbf{t}_1 - c_i \mathbf{t}_2. \tag{228}$$

Furthermore, in order for $\mathbf{p}_r$ and $\mathbf{p}_i$ to have the same magnitude, we must have $a_i = a_r$ This shows that $\mathbf{p}_i$, $\mathbf{p}_r$, and $\mathbf{n}$ all lie in the same plane. Furthermore,

$$\mathbf{p}_i \cdot \mathbf{n} = \mathbf{p}_r \cdot \mathbf{n}, \tag{229}$$

and hence the vectors $\mathbf{p}_i$ and $\mathbf{p}_r$ make the same angle with respect to $\mathbf{n}$. This is precisely what we wanted to prove.

VI. FOURIER OPTICS AND DIFFRACTION THEORY

A. Fresnel Diffraction Theory

Fresnel diffraction theory plays an important role in the theory of beam shaping since it allows us to access the validity of the geometrical optics

approximation. Through the theory of Fresnel diffraction we will be able to turn our physical optics problem into a mathematical problem concerning Fourier transforms. After giving a derivation of the Fresnel approximation, we will outline the conditions necessary for it to be a good approximation.

The Fresnel approximation is concerned with the wave field for $z > 0$ produced when an incoming wave passes through an aperture at $z = 0$. In general the aperture may contain an optical element that changes the amplitude or phase of the incoming wave. Elementary theories of diffraction usually are concerned with the field far from the aperture, and in a narrow solid angle normal to the aperture. The theory of Fresnel diffraction can be outlined in three basic steps.

- Write down an exact expression for determining the wave field for all values of $z > 0$ provided one knows the wavefield at the plane of the aperture $z = 0$.
- Use a paraxial approximation that simplifies this expression assuming the observation point is near the axis.
- Compute the wavefield away from the aperture by using the first two steps along with a very simple assumption concerning the field in the plane of the aperture. The assumption is that at the aperture the wavefield is equal to the undisturbed incoming wavefield (modified by any optical element inside the aperture), and zero everywhere else.

The first of these steps can be carried out rigorously. The second step can be justified quite well using simple asymptotics. The third step is by far the hardest to justify, but it can be argued that it is plausible provided the aperture is large compared to the wavelength of the incoming light.

We begin with a discussion of the Fresnel approximation for the scalar wave equation. Physically we can think of this equation arising from the equations of acoustics. When we present the vector theory of diffraction we will see that this theory can be used to determine the various components of the electric field, but a slight error occurs in the component of the field normal to the aperture. This error is not big as long as we are near the axis.

B. A Fourier Approach to Diffraction Theory

We suppose that function $u(x, y, z)$ satisfies the Helmholtz equation

$$\nabla^2 u + k^2 u = 0, \tag{230a}$$

where

$$k = \frac{2\pi}{\lambda}, \tag{230b}$$

and λ is the wavelength. The function u must also satisfy the boundary condition

$$u(x, y, 0) = f(x, y) \tag{230c}$$

and

$$u(x, y, z) \text{ has no incoming waves as } z \to \infty. \tag{230d}$$

This last boundary condition, sometimes referred to as the Sommerfeld radiation condition, is a somewhat subtle condition, but it is quite straightforward to implement when doing analytical work. It requires that as $z \to \infty$ all of the waves will be traveling away from $z = 0$, not towards it.

We choose to solve these equations by spatially Fourier transforming the function $u(x, y, z)$ in the x and y directions. Let

$$U(k_x, k_y, z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-i(k_x x + k_y y)} u(x, y, z)\, dx\, dy, \tag{231}$$

be the Fourier transform of u, and $F(k_x, k_y)$ be the Fourier transform of $f(x, y)$. The function U must satisfy the equations

$$\frac{d^2 U}{dz^2} + (k^2 - k_x^2 - k_y^2) U = 0, \tag{232a}$$

$$U(k_x, k_y, 0) = F(k_x, k_y), \tag{232b}$$

$$U(k_x, k_y, z) \text{ has only outgoing waves as } z \to \infty. \tag{232c}$$

The solution to this set of equations can be written as

$$U(k_x, k_y, z) = F(k_x, k_y) e^{iz\sqrt{k^2 - k_x^2 - k_y^2}}. \tag{233}$$

The sign of the square root must be chosen so that the field decays as $z \to \infty$, and so that there are no incoming waves from infinity. In order to ensure this we must choose the positive square root for $k^2 - k_x^2 - k_y^2 > 0$, and choose it so that $i\sqrt{k^2 - k_x^2 - k_y^2} < 0$ for $k^2 - k_x^2 - k_y^2 < 0$.

We can now inverse Fourier transform this to get

$$u(x, y, z) = \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{i(k_x x + k_y y)} F(k_x, k_y) e^{ikz\sqrt{1 - k_x^2/k^2 - k_y^2/k^2}}\, dk_x\, dk_y. \tag{234}$$

We have now accomplished the first step in deriving the Fresnel approximation; we have derived an exact expression for the field u in terms of its value at $z = 0$.

This form for the field is sometimes used in diffraction theory. However, both analytical and numerical work is usually much simpler if the square root is approximated by

$$\sqrt{1 - k_x^2/k^2 - k_y^2/k^2} \approx 1 - \frac{k_x^2 + k_y^2}{2k^2}. \tag{235}$$

This approximation is referred to as the paraxial or Fresnel approximation. Using this approximation we can write

$$U(k_x, k_y, z) = F(k_x, k_y) e^{ikz} e^{-iz(k_x^2 + k_y^2/2k)}. \tag{236}$$

Assuming the paraxial approximation, the Fourier convolution theorem tells us that the field $u(x, y, z)$ can be written as the convolution of $f(x, y)$ with the inverse Fourier transform of $e^{ikz} e^{-iz(k_x^2 + k_y^2)/2k}$. The inverse Fourier transform of $e^{ikz} e^{-iz(k_x^2 + k_y^2)/2k}$ is $i(k/2\pi z) e^{ikz} e^{i(k/2z)(x^2 + y^2)}$. It follows that we can write

$$u(x, y, z) = i \frac{k}{2\pi z} e^{ikz} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\xi, \eta) e^{ik((x-\xi)^2 + (y-\eta)^2)/2z} \, d\xi \, d\eta. \tag{237}$$

It is often convenient to write this as

$$u(x, y, z) = \frac{ik}{2\pi z} e^{ikz} e^{ik(x^2 + y^2)/2z} \times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\xi, \eta) e^{ik(\xi^2 + \eta^2)/2z} e^{-ik((x\xi + y\eta)/z} \, d\xi \, d\eta. \tag{238}$$

This is usually referred to as the Fresnel approximation (16). This approximation can greatly simplify both analytical and numerical calculations.

C. Fourier Optics

We will now consider what happens when the beam passes through a lens of focal length f at the aperture. We claim that modifying the field at the aperture by the phase factor $e^{-ik(x^2 + y^2)/2f}$ is equivalent to passing the beam through a lens with focal length f. In order to see this note that if we had a beam of light coming in from infinity, then the field of the incoming light would be constant over the aperture.

$$f(x, y) = A e^{i\psi}. \tag{239}$$

At the plane $z = f$, the field would be given by

$$u(x, y, f) = \frac{ik}{2\pi f} Ae^{i\psi} e^{ikf} e^{ik(x^2+y^2)/2f} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-ik((x\xi+y\eta)/f} \, d\xi \, d\eta, \quad (240)$$

which can be written as

$$u(x, y, f) = \frac{i2\pi k}{f} e^{ikf} e^{ik(x^2+y^2)/2f} Ae^{i\psi} \delta(kx/f, ky/f). \quad (241)$$

This formula assumes that the aperture is infinitely large, and hence goes beyond the limits of validity of the Fresnel approximation. However, we could consider the case of an aperture of finite diameter, and we would get a more complicated but similar result, namely that the field at $z = f$ is all concentrated near the origin $(x, y) = (0, 0)$. This is exactly what a lens of focal length f would do to an incoming field of this sort.

We now consider the case where the incoming beam is not necessarily constant at the aperture, but is equal to $f(x, y)$. We assume that at the aperture $z = 0$ we have a lens with focal length f, which modifies the phase of the incoming beam by a factor $e^{-ik(x^2+y^2)/2f}$. In this case the output will be given by

$$u(x, y, z) = \frac{ik}{2\pi z} e^{ikz} e^{ik(x^2+y^2)/2z} \times \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\xi, \eta) e^{ik(\xi^2+\eta^2)/2z} e^{-ik(\xi^2+\eta^2)/2f} e^{-ik((x\xi+y\eta)/z} \, d\xi \, d\eta. \quad (242)$$

The output at the focal plane is given by

$$u(x, y, z) = \frac{ik}{2\pi f} e^{ikf} e^{ik(x^2+y^2)/2f} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\xi, \eta) e^{-ik((x\xi+y\eta)/f} \, d\xi \, d\eta, \quad (243)$$

which can be written as

$$\frac{ik}{2\pi f} e^{ikf} e^{ik(x^2+y^2)/2f} F(kx/f, ky/f). \quad (244)$$

where $F(\omega_x, \omega_y)$ is the Fourier transform of the function $f(x, y)$. We see that except for the term outside of the integral, the field distribution is given by the Fourier transform of the incoming field distribution. Note that the x and y dependence of the term outside of the integral has only a phase dependence. It follows that if we are only concerned with the irradiance distribution, then we can in fact ignore the terms outside of the integral.

D. Limits of Validity of the Fresnel Approximation

We now comment on the errors introduced by making the Fresnel approximation. We should emphasize that we are only considering the errors introduced in the problem of approximating the field $u(x, y, z)$ assuming we know the field at $z = 0$. In a real diffraction problem we do not know the field at $z = 0$, but approximate it as being the incoming wavefield.

We will now briefly summarize the conditions under which the Fresnel approximation can be assumed to be valid. In what follows R will be the effective dimension of the aperture and λ will be the wavelength of the light. We assume that the aperture lies in the plane $z = 0$, and that we are evaluating the field at a point $(d\cos(\theta), d\sin(\theta), z)$.

The Fresnel approximation always assumes that

$$R \gg \lambda. \tag{245}$$

Assuming that this restriction holds, the following is a summary of the conditions for the validity of the Fresnel approximation.

- The Fresnel approximation will be valid for all values of d if

$$N_F = \frac{2\pi R^2}{\lambda z} \text{ is not small} \tag{246}$$

- The amplitude of the wave predicted by the Fresnel approximation will be valid even if $N_F \ll 1$, provided

$$d/z \ll 1. \tag{247}$$

- Both the phase and amplitude predicted by the Fresnel approximation will be valid when $N_F \ll 1$ if

$$\frac{d^4}{z^4} \ll \frac{1}{kz}. \tag{248}$$

It should be noted that for the most part we are only concerned with the irradiance of the field, so the phase errors introduced by the Fresnel approximation for large values of z will not be important to us. For this reason we will be justified in using the Fresnel approximation provided $R \gg \lambda$, and that $d/z \ll 1$.

We will analyze the two-dimensional case where the aperture and field is independent of the y coordinate. The three-dimensional case is conceptually no more difficult, but the notation and the algebraic manipulations are simpler in two dimensions. We will assume that the incoming wave field is equal to $f(x)$ at the aperture $z = 0$. If $F(k_x)$ is the Fourier transform of

$f(x)$, then the field is given by

$$u(x,z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ik_x x} F(k_x) e^{ikz\sqrt{1-k_x^2/k^2}}\, dk_x. \tag{249}$$

In analyzing the Fresnel approximation we find it fruitful to consider a family of problems where the form of the function $f(x)$ stays the same, but the scaling of the function changes. In particular we will set

$$f(x) = g(x/R). \tag{250}$$

This includes the situation where the function $f(x)$ is equal to 1 inside the aperture, and 0 elsewhere. In this case the parameter R would be the characteristic dimension of the aperture. The Fourier transform of $f(x)$ can be written as

$$F(k_x) = RG(k_x R), \tag{251}$$

where $G(\alpha)$ is the Fourier transform of $g(x)$. The field for $z > 0$ can be written as

$$u(x,z) = R\,\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{ik_x x} G(Rk_x) e^{ikz\sqrt{1-k_x^2/k^2}}\, dk_x. \tag{252}$$

If we make the change of variables

$$\xi = Rk_x, \tag{253}$$

we can write this integral as

$$u(x,z) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{i(\xi x/R)} G(\xi) e^{ikz\sqrt{1-\xi^2/(Rk)^2}}\, d\xi. \tag{254}$$

If $g(\xi)$ is a well-behaved function, the Fourier transform $G(\xi)$ goes to zero as $|\xi| \to \infty$. It follows that our answers will not be very sensitive to how we approximate the term $\sqrt{1-\xi^2/(kR)^2}$ when ξ is large. This means that we only need to approximate this well for $\xi = O(1)$. We now make the approximation that

$$kR \gg 1. \tag{255}$$

This is the first approximation that will be made when doing the Fresnel approximation. This is equivalent to assuming that the aperture is much bigger than the wavelength, an assumption that will have to hold in order to carry out the general plan of diffraction theory. Under this assumption it is reasonable to expand the square root in a Taylor series:

$$kz\sqrt{1-\xi^2/(kR)^2} = kz\left(1 - \frac{\xi^2}{2(kR)^2} + \frac{\xi^4}{8(kR)^4} + \cdots\right). \tag{256}$$

If we ignore the third term and all of the remaining terms in the Taylor series, we will end up with the Fresnel approximation. We will now see when we can ignore these terms, and what sorts of errors we will make when we ignore them. Note that assuming that $kR \gg 1$ and $\xi = O(1)$, these terms will always be small compared to the first two terms. However, it is possible that if $kz \gg 1$, then they will not necessarily be small. For simplicity we will now keep the first three terms, and see when we can ignore the third term. The conditions for ignoring this term will be the same as for ignoring all of the remaining terms. If we keep the first three terms in the Taylor series we get

$$u(x,z) = \frac{1}{2\pi}\, e^{ikz} \int_{-\infty}^{\infty} G(\xi) e^{-i(1/N_F)(-\xi x^* + 1/2\xi^2 - (\xi^4/8(Rk)^2))}\, d\xi, \tag{257}$$

where

$$N_F = \frac{kR^2}{z}, \tag{258}$$

is known as the Fresnel number and

$$x^* = \frac{xkR}{z}. \tag{259}$$

We see that if

$$\frac{1}{N_F k^2 R^2} = \frac{z}{R}\frac{1}{k^3 R^3} \ll 1, \tag{260}$$

then we can ignore the third term in the Taylor series. This means that if we are close to the aperture then the Fresnel approximation will be valid (assuming $kR \gg 1$). In this case there is no restriction on the value of x. If $1/N_F$ is not large, then the Fresnel approximation will hold. This case is not that interesting, because it is essentially the case when the geometrical optics approximation holds and diffraction effects are unimportant.

We now consider the much more interesting case when N_F is small. In this case the phase in the integrand is multiplied by the large parameter $1/N_F$, and we can apply the method of stationary phase to the integral.

The phase will be stationary at the point ξ_0 satisfying

$$-x^* + \xi_0 - \frac{\xi_0^3}{2(kR)^2} = 0. \tag{261}$$

The method of stationary phase predicts that the field will be given by

$$u(x,z) \approx \frac{1}{2\pi}\, e^{ikz} e^{i\pi/4} \sqrt{\frac{2N_F\pi}{\psi_0''}}\, G(\xi_0) e^{-i(1/N_F)\psi_0}, \tag{262}$$

where

$$\psi_0 = -x^*\xi_0 + \frac{\xi_0^2}{2} - \frac{\xi_0^4}{8k^2R^2}, \tag{263}$$

and

$$\psi_0'' = 1 - \frac{3\xi_0^2}{2k^2R^2}. \tag{264}$$

This is the result predicted by the method of stationary phase assuming that $N_F \ll 1$ when we keep the first three terms in the Taylor series expansion of $\sqrt{1 - k_x^2/k^2}$. We would like to know how this compares to the answer we would get if we only kept the first two terms (the Fresnel approximation).

In the Fresnel approximation we would have

$$\xi_0 = x^*. \tag{265}$$

We will have been justified in ignoring the cubic term in the equation for ξ_0 provided

$$\frac{x^{*2}}{k^2R^2} \ll 1. \tag{266}$$

This is equivalent to requiring that

$$\frac{x^2}{z^2} \ll 1. \tag{267}$$

This means that the stationary point when we include the higher-order term will be nearly the same as the stationary point for the Fresnel approximation provided the opening angle from the midpoint of the aperture to the point (x, z) is small.

From the form of the answer in Eq. (262) we see that if we are not concerned with the phase errors, our answer will be accurate provided we have approximated ξ_0 well. This means that the amplitude of the Fresnel approximation will agree with the amplitude of the answer obtained by keeping three terms in the Taylor expansion provided $x^2/z^2 \ll 1$, and $N_F \ll 1$. However, in order for the phase of the answer predicted by the Fresnel approximation to agree with the more refined answer, it is necessary that we also approximate ψ_0/N_F well. The value of ψ_0/N_F predicted by the Fresnel approximation is

$$\psi_0/N_F \approx \frac{1}{N_F}(-x^*\xi_0 + \xi_0^2/2). \tag{268}$$

This will be a good approximation to the more refined answer provided that

$$\frac{1}{N_F}\frac{x^{*4}}{k^2R^2} \ll 1. \tag{269}$$

This can be written as

$$\frac{x^4}{z^4} \ll \frac{1}{kz}. \tag{270}$$

This agrees with the results we have already summarized concerning the errors in the Fresnel approximation.

E. The Vector Theory of Diffraction

The theory we have presented so far is limited to the scalar wave equation. In optics we are concerned with vector fields, the electric and magnetic fields. We begin by outlining the most naive, but nearly correct, approach to the vector theory. We know that each component of the electric and magnetic fields satisfies the scalar wave equation. Just as in the scalar theory we can assume that the field in the aperture is the same as the incoming field, and that the fields vanish elsewhere in the plane of the aperture. Using the scalar theory, we could compute each component of the electric and magnetic fields.

What are some possible difficulties with this approach? Just because each individual component of the field satisfies the wave equation does not mean that the vector field satisfies Maxwell's equations. If each component of the field were chosen exactly right at the plane of the aperture, then this would be the case. However, the assumption that we have made for the fields at the aperture are not necessarily consistant with the correct fields. For this reason we may end up getting inconsistant fields in the far field.

As an example of an inconsistancy, suppose that $z = 0$ is the plane of the aperture, and that the incoming field is a plane wave propagating in the z direction. The naive approach to vector difraction theory would imply that the z components of $\mathbf{E}$ and $\mathbf{B}$ vanish at the aperture, and hence vanish everywhere. A thorough analysis of this situation shows that the z components of the fields do not vanish identically.

This last example merely shows that the results of scalar diffraction theory cannot be exactly right. However, the theory was never intended to give exact answers. Just because the fields are inconsistent does not necessarily mean that they are worse approximations than a theory where the fields satisfy Maxwell's equations. However, the theory that takes into

account the vector nature of the fields is in fact more accurate for large angles.

A more consistent approach to the vector theory can be obtained by noting that it is not possible to arbitrarily specify all three components of **E** and **B** at the aperture. It is only necessary to specify the tangential electric fields at the aperture. We now argue that once these fields are known, we know E_x and E_y for $z > 0$, and we can then determine E_z and **B**.

Clearly both E_x and E_y satisfy the scalar wave equation. It follows that if we know these components at $z = 0$ then we can determine them everywhere for $z > 0$. Once we know E_x and E_y we can use the equation

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0, \tag{271}$$

to determine E_z up to an arbitrary additive function $f(x, y)$. Assuming that we have a finite sized aperture, the field **E**, and in particular the function E_z, must approach zero as $z \to \infty$. This fact allows us to determine this arbitrary function $f(x, y)$. It follows that we can determine E_z. We can now determine **B** by taking the curl of **E** and using Faraday's law. It follows that we can determine all the components of both **E** and **B** once we specify the tangential components of **E** at the aperture.

The vector theory of diffraction (17) approximates the tangential components of the electric field using scalar diffraction, but then computes the z component based on these fields. We will restrict ourselves to the case where the incoming wave has no z component of the electric field. In this case the scalar theory of diffraction predicts that the diffracted field will also have no z component of the electric field. We will now show that in this situation the z component of the electric field can be ignored provided we are only interested in small angles, a condition that we have already assumed in making the Fresnel approximation.

Suppose that at the plane of the aperture the tangential components of the electric field are given by

$$(E_x(x, y, 0), E_y(x, y, 0)) = (g_x(x/R, y/R), g_y(x/R, y/R)). \tag{272}$$

The x and y components of the electric field each satisfy the scalar wave equation. By Fourier transforming the wave equation we can conclude that

$$E_x(x, y, z) = \frac{1}{4\pi^2} R^2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{i(k_x x + k_y y)} G_x(k_x R, k_y R) \times e^{iz\sqrt{k^2 - k_x^2 - k_y^2}}\, dk_x\, dk_y, \tag{273}$$

and

$$E_y(x,y,z) = \frac{1}{4\pi^2} R^2 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{i(k_x x + k_y y)} G_y(k_x R, k_y R)) \times e^{iz\sqrt{k^2-k_x^2-k_y^2}}\, dk_x\, dk_y, \tag{274}$$

where $G_x(k_x, k_y)$ and $G_y(k_x, k_y)$ are the Fourier transforms of $g_x(x, y)$ and $g_y(x, y)$.

Using the fact that $\nabla \cdot \mathbf{E} = 0$, we can write the field E_z as

$$E_z(x,y,z) = -\frac{1}{4\pi^2} R^2 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{i(k_x x + k_y y)} \Gamma(k_x, k_y) e^{iz\sqrt{k^2-k_x^2-k_y^2}}\, dk_x\, dk_y, \tag{275}$$

where

$$\Gamma(k_x, k_y) = \frac{1}{\sqrt{k^2 - k_x^2 - k_y^2}} \left(k_x G_y(k_x R, k_y R) + k_y G_y(k_x R, k_y R)\right). \tag{276}$$

These are exact expressions assuming that we know the tangential electric field at the plane of the aperture.

The expression for E_z is very similar to the expressions for E_x and E_y except that it has $k_x/\sqrt{k^2 - k_x^2 - k_y^2}$ multiplying G_x, and $k_y/\sqrt{k^2 - k_x^2 - k_y^2}$ multiplying G_y. Under the conditions for the Fresnel approximation we can make the approximation

$$\frac{k_x}{\sqrt{k^2 - k_x^2 - k_y^2}} = \frac{k_x}{k}, \tag{277}$$

and

$$\frac{k_y}{\sqrt{k^2 - k_x^2 - k_y^2}} = \frac{k_y}{k}. \tag{278}$$

The Fresnel approximation is based on the assumption that k_x/k and k_y/k are both small in the region of interest. It follows that the factors multiplying G_x and G_y will always make the term E_z negligible compared to E_x and E_y.

For example, if the Fresnel number is small, then we can evaluate these integrals using the method of stationary phase. We could put these integrals in dimensionless form and arrange things so that there was a large

parameter multiplying the phase. However, we can take a short cut and note that in the Fresnel approximation, the phases of the integrands are given by

$$\phi = k_x x + k_y y - z\,\frac{k_x^2 + k_y^2}{2k}. \tag{279}$$

The phase will be stationary when

$$x - k_x z/k = 0, \tag{280}$$

and

$$y - k_y z/k = 0. \tag{281}$$

This shows that when we apply the method of stationary phase, the z component of the electric field can be related to the other two components by

$$E_z(x, y, z) \approx \frac{x}{z}\,E_x(x, y, z) + \frac{y}{z}\,E_y(x, y, z). \tag{282}$$

This shows that provided $|x/z| \ll 1$, and $|y/z| \ll 1$, the z component of the electric field will be negligible compared to the tangential components. This was based on the assumption that the Fresnel number was small. If the Fresnel number is order 1, we can show that the z component will be small provided only that $kR \ll 1$.

VII. GEOMETRICAL THEORY OF BEAM SHAPING

A. One-dimensional Theory

In this section we present a theory of beam shaping based on geometrical optics. Special cases of this theory may be found in the literature on geometrical beam shaping (18). The theory we present is not the most general one using geometrical optics since we assume that the rays are moved around continuously, and in a very orderly manner. In the geometrical optics limit it is possible to accomplish the same goal by moving the rays around in a discontinuous and less orderly manner, but when we analyze beam shaping using diffraction theory we will see that this is very undesirable. We believe that it is very difficult to improve on a beam shaping system designed by the techniques described in this section. However, some systems designed this way will work very well, while others will work very poorly. One must go beyond the geometrical theory and use diffraction theory in order to understand why this is so. That will be the subject of the next section.

We begin by considering the beam shaping problem in one dimension. This theory is directly applicable to cases where the incoming beam has an irradiance distribution that is the direct product of two one-dimensional

distributions. A two-dimensional function $f(x, y)$ is the direct product of one-dimensional functions if

$$f(x, y) = f_1(x) f_2(y). \tag{283}$$

If both the input and the desired output can be written as a direct product, then the problem can be decomposed into two one-dimensional beam shaping problems. This is the case when we try to turn a Gaussian beam into a rectangular flat top beam.

We suppose that an incoming parallel beam of light has an irradiance distribution of $I(x/R)$, and at the plane $z = 0$ the beam passes through a phase element that refracts the beam. We would like to determine the phase element such that the irradiance distribution at the plane $z = f$ is given by $A(R/D)Q(x/D)$, where A is a constant chosen so that the energy of our light beam is conserved.

In our analysis we assume that the aperture contains a lens of focal length f, plus an additional optical element that allows us to shape the beam. In practice these two optical elements can be combined into a single optical element, but this may not be a desirable feature if one wants to use the same element to shape the beam at several different focal planes. We suppose that our beam shaping element introduces a phase shift of $(RD/fc)\phi(x/R)$ at the plane $z = 0$ (c is the speed of light). The goal of our analysis is to determine the function ϕ such that the beam at the plane $z = f$ has the desired shape. This analysis is carried out in three steps.

- Determine the constant A that determines the irradiance of the output beam. This is accomplished by requiring that the total energy of the output beam is the same as the energy of the incoming beam.
- Determine a function that maps rays at the plane of the aperture into rays at the focal plane. In particular, we determine a function $\alpha(\xi)$ such that a ray that passes through the aperture at $x = R\xi$ passes through the focal plane at $x = D\alpha(\xi)$. This step can be carried out by requiring that the energy of any bundle of rays that enters the aperture is the same as the energy of the same bundle of rays as they pass through the focal plane.
- Determine the function $\phi(\xi)$ that gives us the phase shift introduced by our beam shaping element. Once we know the function $\alpha(\xi)$ this step can be carried out by requiring that the time for a ray to get from $z = -\infty$ to the focal plane is consistent with Fermat's principle.

At this point the reader may feel annoyed by our introduction of the lengths R and D. For example, it would be simpler if we said that the input

beam had the irradiance $I(x)$ rather than $I(x/R)$. However, the lengths R and D have been included in the definition of our irradiance profiles, our normalization constant A, and our phase shift ϕ in order to bring out certain scaling properties of beam shaping. These scaling properties will be especially important in the next section when we discuss diffraction effects.

In order to carry out the first step in this process we note that the energy of the incoming beam can be written as

$$E_{in} = \int_{-\infty}^{\infty} I(s/R)\, ds = R \int_{-\infty}^{\infty} I(s)\, ds. \tag{284}$$

The energy of the outgoing beam can be written as

$$E_{out} = \frac{AR}{D} \int_{-\infty}^{\infty} Q(s/D)\, ds = AR \int_{-\infty}^{\infty} Q(s)\, ds. \tag{285}$$

If we equate these two expressions we arrive at the result

$$A = \frac{\displaystyle\int_{-\infty}^{\infty} I(s)\, ds}{\displaystyle\int_{-\infty}^{\infty} Q(s)\, ds}. \tag{286}$$

We have accomplished the first of our three steps. We now determine the function $\alpha(\xi)$ using the conservation of energy.

$$\int_{-\infty}^{R\xi} I(s/R)\, ds = A\, \frac{R}{D} \int_{-\infty}^{D\alpha(\xi)} Q(s/D)\, ds. \tag{287}$$

This is a mathematical statement of the fact that the energy of all of the rays with initial x coordinates less than $R\xi$ must have the same energy as all of the rays at the focal plane that have x coordinate less than $D\alpha(\xi)$. A simple change of variables gives us the equation

$$\int_{-\infty}^{\xi} I(s)\, ds = A \int_{-\infty}^{\alpha(\xi)} Q(s)\, ds. \tag{288}$$

As long as the functions $I(s)$ and $Q(s)$ are both positive, it is clear that the function $\alpha(\xi)$ is uniquely determined by this equation. This follows from the fact that for a given value of ξ we can increase the value α until the integral on the right equals the integral on the left. Since $Q(\xi) > 0$, it is clear that for any value of ξ there is only one value of α such that the two integrals will be equal.

The functions $I(s)$ and $Q(s)$ are both non-negative, but it is possible that they could vanish on certain intervals. This would be the case if we were trying to transform a beam into a beam that had a core of zero irradiance

(such as an annulus). In this case we could have a whole interval of points α that are assigned to the same point ξ. This degenerate case can be thought of as a limiting case of when the functions $I(s)$ and $Q(s)$ are both positive.

Equation (288) determines the functions $\alpha(\xi)$. However, there are a few motivations for differentiating this equation to get

$$AQ(\alpha)\frac{d\alpha}{d\xi} = I(\xi). \tag{289}$$

This gives us a differential equation for the function $\alpha(\xi)$. One way of solving this differential equation is to integrate this equation once to get back to Eq. (288). However, if one needs to solve the equation numerically, it may be more convenient to solve the differential equation than to solve Eq. (288). Another motivation for writing down the differential equation is that when we make the stationary phase approximation to diffraction theory we end up with this differential equation. Yet another motivation comes from the fact when we consider problems that are neither one dimensional nor radially summetric we must revert to a differential equation that is analogous to (289).

In the energy equation (288) we have assumed that the orientation of the incoming rays is the same as the orientation of the rays at the focal plane $z = f$. By this we mean that incoming rays with $\xi \ll 0$ get mapped into rays with $\alpha \ll 0$ at the focal plane, and incoming rays with $\xi \gg 0$ get mapped into rays with $\alpha \gg 0$ at the focal plane. It is possible to reverse the orientation of the rays so that incoming rays with $\xi \gg 0$ end up at the focal plane with $\alpha \ll 0$, and vice versa. In this case the energy equation can be written as

$$\int_{-\infty}^{R\xi} I(s/R)\,ds = A\frac{R}{D}\int_{D\alpha(\xi)}^{\infty} Q(s/D)\,ds. \tag{290}$$

Changing variables in the integrals gives us the equation

$$\int_{-\infty}^{\xi} I(s)\,ds = A\int_{\alpha(\xi)}^{\infty} Q(s)\,ds. \tag{291}$$

If we differentiate this equation we get

$$AQ(\alpha)\frac{d\alpha}{d\xi} = -I(\xi). \tag{292}$$

These two solutions will give identical irradiance distributions as long as we evaluate the irradiance at the plane $z = f$. However, as we move away from the plane $z = f$ these two solutions have very different properties. When we apply the method of stationary phase, the two different types of

solutions appear by choosing different signs of the phase function. These will also be discussed in Sec. IV.E of Chapter 3.

The solutions derived using Eq. (288) or Eq. (291) are the only ones that allow us to shape the beam so that the rays are moved around in a continuous fashion, and so that the function $\xi(\alpha)$ (the inverse of $\alpha(\xi)$) is single valued. When we study the effects of diffraction we will see that beam shaping systems that do not satisfy these requirements will suffer much more from the effects of diffraction than ones that do.

We have now completed the first two steps in our analysis, and we are ready to determine the function $\phi(\xi)$. We assume that the rays that enter the aperture are coming in parallel. For our purposes it is simpler to assume that rays are coming from a distant point source at $(0, -L)$, and we will then let $L \to \infty$. The travel time for a ray to get from the point source to a point $(D\alpha, f)$ consists of three parts:

- The time $t_L(\xi)$ to get from the source at $(0, -L)$ to a point $(R\xi, 0)$ on the aperture;
- The time $t_{\text{delay}}(\xi)$ that it takes to get through the Fourier transform lense, and the beam shaping element at $(R\xi, 0)$;
- The time $t_f(\xi, \alpha)$ that it takes to get from a point $(R\xi, 0)$ on the aperture, to a point $(D\alpha, f)$ at the focal plane.

The total travel time is given by

$$t(\xi, \alpha) = t_L(\xi) + t_{\text{delay}}(\xi) + t_f(\xi, \alpha). \tag{293}$$

Fermat's principle requires that the travel time of a ray that starts out at $(0, -L)$, passes through the aperture at $(R\xi, 0)$, and ends up at $(D\alpha, f)$ must be stationary. This means that it must be stationary compared with the travel time of any nearby ray. In particular it will be stationary with respect to the travel time of a ray that goes from $(0, -L)$, passes through the aperture at $(R\xi + R\,d\xi, 0)$, and then goes straight to the point $(D\alpha, f)$. In order for this to be so we must have

$$\frac{\partial t(\xi, \alpha)}{\partial \xi} = 0. \tag{294}$$

We will now see that this equation allows us to determine $\phi(\xi)$.

The travel time t_L is given by

$$t_L(\xi) = \frac{1}{c}\sqrt{L^2 + \xi^2} \approx \frac{1}{c}\left(L + \frac{\xi^2}{2L}\right). \tag{295}$$

In the limit as $L \to \infty$ we end up with the equation

$$\frac{\partial}{\partial \xi} t_L(\xi) = 0. \tag{296}$$

The travel time $t_{\text{delay}}(\xi)$ is given by

$$t_{\text{delay}}(\xi) = -\xi^2 \frac{R^2}{2fc} + \phi(\xi) \frac{RD}{fc}. \tag{297}$$

The first term on the right gives the time delay introduced by the transform lens, and the second term gives the time delay introduced by the beam shaping element.

Taking the derivative of this we get

$$\frac{\partial}{\partial \xi} t_{\text{delay}}(\xi) = -\xi \frac{R^2}{fc} + \frac{RD}{fc} \frac{\partial}{\partial \xi} \phi(\xi). \tag{298}$$

The travel time $t_f(\xi, \alpha)$ is given by

$$t_f(\xi, \alpha) = \frac{1}{c} \sqrt{f^2 + (R\xi - D\alpha)^2}. \tag{299}$$

The paraxial approximation assumes that $D^2\alpha^2/f^2 \ll 1$, and $R^2\xi^2/f^2 \ll 1$, so that we can make the approximation

$$\sqrt{(\epsilon^2 + f^2)} \approx f + \frac{\epsilon^2}{2f}. \tag{300}$$

In this approximation we get

$$t_f(\xi, \alpha) = \frac{f}{c} + \frac{(D\alpha - R\xi)^2}{2fc}, \tag{301}$$

and hence

$$\frac{\partial}{\partial \xi} t_f(\xi, \alpha) = \frac{R}{fc} (R\xi - D\alpha). \tag{302}$$

Combining our expressions for $\partial/\partial\xi(t_L + t_{\text{delay}} + t_f)$ we end up with the very simple equation

$$\frac{d\phi}{d\xi} = \alpha(\xi). \tag{303}$$

Assuming we know the function $\alpha(\xi)$, the function $\phi(\xi)$ can be determined by quadrature.

We now collect our beam shaping equations into a single set of equations. Given the functions $I(s)$ and $Q(s)$, the phase function $\phi(\xi)$ is

determined by first calculating the constant A

$$A = \frac{\int_{-\infty}^{\infty} I(s)\,ds}{\int_{-\infty}^{\infty} Q(s)\,ds}, \tag{304a}$$

then solving the differential equation

$$AQ(\alpha)\frac{d\alpha}{d\xi} = \pm I(\xi) \tag{304b}$$

in order to determine $\alpha(\xi)$. The sign in this equation depends on whether or not we have reversed the orientation of the rays or not. Finally the function $\phi(\xi)$ is obtained by solving the differential equation

$$\frac{d\phi}{d\xi} = \alpha(\xi). \tag{304c}$$

A very simple scaling property of these equations will now be pointed out. If we determine a beam shaping system for the lengths D and f, then we can use the same phase function $(RD/fc)\phi(\xi)$ for a new beam shaping system with lengths D_1 and f_1, provided $D_1/f_1 = D/f$. This means that we can change the scale of our system by merely using a different quadratic lens, without changing the optical element determined by ϕ. This follows from the fact that the function $\alpha(\xi)$ is independent of the D, f, and R. It follows that the function $\phi(\xi)$ is also independent of these quantities. Clearly the function $(RD/fc)\phi(\xi)$ will not change as long as we keep the ration D/f fixed.

B. Direct Product Distributions

We would once again like to emphasize that the theory of the last section can be applied when both the input and the desired output can be written as direct products. That is, we can use the theory of the last section if we can write

$$I(x, y) = I_1(x)I_2(y),$$

and

$$Q(x, y) = Q_1(x)Q_2(y).$$

In this case the phase function of the beamshaping element can also be written as a direct product.

$$\phi(x, y) = \phi_1(x)\phi_2(y).$$

One very important example of this is when the input is a circular Gaussian,

$$I(x, y) = e^{-(x^2+y^2)/2}$$

and the output is a rectangular flat-top beam:

$$Q(x, y) = \text{Rect}(x/A)\,\text{Rect}(y/B)$$

where

$$\text{Rect}(x) = 1 \quad |x| \leqslant 1$$

$$\text{Rect}(x) = 0 \quad |x| > 1$$

C. Radially Symmetric Problems

We now derive a geometrical theory of beam shaping that applies when we are trying to convert a radially symmetric beam with irradiance profile $I(r/R)$ into a radially symmetric beam with irradiance profile that is proportional to $Q(r/D)$. We assume that the desired output beam has the irradiance $(AR^2/D^2)Q(r/D)$. As in the one-dimensional case we begin by computing the normalization constant A. The total energy of the incoming beam is given by

$$E_{\text{in}} = 2\pi \int_0^\infty I(s/R)s\,ds = 2\pi R^2 \int_0^\infty sI(s)\,ds. \tag{305}$$

The energy of the output beam is given by

$$E_{\text{out}} = \frac{AR^2}{D^2} \int_0^\infty sQ(s/D)\,ds = AR^2 \int_0^\infty sQ(s)\,ds. \tag{306}$$

If we require that the energy of the incoming beam is the same as the outgoing beam we must have

$$A = \frac{\displaystyle\int_0^\infty sI(s)\,ds}{\displaystyle\int_0^\infty sQ(s)\,ds}. \tag{307}$$

We now determine the function $\alpha(\xi)$ such that a ray that encounter our optical element at $(R\xi, 0)$ ends up at $(D\alpha, f)$.

The conservation of energy now implies that

$$\int_{R\xi}^\infty I(s/R)s\,ds = \frac{AR^2}{D^2} \int_{D\alpha(\xi)}^\infty sQ(s/D)\,ds. \tag{308}$$

This equation is a mathematical statement of the fact that the energy of the rays that encounter the plane $z = 0$ with $r > R\xi$ is the same as the energy of the rays that encounter the focal plane with $r > D\alpha(\xi)$.

A simple change of variables gives us the equations

$$\int_{\xi}^{\infty} I(s)s\,dz = A\int_{\alpha(\xi)}^{\infty} sQ(s)\,ds. \tag{309}$$

Just as in the one-dimensional case we can argue that the Eq. (309) uniquely determines the function $\alpha(\xi)$. As in the one-dimensional case it may be convenient to differentiate this equation to get a differential equation for $\alpha(\xi)$.

$$A\alpha Q(\alpha)\,\frac{d\alpha}{d\xi} = \xi I(\xi). \tag{310}$$

This equation assumes that the ray that starts at the axis of symmetry ends up at the axis of symmetry at $z = f$. In analogy to the one-dimensional case we could also consider the case where the ray that started on the axis is sent out infinitely far from the axis when $z = f$. We could devise an optical element that did this, but it would necessarily be quite degenerate and suffer from diffraction effects.

Now that we know the function $\alpha(\xi)$ we can use Fermat's principle to determine the optical thickness $\phi(r/R)$ that can actually accomplish this beam shaping.

Once again, let $-r^2/2fc + RD\phi(r/R)/fc$ be the time delay introduced by our optical element, and $z = f$ be the imaging plane. Fermat's principle requires that

$$\frac{d\phi}{\partial\xi} = \alpha(\xi). \tag{311}$$

This is exactly the same equation we used in the one-dimensional case. Since we know the function $\alpha(\xi)$, we can determine the function $\phi(\xi)$ by quadrature.

Once again we can argue that the function ϕ is independent of the parameters D and f, and hence the time delay $(RD/fc)\phi(r/R)$ depends on D and f only through the ratio D/f.

D. More General Distributions

So far we have considered one-dimensional (applicable to direct product profiles) and radially symmetric beam shaping. In this section we outline how one would determine an optical element that turns an incoming

irradiance profile $I(x/R, y/R)$ into an irradiance distribution that is proportional to $Q(x/D, y/D)$ at the image plane f.

The solution to this problem is much more difficult than the ones we have already encountered. We do not have any first hand experience in actually doing this, but feel that it is worth writing down the equations that would allow one to solve this problem.

We begin by assuming that the irradiance distribution at the focal plane f is equal to $(AR^2/D^2)Q(x/D, y/D)$. In order for the energy of input beam to be the same as the output beam we must have

$$A = \frac{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(s,t)\,ds\,dt}{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} Q(s,t)\,ds\,dt}. \tag{312}$$

We now write down an equation for the conservation of energy of any bundle of rays. Suppose rays that encounter the optical element at $(s, t, 0)$ end up at $(x(s,t), y(s,t), f)$. In order to conserve energy we must have

$$I(s,t) = \pm AQ(x(s,t), y(s,t))J(s,t), \tag{313a}$$

where

$$J(s,t) = \frac{\partial x}{\partial s}\frac{\partial y}{\partial t} - \frac{\partial x}{\partial t}\frac{\partial y}{\partial s}. \tag{313b}$$

This is the generalization of the differential form of the energy equations that we have written down previously. It can be justified by noting that the rays in the area $s < x < s + ds$, $t < y < t + dt$ get mapped into a region with area $J(s,t)\,ds\,dt$ at the focal plane.

If the time delay produced by our beam shaping element is given by $(RD/fc)\phi(x/R, y/R)$, then Fermat's principle shows us that the function $\phi(s,t)$ must satisfy

$$\frac{\partial \phi}{\partial s} = x(s,t), \tag{314a}$$

$$\frac{\partial \phi}{\partial t} = y(s,t). \tag{314b}$$

These two equations can be derived almost identically to the one-dimensional and radial cases. We need two equations because we need to guarantee that the path is stationary with respect to changes in both the x and y directions. Using this last set of equations we can write our energy

equation as

$$I(s,t) = \pm AQ\left(\frac{\partial\phi}{\partial s}, \frac{\partial\phi}{\partial t}\right)\left(\frac{\partial^2\phi}{\partial s^2}\frac{\partial^2\phi}{\partial t^2} - \left(\frac{\partial^2\phi}{\partial s\,\partial t}\right)^2\right). \tag{315}$$

This is a nonlinear partial differential equation for the function $\phi(s,t)$. For the special cases where the profiles are radially symmetric, or can be written as direct products, we end up with our previous results. In general it is not clear that this equation is enough to determine the function $\phi(s,t)$. In order to get a feel for this equation we consider a linearized version of this equation. We will see that the linearized equations end up giving us an equation that is very similar to Poisson's equation. We will see that the linearized equations give us a well-posed mathematical problem, indicating that the same will likely be true of the full non-linear equations.

In order to get a linearized system of equations we suppose that the function $I(s,t)$ is almost identical to the function $Q(x,y)$. This would imply that the function $(x(s,t), y(s,t))$ is very nearly equal to (s,t), and hence

$$\phi(s,t) \approx \frac{s^2+t^2}{2}. \tag{316}$$

This means that the function ϕ is merely reversing the phase difference caused by the lens that focuses the beam at $z = f$. We will now assume that

$$Q(x,y) = I(x,y) + \delta P(x,y), \tag{317}$$

where δ is a very small number. We also assume that

$$A = 1 + \delta a, \tag{318}$$

and

$$\phi(s,t) = \frac{s^2+t^2}{2} + \delta\psi(s,t). \tag{319}$$

To first order in δ we can write

$$Q\left(\frac{\partial\phi}{\partial s}, \frac{\partial\phi}{\partial t}\right) = I(s,t) + \delta(P(s,t) + \nabla\psi(s,t)\cdot\nabla I(s,t)), \tag{320}$$

$$\frac{\partial^2\phi}{\partial s^2}\frac{\partial^2\phi}{\partial t^2} - \left(\frac{\partial^2\phi}{\partial s\,\partial t}\right)^2 = 1 + \delta\left(\frac{\partial^2\psi}{\partial s^2} + \frac{\partial^2\psi}{\partial t^2}\right). \tag{321}$$

If we expand Eq. (315) to first order in δ we end up with the equation

$$P(s,t) + aI(s,t) + I(s,t)\left(\frac{\partial^2\psi}{\partial s^2} + \frac{\partial^2\psi}{\partial t^2}\right) + \nabla\psi\cdot\nabla I = 0, \tag{322}$$

which can be written as

$$\nabla \cdot (I(s,t)\nabla\psi) = -P(s,t) - aI(s,t). \tag{323}$$

If we integrate these equations over the xy plane, we find that the left-hand side vanishes (assuming ψ vanishes at ∞), and hence the constant a must be chosen so that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(s,t)\,ds\,dt + a\int_{-\infty}^{\infty} I(s,t)\,ds\,dt = 0. \tag{324}$$

Once we have chosen a in this way, we can uniquely solve for ψ if we require that ψ vanishes at ∞.

The fact that we can solve the linearized equations is an excellent sign that the nonlinear equations (315) will uniquely determine the function ϕ.

E. Examples

We will now present some concrete examples from the geometrical theory of beam shaping. Some of these examples are important for their own sake, but other examples are presented to illustrate some of the difficulties that can arise when applying the geometrical theory. The difficulties will not appear until we analyze them using diffraction theory.

Example 9 *Turning a Gaussian into a Flat-top Beam—I* Let

$$I(s) = e^{-s^2}, \tag{325}$$

and

$$Q(s) = 1 \text{ for } |s| < 1, \tag{326}$$

$$Q(s) = 0 \text{ for } |s| > 1. \tag{327}$$

The normalization of the energy requires that

$$\int_{-\infty}^{\infty} e^{-s^2}\,ds = 2A, \tag{328}$$

or

$$A = \frac{\sqrt{\pi}}{2}. \tag{329}$$

The function $\alpha(\xi)$ must satisfy

$$Q(\alpha)\,\frac{d\alpha}{d\xi} = \frac{2}{\sqrt{\pi}}\,e^{-\xi^2}. \tag{330}$$

As long as $|\alpha| < 1$ this can be written as

$$\frac{d\alpha}{d\xi} = \frac{2}{\sqrt{\pi}}\, e^{-\xi^2}. \tag{331}$$

The solution to this equation can be written as

$$\alpha(\xi) = \text{erf}(\xi), \tag{332}$$

where

$$\text{erf}(\xi) = \frac{2}{\sqrt{\pi}} \int_0^{\xi} e^{-s^2}\, ds. \tag{333}$$

Since $|\alpha| < 1$ for $-\infty < \xi < \infty$, we conclude that we do not need to consider the case where $Q(\alpha) = 0$.

We now use the equation

$$\frac{d\phi}{d\xi} = \text{erf}(\xi), \tag{334}$$

to find the solution

$$\phi(\xi) = \frac{2}{\sqrt{\pi}} \left(\xi\, \frac{\sqrt{\pi}}{2}\, \text{erf}(\xi) + \tfrac{1}{2} e^{-\xi^2} - \tfrac{1}{2} \right). \tag{335}$$

This example has been presented without any reference to the scalings R and D. If we were trying to turn a beam with the initial distribution $I(x/R)$ into a beam with distribution $Q(x/D)$ at the focal plane f, then our beam shaping element would need to introduce a phase delay of $(RD/fc)\phi(x/R)$.

Example 10 *Turning a Gaussian into a Flat-top Beam—II* We consider the same problem as in the previous example. However, this time we present a solution that reverses the order of the rays.

The function $\alpha(\xi)$ must satisfy

$$Q(\alpha)\, \frac{d\alpha}{d\xi} = -\frac{2}{\sqrt{\pi}}\, e^{-\xi^2}. \tag{336}$$

As long as $|\alpha| < 1$ this can be written as

$$\frac{d\alpha}{d\xi} = -\frac{2}{\sqrt{\pi}}\, e^{-\xi^2}. \tag{337}$$

The solution to this equation can be written as

$$\alpha(\xi) = -\text{erf}(\xi). \tag{338}$$

We now use the equation

$$\frac{d\phi}{d\xi} = -\text{erf}(\xi), \tag{339}$$

to find the solution

$$\phi(\xi) = -\frac{2}{\sqrt{\pi}}\left(\xi\,\frac{\sqrt{\pi}}{2}\,\text{erf}(\xi) + \tfrac{1}{2}e^{-\xi^2} - \tfrac{1}{2}\right). \tag{340}$$

Example 11 *Turning a Radial Gaussian into a Radial Flat top* We now consider the problem of turning a radial Gaussian into a radial flat top. In particular suppose $I(s) = e^{-s^2}$, and

$$Q(s) = 1 \text{ if } s < 1, \tag{341}$$

$$Q(s) = 0 \text{ if } s > 1. \tag{342}$$

In this case we must choose the constant A so that

$$A\pi = \int_0^\infty s e^{-s^2}\, ds. \tag{343}$$

It follows that

$$A = \frac{1}{2\pi}. \tag{344}$$

Equation (310) implies

$$\alpha\,\frac{d\alpha}{d\xi} = 2\pi\xi e^{-\xi^2}. \tag{345}$$

If we require that $\alpha(0) = 0$, this equation implies that

$$\alpha(\xi) = \sqrt{2\pi}\sqrt{1 - e^{-\xi^2}}. \tag{346}$$

Equation (311) for ϕ now implies that

$$\phi(\xi) = \sqrt{2\pi}\int_0^\xi \sqrt{1 - e^{-s^2}}\, ds. \tag{347}$$

Example 12 *Turning a Gaussian into a Stairstep* We consider the case where the input beam is a Gaussian

$$I(s) = e^{-s^2}, \tag{348}$$

and the desired output beam is a stair step function.

$$Q(s) = \gamma|s| < \alpha_0, \tag{349}$$

$$Q(s) = 1 \ \alpha_0 < |s| < 1, \tag{350}$$

$$Q(s) = 0 \ |s| > 1. \tag{351}$$

This situation clearly is symmetrical, so that the phase function $\phi(\xi) = \phi(-\xi)$, and $\alpha(\xi) = -\alpha(-\xi)$. For this reason we will only concern ourselves with finding ϕ and α for $\xi > 0$.

The normalization condition requires that

$$A = \frac{\sqrt{\pi}}{2\alpha_0(\gamma - 1) + 2}. \tag{352}$$

There will be a point ξ_0 that separates the rays that get sent into the first step from those that get sent into the second step. We do not know this point ahead of time, but must calculate its value given the parmeters γ and α_0. The function $\alpha(\xi)$ must satisfy

$$\frac{d\alpha}{d\xi} = \frac{1}{A\gamma} e^{-\xi^2} \text{ for } \xi < \xi_0. \tag{353}$$

This equation is valid for $\alpha < \alpha_0$. We also have

$$\frac{d\alpha}{d\xi} = \frac{1}{A} e^{-\xi^2} \text{ for } \xi > \xi_0 \tag{354}$$

This equation is valid for $\alpha_0 < \alpha < 1$.

The first of these equations can be integrated from 0 to ξ_0 to give

$$\text{erf}(\xi_0) = \frac{2}{\sqrt{\pi}} A\gamma\alpha_0. \tag{355}$$

This is not an explicit expression for ξ_0, but it can very quickly be determined using an iterative method such as Newton's method. Once we have determined ξ_0 and A we have explicit expressions for $\alpha(\xi)$. We can now determine the function $\phi(\xi)$ by solving the equation

$$\frac{d^2\phi}{d\xi^2} = \frac{1}{A\gamma} e^{-\xi^2} \text{ for } \xi < \xi_0, \tag{356}$$

$$\frac{d^2\phi}{d\xi^2} = \frac{1}{A} e^{-\xi^2} \text{ for } \xi > \xi_0, \tag{357}$$

along with the requirements

$$\phi(0) = 0, \tag{358}$$

and the requirement that ϕ and its derivative are continuous at ξ_0. These equations are almost identical for those of turning a Gaussian into a flat top. Let $\phi_0(\xi)$ be given by

$$\phi_0(\xi) = \xi \frac{\sqrt{\pi}}{2} \operatorname{erf}(\xi) + \tfrac{1}{2} e^{-\xi^2} - \tfrac{1}{2}. \tag{359a}$$

Then the phase function for the stair step can be written as

$$\phi(\xi) = \frac{1}{A\gamma} \phi_0(\xi) \text{ for } \xi < \xi_0, \tag{359b}$$

and

$$\phi(\xi) = \frac{1}{A\gamma} \left(\gamma\phi_0(\xi) + (1-\gamma)\phi_0(\xi_0) + (\xi - \xi_0)(1-\gamma)\operatorname{erf}(\xi_0) \frac{\sqrt{\pi}}{2} \right)$$
$$\text{for } \xi \geqslant \xi_0. \tag{359c}$$

Example 13 *Numerical Solutions for Symmetrical Profiles* There are many situations where it is very cumbersome, or impossible to obtain closed form analytical solutions for the function $\phi(\xi)$. However, it is not difficult to write a computer code that solves for ϕ. We now consider how to write a code for the special case where both $I(s)$ and $Q(s)$ are symmetric with respect to reflections in s. That is

$$I(s) = I(-s), \tag{360a}$$

and

$$Q(s) = Q(-s). \tag{360b}$$

In this case we can argue that

$$\alpha(-\xi) = -\alpha(\xi), \tag{361a}$$

and

$$\phi(-\xi) = \phi(\xi). \tag{361b}$$

This means that we can solve for α and ϕ on the interval $\xi > 0$, and this will allow us to determine these functions everywhere.

We now outline how one can use an ODE (ordinary differental equation) solver to determine the function ϕ, given the functions $Q(s)$ and $I(s)$. In order to do this we first determine the constant A.

$$A = \frac{\displaystyle\int_{-\infty}^{\infty} I(s)\, ds}{\displaystyle\int_{-\infty}^{\infty} Q(s)\, ds}. \tag{362}$$

In many situations, this constant can be determine analytically, even when the function $\phi(\xi)$ cannot. In these situations, one can analytically compute A. In general, one can use the ODE solver to compute the integrals in both the numerator and the denominator. Once the constant A has been determined, we use the ODE solver to solve the following intial value problem.

$$\frac{d\alpha}{d\xi} = \pm \frac{1}{AQ(\alpha)} I(\xi), \tag{363a}$$

$$\frac{d\phi}{d\xi} = \alpha(\xi). \tag{363b}$$

Either sign can be taken in the first of these equations. As we have already mentioned, each sign corresponds to a physically different solution.

These initial conditions for these equations can be written as

$$\alpha(0) = 0, \tag{364a}$$

$$\phi(0) = 0. \tag{364b}$$

These equations can now be integrated out to any value of ξ that you want. A plot of $\phi(\xi)$ can be made by outputting the values as the integration procedes.

VIII. DIFFRACTIVE THEORY OF LOSSLESS BEAM SHAPING

A. Scaling Properties

We now present a theory of lossless beam shaping that is based on diffraction theory (19). In the geometrical theory of beam shaping it is possible to turn a beam with one irradiance distribution into a beam with any desired irradiance distribution, provided only that the energies of the incoming and outgoing beams are the same. However, when diffraction effects are taken into account, this is no longer possible. The geometrical theory is valid provided the wavelength is small. The major goal of this section is to quantify what we mean by a small wavelength. As in our discussion of geometrical beam shaping, we are interested in turning a beam with an incoming irradiance distribution of $I(x/R, y/R)$ at the plane $z = 0$ into a beam with an irradiance distribution of $Q(x/D, y/D)$ at the plane $z = f$. We will see that the parameter

$$\beta = \frac{2\pi RD}{f\lambda} \tag{365}$$

is a dimensionless measure of how small the wavelength λ is. If this parameter is large, then the results from the geometrical theory of beam shaping should be valid. If it is small, then diffractive effects will be important. The parameter β is one, but not the only, measure of how difficult our beam shaping problem is. We will see that the smoothness properties of our input and output beam is another important measure of the difficulty of the beam shaping problem.

Suppose that at the plane $z = 0$ the incoming wave field is given by $g(x/R, y/R)$, and we have an aperture that has a lens with focal length f along with an additional phase element $\psi(x/R, y/R)$. The theory of Fourier optics shows us that the wavefield at $z = f$ is given by

$$U(x_f, y_f, f) = \frac{1}{i\lambda f}\, e^{ikf} e^{ik(x_f^2+y_f^2)/2f} \times \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x/R, y/R) e^{i\psi(x/R,y/R)} e^{-ik(x_f x + y_f y)/f}\, dx\, dy. \tag{366}$$

We would like to determine a function ψ such that the output $U(x_f, y_f)$ satisfies

$$|U(x_f, y_f)|^2 = A\, \frac{R^2}{D^2}\, Q(x_f/D, y_f/D), \tag{367}$$

where the function Q determines the shape of the desired irradiance distribution, D determines the scale of the desired irradiance distribution, and A is a scaling factor that guarantees that the energy of the output beam is the same as that of the incoming beam. At this point our problem has the parameters $\lambda = 2\pi/k$, f, R, and D, and it is not clear what we mean when we say the wavelength is small. We can collect all of our parameters into a single parameter by introducing dimensionless coordinates. In particular, assuming we could choose ψ so that our desired output had exactly the right shape, we would have

$$|G(\omega_x, \omega_y)|^2 = \frac{4\pi^2 A}{\beta^2}\, Q(\omega_x/\beta, \omega_y/\beta), \tag{368a}$$

where

$$G(\omega_x, \omega_y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(\xi, \eta) e^{-i(\omega_x \xi + \omega_y \eta)} e^{i\beta\phi(\xi,\eta)}\, d\xi\, d\eta, \tag{368b}$$

where we have introduced the variables

$$\xi = x/R, \tag{369a}$$

$$\eta = y/R, \tag{369b}$$

$$\omega_x = x_f Rk/f, \tag{369c}$$

$$\omega_y = y_f Rk/f, \tag{369d}$$

and

$$\psi(\xi, \eta) = \beta\phi(\xi, \eta). \tag{369e}$$

We have chosen to write the phase as $\beta\phi$ rather than as ψ. This will be convenient when we are doing the large β approximation. We will refer to the Eqs. (368) as the dimensionless beam shaping equations. Given the function g, the function Q, and the parameter β our goal is to determine a constant A and a function $\phi(\xi, \eta)$ such that Eqs. (368) are satisfied. This statement of the beam shaping problem is very nice because we have collected all of our parameters into the single parameter β.

B. One-dimensional Beam Shaping

As in the theory of geometrical beam shaping, we now consider problems where the incoming beam $g(\xi, \eta)$ and the desired output $Q(s, t)$ can be written as a direct product. This allows us to separate the beam shaping problem into two one-dimensional problems. In particular we are trying to find a function ϕ and a constant A such that for a given $g(\xi)$, $Q(s)$ and β we have

$$|G(\omega)|^2 = A\,\frac{2\pi}{\beta}\,Q(\omega/\beta), \tag{370a}$$

where

$$G(\omega) = \int_{-\infty}^{\infty} g(\xi) e^{-i\omega\xi} e^{i\beta\phi(\xi)} d\xi. \tag{370b}$$

In general it is not possible to choose ϕ so that Eqs. (370) are satisfied exactly. For example, if β is small, then we would need the Fourier transform of $g(\xi)e^{i\phi(\xi)}$ to be very concentrated around the origin. This would contradict the uncertainty principle. To make this statement more precise, we can apply the uncertainty principle to the function $g(\xi)e^{i\beta\phi(\xi)}$ and its desired Fourier transform to get

$$\Delta_g \Delta_G \geq \tfrac{1}{4}, \tag{371}$$

where

$$\Delta_G = \frac{\int_{-\infty}^{\infty} \omega^2 |G(\omega)|^2 \, d\omega}{\int_{-\infty}^{\infty} |G(\omega)|^2 \, d\omega}, \tag{372}$$

and

$$\Delta_g = \frac{\int_{-\infty}^{\infty} \xi^2 |g(\xi)|^2 \, d\xi}{\int_{-\infty}^{\infty} |g(\xi)|^2 \, d\xi}. \tag{373}$$

If we could choose ϕ so that we accomplished our beam shaping exactly, we would have

$$|G(\omega)|^2 = \frac{A2\pi}{\beta} Q(\omega/\beta). \tag{374}$$

This would imply that

$$\Delta_G = \beta^2 \Delta_Q, \tag{375}$$

where

$$\Delta_Q = \frac{\int_{-\infty}^{\infty} \omega^2 |Q(\omega)|^2 \, d\omega}{\int_{-\infty}^{\infty} |Q(\omega)|^2 \, d\omega}, \tag{376}$$

and hence,

$$\beta^2 \Delta_g \, \Delta_Q \geq \tfrac{1}{4}. \tag{377}$$

This inequality cannot be satisfied if β is too small. It should be evident that if β is very small, then it will not even be possible to turn the beam into a profile that is even near the desired profile. This shows that it is not possible to do a good job of beam shaping if the parameter β is small. We now consider the case where β is large, and show that in this case if we choose ϕ to be the function obtained from using geometrical beam shaping, then this will nearly satisfy our beam shaping problem.

We begin our analysis of the beam shaping problem by commenting on our decision to write the phase delay as $\beta\phi(\xi)$. This scaling will allow us to use the method of stationary phase to determine the behavior for large values of β. It should be noted that this scaling predicts that the phase function grows linearly with the frequency of light that we are using, a result

that would hold if we designed a lens based on geometrical optics, and kept the same lens for all frequencies of light.

If we use the variable

$$\alpha = \frac{\omega}{\beta}, \tag{378}$$

our beam shaping problem can be written as follows: given the function g, the function Q, and the parameter β, try to determine the constant A, and the function ϕ such that

$$G(\alpha) = \int_{-\infty}^{\infty} g(\xi) e^{i\beta(\phi(\xi)-\alpha\xi)}\, d\xi, \tag{379a}$$

$$|G(\alpha)|^2 = \frac{2\pi A}{\beta}\, Q(\alpha). \tag{379b}$$

The integral in Eq. (2.379a) is in a form that can be evaluated using the method of stationary phase. To lowest order in β, the method of stationary phase shows us that the integral is given by

$$G(\alpha) \approx e^{i\pi/4} e^{i\beta(\phi(\xi(\alpha))-\xi(\alpha))} \sqrt{2\pi}\, \frac{g(\xi(\alpha))}{\sqrt{\beta\phi''(\xi(\alpha))}}, \tag{380}$$

where the function $\xi(\alpha)$ is determined implicitly by the equation

$$\frac{d}{d\xi}\, \phi(\xi(\alpha)) - \alpha = 0. \tag{381a}$$

If we have chosen ϕ so that the beam has the desired output, then we have

$$AQ(\alpha) = \frac{g^2(\xi(\alpha))}{\phi''(\xi(\alpha))}. \tag{381b}$$

With a little bit of manipulation we can make these equations identical to the equations for geometrical beam shaping. In order to do this we begin by differentiating the first of Eqs. (381) with respect to α. This gives us

$$\frac{d^2\phi(\xi(\alpha))}{d\xi^2}\, \frac{d\xi(\alpha)}{d\alpha} = 1. \tag{382}$$

Using this equation, the Eq. (381b) can be written as

$$\frac{d\xi(\alpha)}{d\alpha}\, g^2(\xi) = AQ(\alpha). \tag{383}$$

If we use the fact that the irradiance of the incoming beam is given by

$|g(\xi)|^2 = I(\xi)$, we get the system of equations

$$\frac{d\xi}{d\alpha} I(\xi) = AQ(\alpha), \tag{384a}$$

$$\frac{d}{d\xi} \phi(\xi(\alpha)) - \alpha = 0. \tag{384b}$$

If we integrate the first of these equations from $-\infty$ to ∞ we find the normalization condition

$$A = \frac{\int_{-\infty}^{\infty} I(\xi)\, d\xi}{\int_{-\infty}^{\infty} Q(\alpha)\, d\alpha}. \tag{384c}$$

These equations are identical to Eqs. (304) derived using the geometrical theory of beam shaping.

C. Two-dimensional Beam Shaping

We will now quickly summarize how our results can be extended to apply to arbitrary beam shape problems, that is, ones that are not separable. In general we want to find a function $\phi(\xi, \eta)$ such that

$$G(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(\xi) e^{i\beta(\phi(\xi,\eta) - x\xi - y\eta)}\, d\xi\, d\eta, \tag{385a}$$

$$|G(x, y)|^2 = \frac{4\pi^2 A}{\beta^2} Q(x, y). \tag{385b}$$

An argument almost identical to that used in the separable case shows that the uncertainty principle requires that

$$\beta^2 \Delta_g \Delta_Q \geq 1, \tag{386a}$$

where

$$\Delta_g = \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\xi^2 + \eta^2) |g(\xi, \eta)|^2\, d\xi\, d\eta}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |g(\xi, \eta)|^2\, d\xi\, d\eta}, \tag{386b}$$

and

$$\Delta_Q = \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\omega_x^2 + \omega_y^2) |Q(\omega_x, \omega_y)|^2\, d\omega_x\, d\omega_y}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |Q(\omega_x, \omega_y)|^2\, d\omega_x\, d\omega_y}. \tag{386c}$$

As in the separable case, this inequality cannot be satisfied if β is too small. We now consider the limit of the integral in (385a) as $\beta \to \infty$. Using the two-dimensional method of stationary phase we find that

$$|G(x,y)|^2 \approx \frac{4\pi^2}{\beta^2 J(\xi_0,\eta_0)} |g(\xi_0,\eta_0)|^2, \tag{387}$$

where (ξ_0,η_0) are determined implicitly by the stationarity conditions

$$\frac{\partial}{\partial\xi}\phi(\xi_0,\eta_0) = x, \tag{388a}$$

$$\frac{\partial}{\partial\eta}\phi(\xi_0,\eta_0) = y. \tag{388b}$$

and the function J is defined by

$$J(\xi_0,\eta_0) = \frac{\partial^2\phi(\xi_0,\eta_0)}{\partial\xi^2}\frac{\partial^2\phi(\xi_0,\eta_0)}{\partial\eta^2} - \left(\frac{\partial^2\phi(\xi_0,\eta_0)}{\partial\xi\,\partial\eta}\right)^2. \tag{389}$$

If we use the stationarity conditions, we can write the function J as

$$J(\xi_0,\eta_0) = \frac{\partial x(\xi_0,\eta_0)}{\partial\xi}\frac{\partial y(\xi_0,\eta_0)}{\partial\eta} - \frac{\partial x(\xi_0,\eta_0)}{\partial\eta}\frac{\partial y(\xi_0,\eta_0)}{\partial\xi}. \tag{390a}$$

If we require that the function $|G(x,y)|^2$ has the desired output we arrive at the equation

$$I(\xi_0,\eta_0) = AQ(x,y)J(\xi_0,\eta_0). \tag{390b}$$

These last two equations along with the stationarity conditions in Eq. (388) are identical to the two-dimensional equations that we derived using geometrical optics.

D. Radially Symmetric Problems

In our section on geometrical beam shaping we considered problems that have radial symmetry. We now consider how to analyze these problems for the effect of diffraction. Problems with radial symmetry can be considered as a special case of the general theory of two-dimensional beam shaping. These problems are important enough that they deserve some special attention. Suppose both the input beam g and the desired output beam Q have radial symmetry. In this case the phase function ϕ will also have radial symmetry, and we can replace our two-dimensional Fourier transforms with Hankel transforms (see Sec. II.C in this chapter). The theory of Hankel transforms shows that our beam shaping problems can be phrased as follows.

Given a function $g(\xi)$, a function $Q(\alpha)$, and a parameter β, find a function $\phi(\xi)$ such that

$$G(\alpha) = 2\pi \int_0^\infty g(\xi)\xi e^{i\beta\phi(\xi)} J_0(\alpha\xi)\, d\xi \tag{391a}$$

satisfies

$$|G(\alpha)|^2 = \frac{4\pi^2 A}{\beta^2}\, Q(\alpha/\beta). \tag{391b}$$

We already know that a lens designed using the first-order term in the stationary phase approximation gives the same lens as one designed using geometrical optics. Since radially symmetric problems are special cases of the two-dimensional case, if we design a radially symmetric lens using the large β limit, we should get the same lens as when we design it using geometrical optics. We conclude that the function $\phi(\xi)$ can be obtained by using the techniques described in our section on the geometrical theory of beam shaping. Once we have obtained this function, we can use Eqs. (391) to see how our system performs with a finite value of β. To carry this out in practice, we have used ODE solvers in order to compute the function ϕ, and to perform the integration in the definition of the Hankel transform.

E. The Continutity of ϕ

We have seen that the first term in the method of stationary phase is identical to the results obtained using geometrical optics. In order for us to know how well the geometrical optics approximation is working, it is necessary to understand the next order term in the stationary phase approximation. We discussed the higher-order terms in the method of stationary phase in Sec. III. There we saw that if the functions ϕ and g are infinitely differentiable, then the next order term in the method of stationary phase is $1/\beta$ times the size of the first term. However, if the third derivative of ϕ (or g) are discontinuous, then the next order term will only be $1/\sqrt{\beta}$ times smaller than the first-order term. If ϕ has a discontinuity in a lower derivative, we get even worse convergence.

We now consider what class of functions $Q(\alpha)$ will lead to discontinuities in the phase function $\phi(\xi)$ designed by using geometrical optics. We will assume that the function $I(\xi)$ is smooth (such as a Gaussian). Equations (384) show that the derivative of ϕ has the same continuity properties as the function $\alpha(\xi)$. If we take the derivative of the first of Eqs. (384) with respect

to ξ we find

$$A\left(\frac{dQ}{d\alpha}\left(\frac{d\alpha}{d\xi}\right)^2 + Q(\alpha)\frac{d^2\alpha}{d\xi^2}\right) = \frac{dI}{d\xi}. \tag{392}$$

We see that if the function $Q(\alpha)$ has a discontinuous derivative at a point $\alpha = \alpha(\xi_0)$ where $I(\xi_0) \neq 0$, then this will lead to a discontinuity in the second derivative of α with respect to ξ, and hence to a discontinuity in the third derivative of ϕ. It follows that discontinuities in the derivatives of Q or I will slow down the convergence towards the geometrical optics limit.

Note that we excluded the case where the discontinuity in Q occurs at a point where I vanishes. In this case we must have $d\alpha/d\xi = 0$, and when we look at our expression for the second derivative of α we find that it does not have a discontinuity. Similar arguments hold for the case where Q itself is discontinuous at a point where I vanishes. A very important example of this is the case where one turns a Gaussian profile into a flat-top beam. In that case the phase function is infinitely differentiable, even though the function $Q(\alpha)$ has a discontinuity in it. This is because the discontinuity in Q occurs as $\xi \to \infty$, and hence at a point where $I(\xi) = 0$.

For the case where the incoming distribution $I(\xi)$ is a Gaussian, we see that discontinuities in the first derivative of Q will lead to discontinuities in the third derivative of ϕ, unless the discontinuity in Q occurs at an extremity. By an extremity we mean a point where the rays reaching this point have come from points infinitely far off the axis.

F. One-Dimensional Examples

In order to illustrate the principles of beam shaping a computer code was written that allows us to compute the function ϕ as well as the effects of using a finite value of β. In these examples we calculate

$$G(\alpha) = \int_{-\infty}^{\infty} g(\xi) e^{i\beta(\phi(\xi)-\alpha\xi)}\, d\xi, \tag{393}$$

by using an ODE integrator. When an analytical expression for ϕ cannot be found, we compute ϕ with the ODE integrator as we are computing the integral. We output the quantity

$$\Gamma(\alpha, \beta) = \frac{2\pi A}{\beta} |G(\alpha)|^2, \tag{394}$$

where

$$A = \frac{\int_{-\infty}^{\infty} I(\xi)\, d\xi}{\int_{-\infty}^{\infty} Q(\alpha)\, d\alpha}. \tag{395}$$

If the effects of diffraction are negligible, the function $\Gamma(\alpha, \beta)$ should be very close to $Q(\alpha)$.

We could have used a code that computed the function ϕ using the technique described in the section on geometrical beam shaping, and then fed this input into an FFT for computing the effects of a finite value of β.

In all of the examples we present we will use the function

$$g(\xi) = e^{-\xi^2/2}, \tag{396}$$

and hence

$$I(\xi) = e^{-\xi^2}. \tag{397}$$

Example 14 *Turning a Gaussian Into a Flat top* We want to turn the output beam into a flat top with

$$Q(\alpha) = 1 \text{ for } |\alpha| < 1 \tag{398}$$

$$Q(\alpha) = 0 \text{ for } |\alpha| > 1. \tag{399}$$

We have already considered this example in our section on geometrical beam shaping, where it was shown that the function ϕ is given by

$$\phi(\xi) = \frac{2}{\sqrt{\pi}} \left(\xi \frac{\sqrt{\pi}}{2} \operatorname{erf}(\xi) + \tfrac{1}{2} e^{-\xi^2} - \tfrac{1}{2} \right). \tag{400}$$

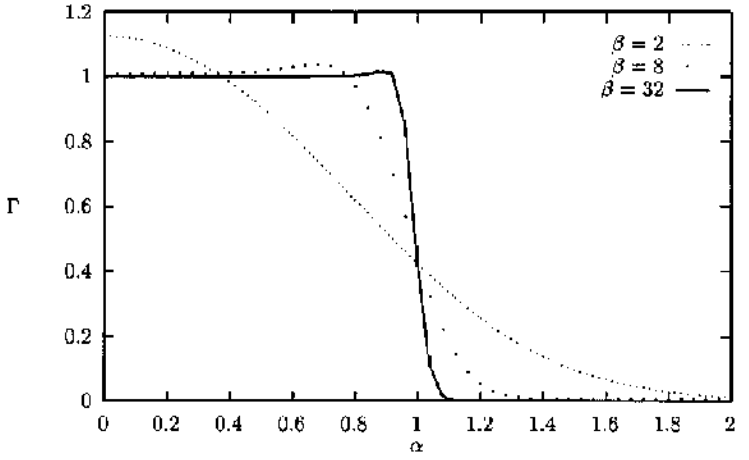
We will be able to see the effects of having a finite value of β. Figure 5a shows plots of $\Gamma(\alpha, \beta)$ for various values of β. We see that for $\beta = 2$ the answer does not look at all like a square pulse, while for $\beta = 32$ the answer is starting to look very good.

Figure 5b shows a plot of the function $\phi(\xi)$.

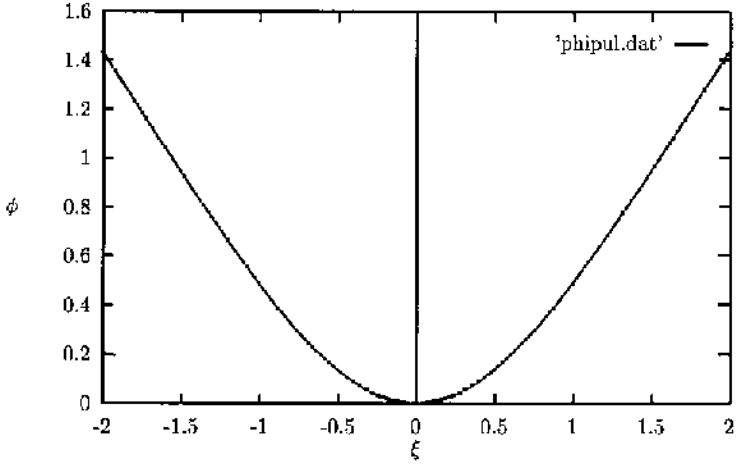
Example 15 *A Polynomial Output—I* We will now let the output beam be a polynomial that has a hump in it.

$$Q(\alpha) = (1 - \alpha^2)(\alpha^2 + \delta) \text{ for } |\alpha| < 1, \tag{401}$$

$$Q(\alpha) = 0 \text{ for } |\alpha| > 1. \tag{402}$$



(a)



(b)

Figure 5 (a) The intensity distribution for different values of β for the problem of turning a one-dimensional Gaussian into a flat-top beam (Example 14). (b) The function $\phi(\xi)$ that accomplishes this exactly in the geometrical optics limit.

The constant A is easily computed to be

$$A = \frac{15\sqrt{\pi}}{4 + 20\delta}. \tag{403}$$

We will choose $\delta = 1$, for this example. Once we know the constant A, we use the ODE solver to compute the function ϕ and the function $\Gamma(\alpha, \beta)$ for various values of β. Figure 6a shows plots of $\Gamma(\alpha, \beta)$ for various values of β. Once again, the results are not good for $\beta = 2$, but get progressively better as we increase the value of β. A careful analysis of the data shows that the relative error

$$e(\alpha, \beta) = \frac{\Gamma(\alpha, \beta) - Q(\alpha)}{Q(\alpha)}, \tag{404}$$

is going to zero like $1/\beta$ everywhere except right at the endpoints $\alpha = \pm 1$.

Figure 6b shows a plot of the function $\phi(\xi)$.

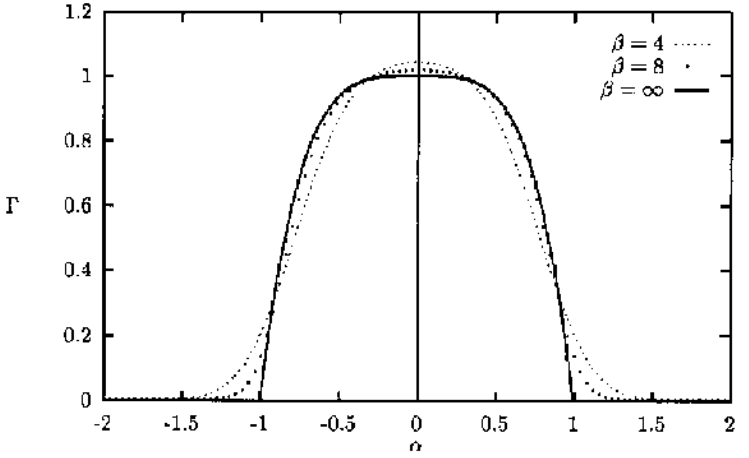
Example 16 *A Polynomial Output—II* This example is the same as the last example except that we have chosen a value of $\delta = 0.25$ in the function $Q(\alpha)$. This causes the function Q to have two humps in it. Figure 7a shows plots of $\Gamma(\alpha, \beta)$ for varous values of β, and figure 7b shows a plot of the function ϕ. The relative error is dying down faster than $1/\beta^2$ almost everywhere. Once again right at the ends ($\alpha = \pm 1$), we do not get this behavior, and in the middle ($\alpha = 0$) the convergence is somewhat slower than $1/\beta^2$. The slow convergence at this point does not appear to be illustrating any fundamental principle, but appears to go away if we choose a large enough value of β.

Example 17 *A Triangle Function*

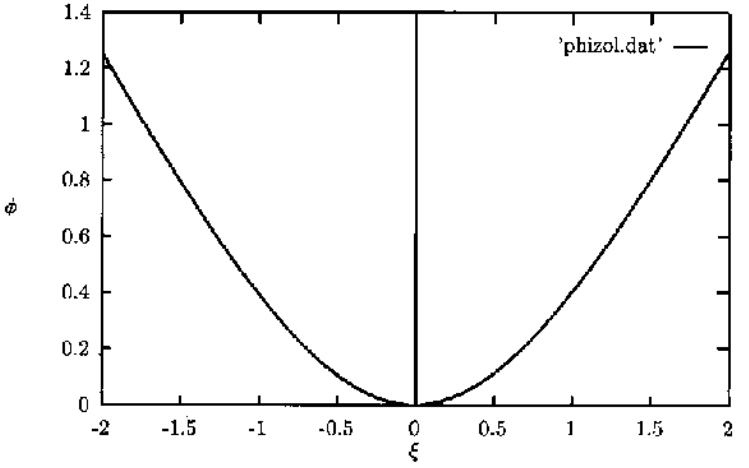
$$Q(\alpha) = 1 - |\alpha| \text{ for } |\alpha| < 1, \tag{405}$$

$$Q(\alpha) = 0 \text{ for } |\alpha| > 1. \tag{406}$$

This discontinuity in the derivative of the function $Q(\alpha)$ at $\alpha = 0$ causes the function ϕ to have a discontinuity in its third derivative. Figure 8a shows plots of the function $\Gamma(\alpha, \beta)$ for various values of β. At the point $\alpha = 0$ the convergence towards the function $Q(\alpha)$ can be seen to be going like $1/\sqrt{\beta}$. Figure 8b shows a plot of the function ϕ.

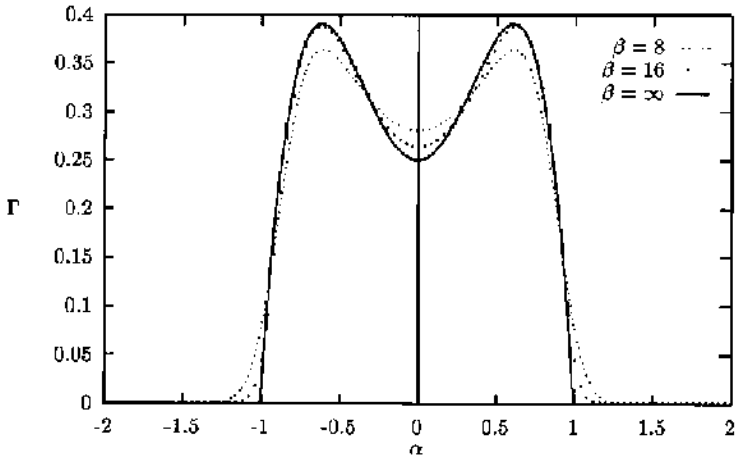


(a)

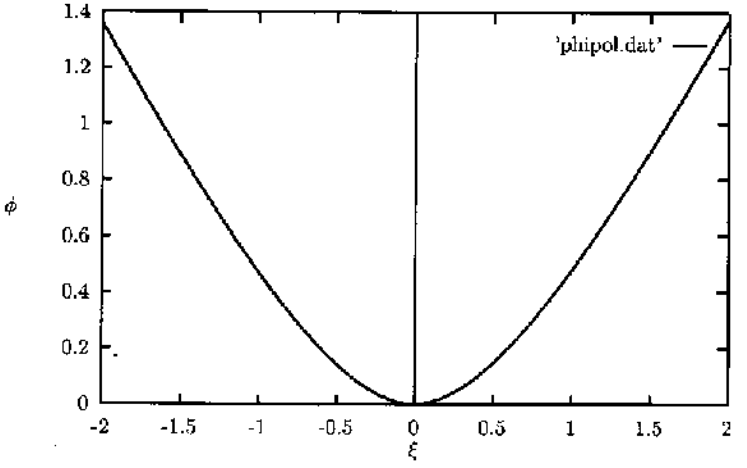


(b)

Figure 6 (a) The intensity distribution for different values of β for the problem of turning a Gaussian into the output $Q(\alpha) = (1 - \alpha^2)(1 + \alpha^2)$ for $|\alpha| < 1$, $Q(\alpha) = 0$ for $|\alpha| > 1$. (b) The function $\phi(\xi)$ that accomplishes this in the geometrical optics limit.

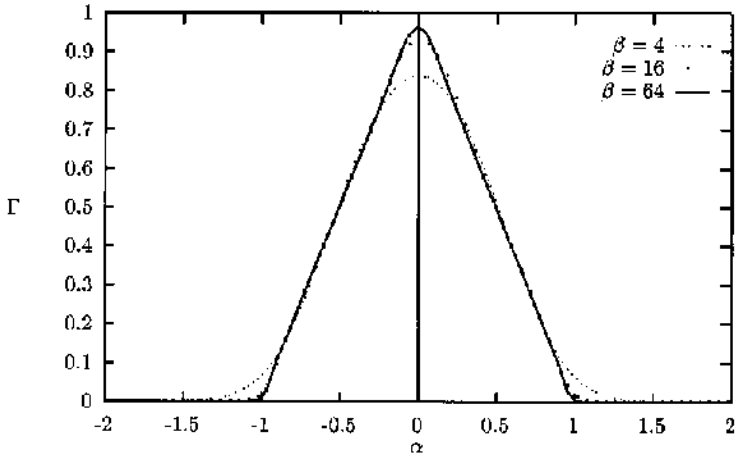


(a)



(b)

Figure 7 (a) The intensity distribution for different values of β for the problem of turning a Gaussian into the output $Q(\alpha) = (1 - \alpha^2)(1/4 + \alpha^2)$ for $|\alpha| < 1$, $Q(\alpha) = 0$ for $|\alpha| > 1$. (b) The function $\phi(\xi)$ that accomplishes this in the geometrical optics limit.



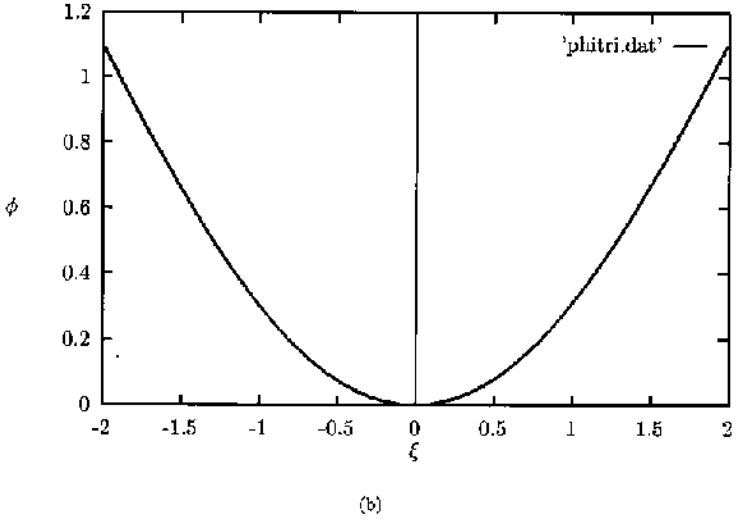


Figure 8 (a) The intensity distribution for different values of β for the problem of turning a Gaussian into a triangle function $Q(\alpha) = 1 - |\alpha|$ for $|\alpha| < 1$, $Q(\alpha) = 0$ for $|\alpha| > 1$. (b) The function $\phi(\xi)$ that accomplishes this in the geometrical optics limit.

Example 18 *A Stairstep Function—I* We now consider the case where $Q(\alpha)$ is a stair step function.

$$Q(\alpha) = \gamma \text{ for } |\alpha| < 1/2, \tag{407a}$$

$$Q(\alpha) = 1 \text{ for } 1/2 < |\alpha| < 1, \tag{407b}$$

$$Q(\alpha) = 0 \text{ for } |\alpha| > 1. \tag{407c}$$

In this example we choose $\gamma = 3/4$. The discontinuity in the function Q at $\alpha = \pm 1/2$ causes the function ϕ to have a discontinuity in its second derivative. Figure 9a shows plots of the function $\Gamma(\alpha, \beta)$. The convergences towards the solution $Q(\alpha)$ is extremely slow. Figure 9b shows a plot of the function ϕ.

Example 19 *A Stairstep function—II* This is the same as in the last example except we choose the parameter γ in the function Q to be equal to zero. This causes the function ϕ to have a discontinuity in the first derivative. Figure 10a shows plots of the function $\Gamma(\alpha, \beta)$. We see that the convergence towards $Q(\alpha)$ is extremely slow in this case. Figure 10b shows a plot of ϕ.

G. An Axisymmetric Example

In our section on geometric beam shaping we considered the problem of turning a circular Gaussian beam into an axisymmetric flat-top beam. In this case the input beam $g(\xi, \eta)$ is given by

$$g(\xi, \eta) = e^{-\xi^2 - \eta^2}, \tag{408}$$

and the desired output is given by

$$Q(x, y) = 1 \text{ for } x^2 + y^2 < 1, \tag{409}$$

$$Q(x, y) = 0 \text{ for } x^2 + y^2 > 1. \tag{410}$$

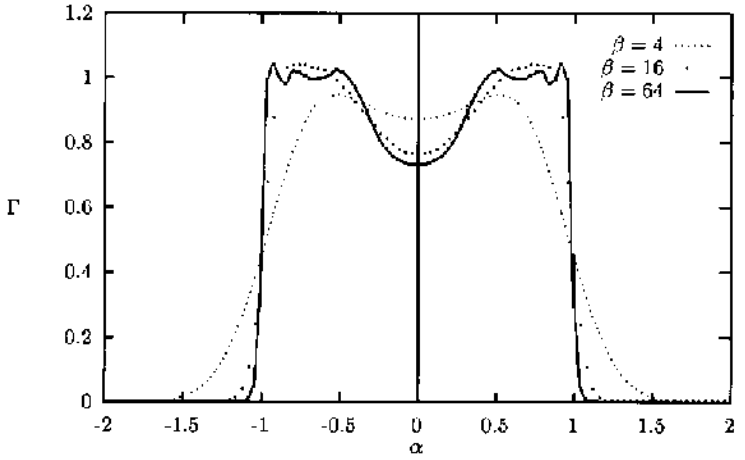
The radially symmetric beam shaping equations give us the normalization constant

$$A = 1. \tag{411}$$

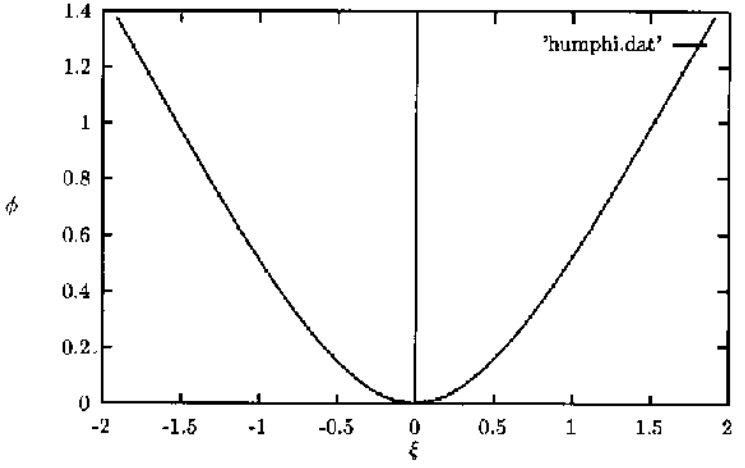
The phase function is given by

$$\phi(r) = \int_0^r \sqrt{1 - e^{-\xi^2}}\, d\xi, \tag{412}$$

where $r^2 = \xi^2 + \eta^2$. In order to analyze the effects of diffraction we compute the radially symmetric Fourier transform. In our section on math-



(a)



(b)

Figure 9 (a) The intensity distribution for different values of β for the problem of turning a Gaussian into a step function $Q(\alpha) = 3/4$ for $|\alpha| < 1/2$, $Q(\alpha) = 1$ for $1/2 < |\alpha| < 1$, $Q(\alpha) = 0$ for $|\alpha| > 1$. (b) The function $\phi(\xi)$ that accomplishes this in the geometrical optics limit.

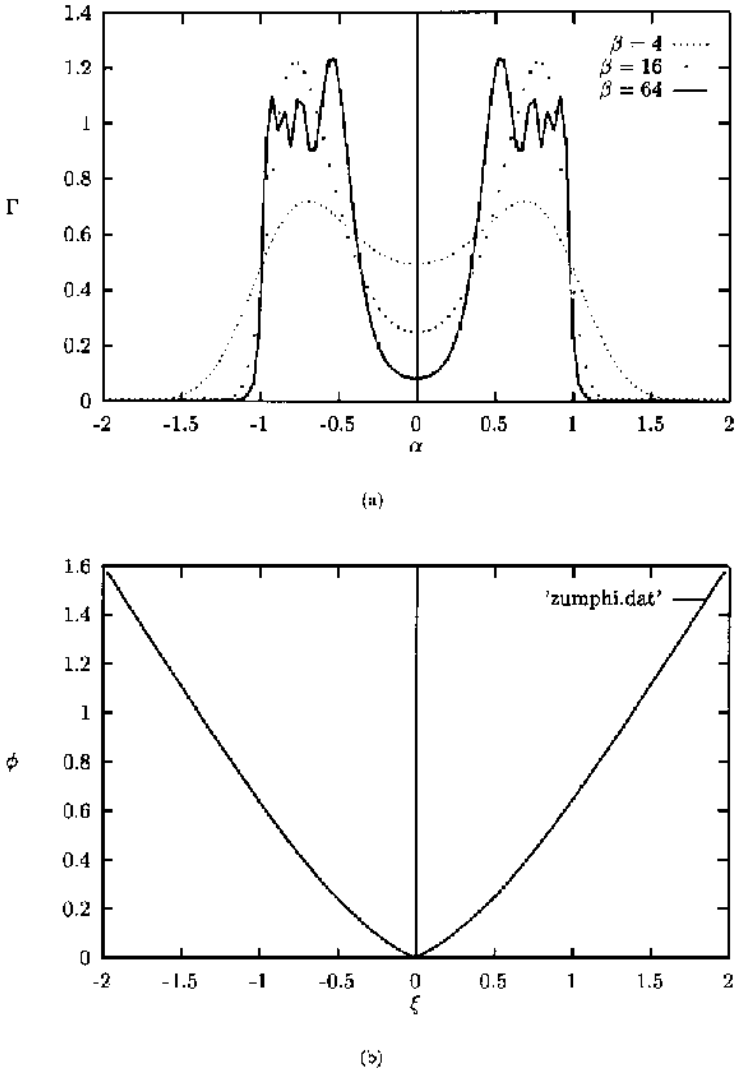
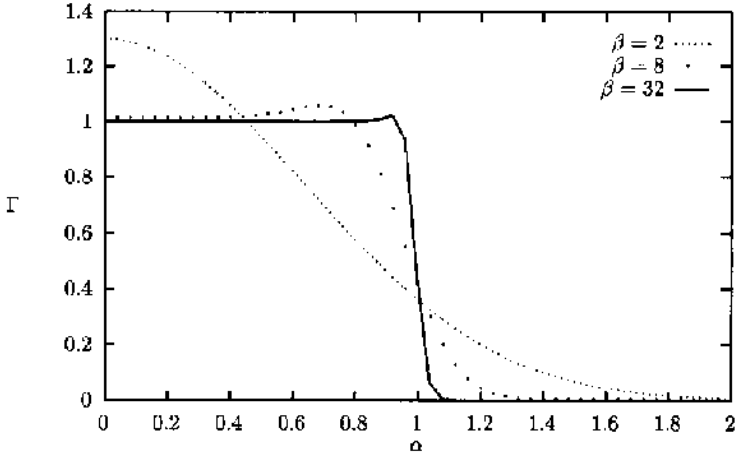
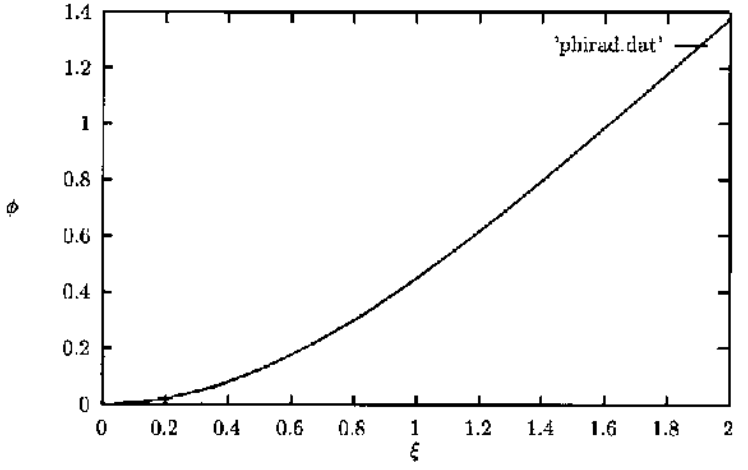


Figure 10 (a) The intensity distribution for different values of β for the problem of turning a Gaussian into a step function $Q(\alpha) = 0$ for $|\alpha| < 1/2$, $Q(\alpha) = 1$ for $1/2 < |\alpha| < 1$, $Q(\alpha) = 0$ for $|\alpha| > 1$. (b) The function $\phi(\xi)$ that accomplishes this in the geometrical optics limit.



(a)



(b)

Figure 11 (a) The intensity distribution for different values of β for the problem of turning a radially symmetric Gaussian into a radially symmetric flat-top beam. (b) The function $\phi(\xi)$ that accomplishes this in the geometrical optics limit.

ematical preliminaries, we showed that this can be done using the Hankel transform.

$$G(\alpha) = 2\pi \int_0^\infty e^{i\beta\phi(r)} r J_0(\alpha\beta r) g(r)\, dr. \tag{413}$$

We are interested in the normalized irradiance of this function.

$$\Gamma(\alpha, \beta) = \frac{4\pi^2}{\beta^2} |G(\alpha)|^2. \tag{414}$$

If the effects of diffraction are negligible, then the function $\Gamma(\alpha, \beta)$ should be nearly equal to $Q(\alpha)$.

Figure 11a shows a plot of $\Gamma(\alpha, \beta)$ for various values of β. We see that the results are quite similar to the one dimensional case. Figure 11b shows a plot of the function ϕ.

REFERENCES

1. GK Batchelor. An Introduction to Fluid Mechanics. Cambridge: Cambridge University Press, 1967.
2. PW Bridgeman. Dimensional Analysis. New Haven: Yale University Press, 1922.
3. RN Bracewell. The Fourier Transform and its Applications. New York: McGraw-Hill, 1978.
4. RV Churchill. Operational Mathematics. New York: McGraw-Hill, 1972.
5. LE Franks. Signal Theory, Englewood Cliffs, NJ: Prentice Hall, 1969.
6. A Messiah. Quantum Mechanics. Amsterdam: North Holland, 1958.
7. A Sommerfeld. Partial Differential Equations in Physics. New York: Academic Press, 1949.
8. N Bleistein and RA Handelsman. Asymptotic Expansions of Integrals. New York: Dover, 1986.
9. MJ Lighthill. Waves in Fluids. Cambridge: Cambridge University Press, 1978.
10. GB Whitham. Linear and Nonlinear Waves. New York: John Wiley, 1974.
11. M Born, E Wolf. Principles of Optics. London: Permagon Press, 1964.
12. J Lighthill. Fourier Analysis and Generalized Functions. London: Cambridge University Press, 1958.
13. JD Jackson. Classical Electrodynamics. New York: John Wiley, 1962.
14. PR Garabedian. Partial Differential Equations. New York: John Wiley, 1964.
15. N Bleistein. Mathematical Methods for Wave Propagation. Orlando, FL: Academic Press, 1984.
16. JW Goodman. An introduction to Fourier Optics. New York: McGraw-Hill, 1968.
17. JA Stratton. Elecrromagnetic Theory. New York: McGraw-Hill, 1941.

18. CC Aleksoff, KK Ellis, BD Neagle. Holographic conversion of a Gaussian beam to a near-field uniform beam. Opt Eng 30:537–543, 1991.
19. LA Romero, FM Dickey, Lossless laser beam shaping. J Opt Soc Amer A 13(4):751–760, 1996.

3

Gaussian Beam Shaping: Diffraction Theory and Design

Fred M. Dickey and Scott C. Holswade
Sandia National Laboratories, Albuquerque, New Mexico

I. INTRODUCTION

This chapter describes a diffraction-based method for converting single-mode Gaussian beams into beams with uniform irradiance profiles. The design is based on a Fourier transform relation between the input and output beam functions. This solution can be obtained using geometrical optics methods. However, the diffraction approach introduces a parameter that contains the product of the widths of the input and output beams. This parameter is a significant part of the physical optics solution. The efficacy of the solution is shown to depend on this parameter. The quality of the solution improves asymptotically with increasing value of the parameter.

Many experiments and industrial applications require a laser beam irradiance that is nominally constant over a specified area. Such applications include laser/material processing, laser/material interaction studies, optical data/image processing and lithography. In many cases it is desirable, for obvious reasons, that the beam shaping operation conserves energy.

The multifaceted integrator approach to laser beam shaping is especially suitable to laser beams with highly irregular (multi-mode) irradiance distributions (1,2). The number and size of the facets is selected to accomplish the required integration or, equivalently, averaging. Doherty (3) has treated the problem of irradiance mapping for laser beams with radial symmetry and regular irradiance distributions. Dickey and O'Neil (4) give a general formulation of the multi-faceted beam integrator problem and introduce a configuration that minimizes deleterious diffraction effects.

For single-mode beams with a Gaussian profile it is possible to map the beam into a uniform intensity profile with steep skirts. This mapping can be accomplished with simpler optics that is more flexible with respect to scaling and does not have the interference patterns inherent in multi-faceted beam integration. Several authors address the problem of mapping a Gaussian beam into one with a uniform irradiance distribution. The earliest paper known to the authors that addresses the lossless shaping of a single mode laser beam is the paper by Frieden (5). Lee (6) employs an iterative technique to design a phase filter to convert a Gaussian beam into a more uniform irradiance distribution. Veldkamp (7,8) uses an iterative technique to design binary gratings to accomplish the profile shaping. Aleksoff et al. (9) uses the geometrical optics approximation to develop a holographic system that maps a Gaussian beam into a rectangularly shaped beam with uniform amplitude and phase. Kosoburd and Kedmi (10) use geometrical optics to design a diffractive system that maps Gaussian beams into beams with uniform irradiance. Eismann et al. (11) applies the Gershberg–Saxton algorithm, or equivalently, phase retrieval to synthesize a two element design that produces a beam with uniform amplitude and phase. In a recent paper, Golub et al. (12) present numerical and experimental results for a diffractive beam shaper based on a geometrically derived phase function.

In this chapter, we give a solution to the problem of mapping a Gaussian laser beam into a beam with uniform irradiance profile. The configuration analyzed exploits the Fourier transform properties of lenses. That is, the output optical field is the Fourier transform of the input optical field and a phase function. This configuration has the advantage of being able to change the output size or the working distance by changing the transform lens. In Sec. II we define the general problem of converting a Gaussian beam into a beam with uniform irradiance and give a solution to the problem. Bounds set by the uncertainty principle are also discussed in this section. The problem of creating a collimated beam, a beam with both a uniform irradiance and a uniform phase, is addressed in Sec. III. Again, the uncertainty principle has implications. It is used to define a generalized Rayleigh range for shaped beams. In Sec. IV considerations associated with designing a shaping element are discussed. Sensitivity to alignment and scaling errors are treated in Sec. V. Section VI discusses the application of the design methodology to a particular problem. In Sec. VII we present the results of the design and testing of a prototype system. A summary of the chapter is given in Sec. VIII. This chapter is based on the author's papers. (13,14).

II. THE ANALYTICAL SOLUTION

The general beam shaping problem is shown schematically in Fig. 1. In the figure the beam to be shaped enters the proverbial black box from the left

and exits on the right, diffracting to the design irradiance pattern. The black box may contain a single optical element or a combination of several optical components of differing types such as lenses, mirrors, prisms, diffractive optics and holograms. One approach to solving a beam shaping problem would be to assume an optical configuration and develop a solution around this configuration. An example of this approach would be iterative techniques that provide a solution for a single element diffractive optic. A more general approach would be to obtain a general solution for a shaping function, amplitude and phase, using diffraction theory, and then develop an optical design realizing the shaping function. This approach is usually the most difficult. An approach somewhere between the two is commonly what is taken. Although a given solution might be realizable with a single optical element, it is frequently the case that a more versatile, more practicable, and less expensive design is obtained using multiple elements.

A. Optical Configuration

Our approach to lossless beam shaping, illustrated in Fig. 2, consists of a phase element in conjunction with a Fourier transform lens. The optical field at the focal plane of the transform lens is proportional to the

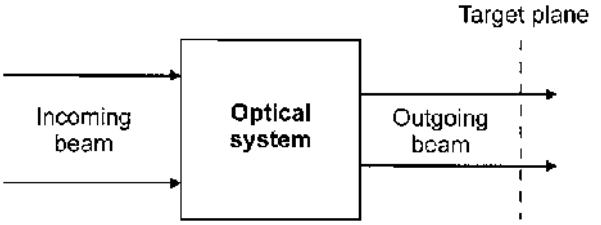


Figure 1 Schematic of the beam shaping problem.

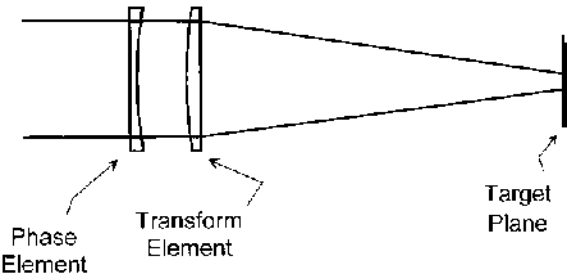


Figure 2 Fourier transform beam shaping system. (From Ref. 13)

Fourier transform of the product of the input optical field and phase of the phase element (15). This configuration has several advantages. The phase element can be changed to control both the scale and shape of the output irradiance. The transform lens can be changed to modify the working distance, with a corresponding change in scale of the output. Finally, although the phase element and transform lens could be designed as one optical element, it is generally easier to design and fabricate the two components if their functions are kept separate.

It should be noted that this configuration is more general than it might first appear. Any solution that could be obtained using a Fresnel integral can be obtained using the Fourier transform system of Fig. 2. This can be seen from the fact that the Fresnel integral can be written as a Fourier transform of the product of the input aperture function and a pure (quadratic) phase factor (see Goodman (15), Chapter 4, Eq. (4–10)). The quadratic phase function becomes part of the beam shaping element.

B. Minimum Mean Square Error Formulation

Given the configuration of Fig. 2, the problem is to design the phase element. The direct approach would be to solve for the phase function that minimizes the mean square difference between the desired irradiance and the irradiance produced by the phase element. That is, we want to find ϕ that minimizes an integral of the form

$$R = \int ||\Im[(2/\sqrt{\pi})^{1/2} e^{-x^2} e^{i\phi}]|^2 - (1/\alpha)^{1/2} \operatorname{rect}(f/\alpha)|^2 \, df, \tag{1}$$

where $\Im$ denotes a Fourier transform operation, f denotes the corresponding frequency domain variable, α defines the size of the output, and the problem is scaled to a unit width $(1/e^2)$ Gaussian beam function. Here, the problem is formulated in one dimension, which is appropriate to the separable problem of converting a circular Gaussian beam into a uniform beam with a square cross section. In general, a single variable will be used to represent a one- or two-dimensional variable. Unfortunately, we were not able to obtain a global solution to Eq. (1). We were able to obtain solutions to the problem using the method of stationary phase. Before presenting the stationary phase solution, it is interesting to discuss the solution to a related problem of requiring both the amplitude and phase of the output to be constant over the region of interest and zero elsewhere.

The solution to the separable uniform amplitude and phase problem can be obtained by determining the phase ϕ that minimizes

the functional

$$R = \int |\Im[(2/\pi)^{1/2} e^{-x^2} e^{i\phi}] - (1/\alpha)^{1/2} \operatorname{rect}(f/\alpha)|^2 \, df$$

$$= 2 - 2 \operatorname{Re} \int \Im[(2/\sqrt{\pi})^{1/2} e^{-x^2} e^{i\phi}](1/\alpha)^{1/2} \operatorname{rect}(f/\alpha) \, df. \tag{2}$$

Note that this equation differs from that of Eq. (1) in that it involves the differences of fields (complex) functions, while Eq. (1) is the difference of intensities (magnitude squared) functions. Thus, the problem described in Eq. (1) is less constrained since the phase of the output is a free parameter. This allows for a broader range of solutions. In Eq. (2) the input Gaussian function and the output rect function are normalized to unit energy. This insures that the mean square difference in Eq. (2) depends on the variation in the shape of the input and output functions and not on any relative amplitude difference between the two functions. As a result, only the integral of the cross terms needs to be evaluated, since the normalization forces the other two integrals to unity.

The solution to Eq. (2) is readily obtained by applying Parseval's theorem and expanding the integrand. Integrating the magnitude squared terms gives

$$R = 2 - 2 \operatorname{Re} \int (2\alpha/\sqrt{\pi})^{1/2} e^{-i\phi} \operatorname{sinc}(\alpha x) e^{-x^2} \, dx. \tag{3}$$

Here, Re z denotes the real part of z. Clearly, R is minimized if the integral is maximized. This is obtained if ϕ is set equal to the phase of the sinc function. Since the phase of the sinc function is a binary function with values of 0 and π, the optimum phase function is a binary function. All that remains is to determine the value of α that maximizes the integral in Eq. (3) with ϕ set equal to the phase of the sinc function. This can be evaluated numerically to give $\alpha = 0.710$. The beam profile irradiance for this solution is shown in Fig. 3. This is just the optimum solution, in the sense of Eq. (2), of the problem posed by Veldkamp (7). It is interesting to note that the solution is not as flat as might be expected. As mentioned above, α determines the size of the shaped output. In obtaining the minimum mean square error solution, we have let the size of the output be a free parameter. The value $\alpha = 0.710$ corresponds to small β discussed in the next section. If the output beam size is allowed to increase, one can get a flatter looking output beam. However the mean square error between the shaped beam and the desired output is larger than that of Fig. 3. A careful inspection of Eq. (3) shows that the mean square error increases to a maximum with increasing α (output beam size). This is due to the fact that, with respect to Eq. (1), Eq. (2) is

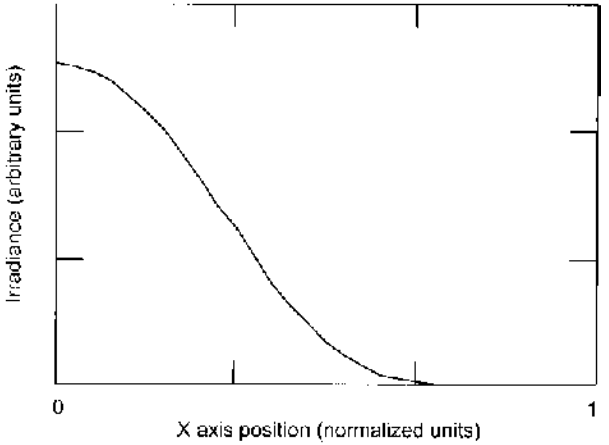


Figure 3 Beam irradiance profile (right half) for the optimum solution of Eq. (2). (From Ref. 13)

overly constrained. That is, in Eq. (2) both the phase and amplitude of the output are required to be uniform, whereas only the irradiance of the output in Eq. (1) is required to be uniform. It can also be noted that this solution corresponds to a value of $R = 1.199$, which is a significant fraction of the maximum of $R = 2$.

C. The Uncertainty Principle

There are fundamental constraints on the on the beam shaping problem that can be traced to electromagnetic theory. It is difficult to develop constraints on a problem without a degree of specificity. However, the uncertainty principle of quantum mechanics or, equivalently, the time-bandwidth inequality associated with signal processing can be applied to the beam shaping configuration defined by Fig. 2. The uncertainty principle is a constraint on the lower limit of the product of the root-mean-square width of a function and its root-mean-square bandwidth (16,17),

$$\Delta_x \Lambda_v \geq \frac{1}{4\pi}. \tag{4}$$

The respective widths are defined by

$$(\Delta_x)^2 = \frac{\int_{\infty}^{\infty}\int_{\infty}^{\infty}(x-\bar{x})^2|u(x,y)|^2\,dx\,dy}{\int_{\infty}^{\infty}\int_{\infty}^{\infty}|u(x,y)|^2\,dx\,dy}, \tag{5}$$

where

$$\bar{x} = \frac{\int_{\infty}^{\infty}\int_{\infty}^{\infty} x|u(x,y)|^2\,dx\,dy}{\int_{\infty}^{\infty}\int_{\infty}^{\infty} |u(x,y)|^2\,dx\,dy}, \tag{6}$$

and

$$(\Delta_v)^2 = \frac{\int_{\infty}^{\infty}\int_{\infty}^{\infty} (v-\bar{v})^2|U(v_x,v_y)|^2\,dv_x\,dv_y}{\int_{\infty}^{\infty}\int_{\infty}^{\infty} |U(v_x,v_y)|^2\,dv_x\,dv_y}, \tag{7}$$

where

$$\bar{v}_x = \frac{\int_{\infty}^{\infty}\int_{\infty}^{\infty} v_x|U(v_x,v_y)|^2\,dv_x\,dv_y}{\int_{\infty}^{\infty}\int_{\infty}^{\infty} |U(v_x,v_y|^2\,dv_x\,dv_y}. \tag{8}$$

In the last two equations upper and lower case letters denote the field function and its Fourier transform, respectively. The uncertainty principle stated in Eq. (4) is obtained from Eq. (5) through Eq. (8) using the Cauchy–Schwarz inequality, Parseval's theorem, and the Fourier transform correspondence $\partial u/\partial x \leftrightarrow i2\pi v$.

The field distribution at the focal plane of an ideal lens is proportional to a Fourier transform. The Fourier transform variable is related to the physical variables by the equation (15)

$$v = \frac{x'}{\lambda f}, \tag{9}$$

where f is the focal length, λ is the wavelength, and x' is the coordinate in the focal plane. Using Eq. (9), Eq. (4) can be written as

$$\frac{\Delta_x \Delta_{x'}}{\lambda f} \geq \frac{1}{4\pi}. \tag{10}$$

Further, converting the widths on the left side of the above equation to a $1/e^2$ radius for the input Gaussian beam, a full radius for the shaped beam, and multiplying both sides by $2\sqrt{2\pi}$ Eq. (10) becomes

$$\frac{2\sqrt{2\pi} r_0 y_0}{f\lambda} \geq 0.69. \tag{11}$$

where r_0 is the Gaussian beam radius and y_0 is the shaped profile radius. The left-hand side of the above inequality is the β defined in Sec. III.D. It should be noted that this result is strictly true for separable input and output beam functions. In fact, for the non-separable case the constant on the right-hand side would be greater because the radii are averaged with respect to the orthogonal coordinate.

Applying the uncertainty principle to the beam shaping problem requires some thought. As derived above, the inequality is strictly applicable to the product of the input and output beam radii. However, it does prohibit focusing a beam to a radius smaller than the lower limit given by the inequality. It is reasonable to expect that good shaping results would not be obtained for beam radii determined by the equality in Eq. (11). In fact, it is expected that good shaping would not be obtained unless the equality was exceeded by a factor of 3 or more. This is supported by the fact that the equality in Eq. (4) obtains only when the input and output beam are both Gaussian beams that are Fourier transform pairs (16). Further, since the beam shaping problem addressed here can be expressed as a convolution in the output plane of the Fourier transform of the Gaussian input beam and the Fourier transform of phase function, one would expect that the uncertainty principle would also be an indication of the ability to achieve steep skirts on the edges of the beam profile. Finally, since the constant on the right-hand side of Eq. (11) would be greater for a non-separable function it would be more difficult to produce flat circular beams with steep skirts than it would be to produce corresponding beams with a square cross section. These results are compatible with our numerical modeling of the beam shaping problem, see Sec. IV.

D. Stationary Phase Solution

Solutions to the problem defined by Fig. 2 and Eq. (1) can be obtained by application of the method of stationary phase. Before giving the stationary phase solutions, we present a brief introduction to the one-dimensional stationary phase formula (18,19). Stamnes (20) provides an extensive discussion of the method of stationary phase and its application to diffraction problems. Walther (21) applies the method of stationary phase to the wave theory of lenses. A treatment of stationary phase is given in Chapter 2.

The method of stationary phase gives an asymptotic approximation to integrals of the form

$$I(\beta) = \int_a^b f(x) e^{i\beta\phi(x)}\, dx, \tag{12}$$

where β is a dimensionless parameter. The first term in the asymptotic phase approximation to the integral in Eq. (12) is given by

$$I_c(\beta) \sim e^{i\{\beta\phi(c)+\mu\pi/4\}} f(c) \left[\frac{2\pi}{\beta|\phi''(c)|}\right]^{1/2}, \tag{13}$$

where primes denote derivatives,

$$\mu = \operatorname{sign} \phi''(c), \tag{14}$$

and c is a simple stationary point defined by

$$\phi'(c) = 0, \ \phi''(c) \neq 0. \tag{15}$$

Equation (13) is commonly referred to as the *stationary phase formula*. Similar results are obtained in two dimensions with $\phi''(c)$ replaced by the Hessian matrix for ϕ.

The essence of the beam shaping problem is to equate $|I_c(\beta)|^2$ with the desired irradiance in the output plane of Fig. 2. That is, the magnitude squared of the right-hand side of Eq. (13) is equal to the desired output irradiance. Using this condition with Eq. (15) leads to a second order differential equation for the beam shaping phase function $\phi(x)$. The details of obtaining the explicit form of the differential equation from Eq. (13) and Eq. (15) are rather tedious (25) (see Chapter 2). Care must be taken with respect to the absolute value of ϕ'' in the denominator of Eq. (13). This condition requires that the phase $\phi(x)$ is a convex function, a function whose second derivative is either positive or negative everywhere. This turns out not to be a problem for the case of mapping a Gaussian into a *rect* function. This can be seen from the geometrical optics representation of the beam shaping problem illustrated in Fig. 4. In the figure, the input beam consists of collimated rays whose density is accurately scaled to be proportional to a Gaussian irradiance profile. These rays are bent, in the shaping plane, to form a uniform irradiance distribution in the output plane. Near the shaping plane one can form a phase front for the converging beam by integrating the reciprocal of the slope of the rays (wave normals). The curved line in the figure represents the phase front. It can seen that the slope of the phase front, derivative of the phase function, is a monotonic function giving a phase function with a positive (or negative) second derivative. It may be noted the construction of Fig. 4 provides an algorithm for a geometrical optics solution to the beam shaping problem (9). When the phase of the phase function is not a convex function over the entire input beam the problem becomes more difficult. In this case, a solution to the problem can be obtained by dividing the input beam into regions over which

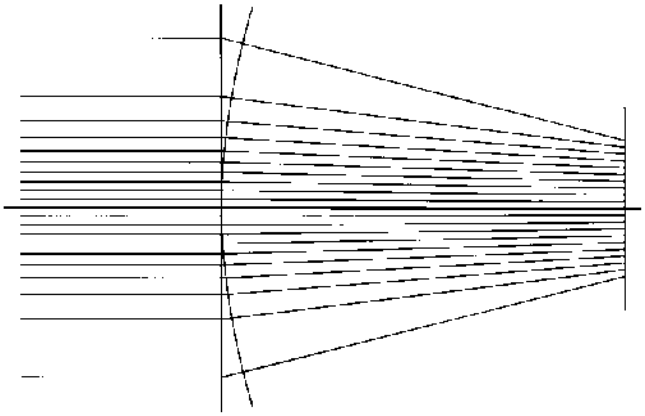


Figure 4 Geometrical optics representation of the Gaussian to rect function beam shaping problem.

the phase function is convex and combining the solutions for these regions in a seamless manner. In practice, this may be a very difficult problem.

A very simple and useful application of the method of stationary phase is outlined in Appendix A. In this appendix, we treat the problem of the lossless mapping of a uniform amplitude and phase beam into a uniform irradiance profile at the focal plane of a transform lens using a phase element as shown in Fig. 2. In this case, the determination of the differential equation is quite simple. The solution for the phase element is just a quadratic phase function, a thin (ideal) lens. This problem provides the diffraction theory basis for the "fly's eye lens" beam integration system (22,23) (see Chapter 7). If a small array of these elements were placed before the Fourier transform lens, the uniform patterns for each phase element would be superimposed in the focal plane of the Fourier transform lens. This solution is very closely related to the problem if Fourier analysis of chirped or linear frequency modulated signals occurring in synthetic aperture radar system theory (24).

In two dimensions, the general form of the equation to be solved is

$$F(\omega_x, \omega_y) = \frac{1}{2\pi} \int_{\infty}^{\infty} \int_{-\infty}^{\infty} f(\xi, \eta) \exp(i(\beta\phi(\xi, \eta) - \xi\omega_x - \eta\omega_y))\, d\xi\, d\eta, \quad (16)$$

where $\xi = x/r_i$ and $\eta = y/r_i$ are normalized input variables with r_i defining the length scale, and $\omega_x = x_f/r_0$ and $\omega_y = y_f/r_0$ are normalized output variables in the focal plane of the Fourier transform lens with r_0 defining the length scale. The stationary phase solution improves asymptotically with increasing dimensionless parameter $\beta = 2\pi r_i R_0/f\lambda$ where R_0 is the size of

the output beam, λ is the optical wavelength and f is the focal length of the transform lens.

The stationary phase evaluation of integrals of the type given by Eq. (16), generally, leads to second order partial differential equation for the phase function ϕ. The resulting partial differential equation can then be solved for ϕ, subject to an energy boundary condition determined by Parseval's theorem. The partial differential equation reduces to a second order ordinary differential equation for the separable and circularly symmetric problem. The optical element is then designed to realize $\beta\phi$.

The stationary phase evaluation of Eq. (16) allows for the mapping of arbitrary single mode laser beams into arbitrary irradiance profiles using the system in Fig. 2. However, some irradiance profiles may be more mathematically difficult to realize. Romero and Dickey (25) (see Chapter 2) have obtained solutions for the separable problem of converting circular Gaussian beams to uniform profiles with rectangular cross-sections and the problem of converting circular Gaussian beams to uniform beams with circular cross-sections.

For a circular Gaussian beam input, the problem of turning a Gaussian beam into a flat-top beam with rectangular cross section is separable. That is, the solution is the product of two one-dimensional solutions. β and $\phi(\xi)$ are thus calculated for each dimension. The phase element will then produce the sum of these phases $(\beta_x\phi_x(x) + \beta_y\phi_y(y))$. The corresponding one dimensional solution for ϕ is (25)

$$\phi(\xi) = \frac{\sqrt{\pi}}{2}\,\xi\,\mathrm{erf}(\xi) + \tfrac{1}{2}\exp(-\xi^2) - \tfrac{1}{2}, \tag{17}$$

where

$$\xi = \frac{\sqrt{2}x}{r_0} \quad \text{or} \quad \xi = \frac{\sqrt{2}y}{r_0},$$

and $r_0 = 1/e^2$ radius of the incoming Gaussian beam.

The solution for the problem of turning a circular Gaussian beam into a flat-top beam with circular cross-section is (25)

$$\phi(\xi) = \frac{\sqrt{\pi}}{2}\int_0^{\xi} \sqrt{1 - \exp(-\rho^2)}\,d\rho, \tag{18}$$

where

$$\xi = \frac{\sqrt{2}r}{r_0},$$

and $r =$ radial distance from the optical axis.

As previously mentioned, the quality of these solutions depend strongly on the parameter β. For the two solutions given in Eq. (17) and Eq. (18), β is given by

$$\beta = \frac{2\sqrt{2\pi} r_0 y_0}{f\lambda}, \tag{19}$$

where:

$r_0 = 1/e^2$ radius of incoming Gaussian beam,
$y_0 =$ half-width of desired spot size (the radius for a circular spot, or half the width of a square or rectangular spot).

Examples of the dependence on β will be given later.

E. Positive and Negative Solutions

An interesting property of the configuration in Fig. 2 is that if ϕ is any even function solution then $-\phi$ is also a solution. If the input beam and the output beams are even functions then ϕ will be an even function. This is easily demonstrated in one dimension and the development is readily extended to two dimensions. In simplest form the input/output optical fields in Fig. 2 are related by the Fourier transform

$$G(\omega) = \int f(x) e^{i\beta\phi(x)} e^{-i\omega x}\, dx, \tag{20}$$

where $G(\omega)$ is the output field and $f(x)$ is the input field. In this equation the integrand can be expanded to give

$$\begin{aligned} G(\omega) &= \int f(x)[\cos\beta\phi(x) + i\sin\beta\phi(x)][\cos\omega x - i\sin\omega x]\, dx, \\ &= \int f(x)\cos\beta\phi(x)\cos\omega x\, dx, \\ &= \int f(x)\cos[-\beta\phi(x)]\cos\omega x\, dx. \end{aligned} \tag{21}$$

The equivalence of the positive and negative solutions follows from the fact that the odd terms in the integrand integrate to zero and the cosine function is an even function. This result has practical implications for system design. The positive solution produces a beam that converges to a small diameter after the output plane, and the negative solution gives a beam that converges to a small diameter before the output plane. It should be noted that this result is independent of the type of solution method such as the method of stationary phase. It depends only on the symmetry assumptions stated at the

start of this section. However, the convexity problem discussed following Eq. (15) is related in that it allows for both a positive and a negative solution.

F. Quadratic Phase Correction

The solutions described in Sec. II.D assume that the input Gaussian beam has a uniform (constant) phase at the beam shaping element. For a Gaussian beam this condition is obtained at the beam waist, and it is not convenient or practical to always locate the beam waist at the shaping element. One solution is to build into the shaping element a phase conjugate to the input beam phase. A more practicable solution is to exploit the fact that the Gaussian beam phase causes a shift in the location of the output plane. That is, the desired profile is located a distance from the focal plane of the transform lens. There is also a slight magnification associated with the shift of the output plane. These assertions can be proved using the Fresnel integral and the general form for Gaussian beams.

Gaussian beams propagate with a phase function given by (26)

$$f(x,y) = e^{(-\sigma+i\gamma)(x^2+y^2)}, \tag{22}$$

where σ and γ are functions of the distance from the beam waist, and $\gamma = 0$ at the beam waist. The solutions in Sec. II.D assume that $\gamma = 0$ and the output is the Fourier transform of the product of a Gaussian and the beam shaping phase function given by

$$U(x_f,y_f) = Ae^{i(k/2f)(x_f^2+y_f^2)} \iint e^{-\sigma(x^2+y^2)} e^{i\beta\phi} e^{i(2\pi/\lambda f)(xx_f+yy_f)}\,dx\,dy, \tag{23}$$

where x, y are the input coordinates and x_f, y_f are the output coordinates in the focal plane of the transform lens. We can arrive at an equivalent expression by applying the Fresnel integral to the field after the lens. If the lens function is given by

$$t_l(x,y) = e^{-i(k/2f)(x^2+y^2)}, \tag{24}$$

the Fresnel integral gives

$$U(x_o,y_o) = \frac{e^{ikz}}{i\lambda z} e^{i(k/2z)(x_o^2+y_o^2)} \iint e^{(-\sigma+i\gamma)(x^2+y^2)} e^{i\beta\phi} e^{[(ik/2z)-(ik/2f)](x^2+y^2)} \times e^{-i(k/z)(xx_o+yy_o)}\,dx\,dy, \tag{25}$$

where x_o, y_o are coordinates in a plane a distance z from the transform lens. If $z = z_o$ is the solution to

$$\gamma - \frac{k}{2f} + \frac{k}{2z} = 0, \tag{26}$$

then Eq. (25) reduces to

$$U(x_o, y_o) = \frac{e^{ikz_o}}{i\lambda z_o} e^{i(k/2z_o)(x_o^2+y_o^2)} \iint e^{-\sigma(x^2+y^2)} e^{i\beta\phi} e^{i(k/z_o)(xx_o+yy_o)}\, dx\, dy. \tag{27}$$

The integral in Eq. (27) is a scaled version of that in Eq. (23). Thus both equations produce the same intensity pattern except for scaling and amplitude factors (phase factors do not effect the irradiance pattern).

III. COLLIMATED UNIFORM IRRADIANCE BEAMS

A drawback of existing beam shaping systems is the limited depth of field. The uniform profile appears only at the target plane, and the profile quickly degrades beyond it. This is due to the fact that the configuration of Fig. 2 cannot produce a beam with both a uniform phase and a uniform amplitude without including a loss mechanism (amplitude control). This can readily be seen from the Fourier transform relation between the field in the output plane and the field just before the transform lens (after the shaping element). For example, for the one-dimensional case, if the beam has a uniform amplitude and phase in the output plane it must be a sinc function in the input plane. A desirable extension would be to create a uniform beam that could propagate for considerable distances. In other words, besides a uniform profile at the target plane, what is desired is a uniform phase front (see Theorem 11 in Sec. II.B of Chapter 2). The uniform profile would then continue to propagate subject only to diffraction effects due to its finite size. In addition, applications such as optical lithography, nonlinear optics, and optical data (image) processing may require beams with uniform phase as well as amplitude.

A. Conjugate Phase Plate

A uniform phase and amplitude beam can be obtained by adding a conjugate phase plate at the output plane of the beam shaping system as shown in Fig. 5. The phase of the conjugate phase plate is designed to cancel the phase of the uniform irradiance beam at the output plane of the beam shaping optics, producing a collimated beam to the right of the output

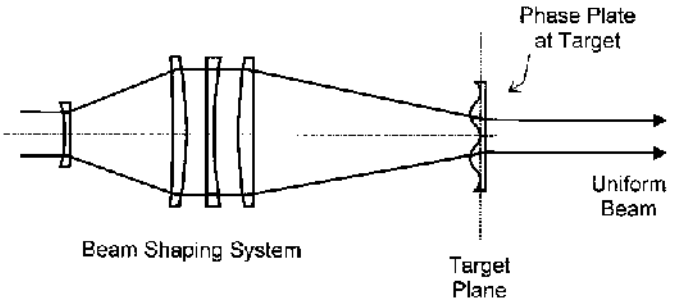


Figure 5 Optical system for creating a uniform phase beam. (From Ref. 14)

plane. Given the properties of the input beam and the solution for the shaping element, it is theoretically possible to compute the phase of the conjugate phase plate. However, in some cases, it might be more practicable to design the beam shaping system and then measure the phase of the shaped beam. The phase plate would then be designed to give the conjugate of the measured phase.

In Appendix B, we show that a beam obeying the scalar wave equation has a minimum root-mean-square (r.m.s.) radius at a plane of uniform phase, and the beam radius as a function of z (the beam axis coordinate) is a quadratic function given by

$$(\Delta\rho)^2 = a + cz^2. \tag{28}$$

where a is the minimum radius squared, $(\Delta\rho_{\min})^2$, in the plane $z = 0$. It can be seen from Eq. (28) that minimizing c minimizes the spread of the beam in the region of the plane of uniform phase. Further, it is shown, using the uncertainty principle, that the Rayleigh range for beams obeying the scalar wave equation is constrained by

$$z_0 \leq \frac{4\pi a}{\lambda} = \frac{4\pi(\Delta\rho_{\min})^2}{\lambda}. \tag{29}$$

The Rayleigh range in this equation is a generalized Rayleigh range defined as the distance (measured from $z = 0$) over which the r.m.s. beam radius increases by a factor of $\sqrt{2}$. Equation (29) is just a quantitative statement of the intuitive concept that to obtain a beam with a large depth of field one wants a large beam width with uniform phase. Since the components of the vector wave equation obey the scalar wave equation, these results can be extended to include solutions of the vector wave equation.

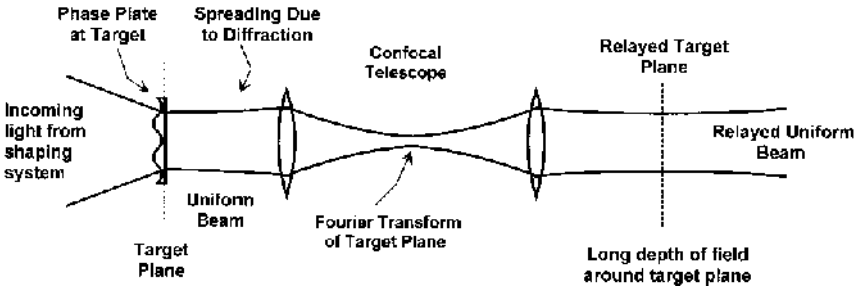


Figure 6 Relay system for extending the depth of field of uniform irradiance beams. (From Ref. 14)

B. Relay Optics

In many applications, especially high-power applications such as material processing, it is not desirable to have a phase plate at the uniform irradiance plane. A solution is to use relay optics to image the beam in the vicinity of the output plane. A relay system for this purpose is illustrated in Fig. 6. The relay optics configuration in the figure has the additional advantages that it doubles the depth of field and can also be used to magnify the size of the shaped beam. These advantages are, of course, obtained at the expense of additional optics.

It should be noted that the relay system in Fig. 6 consists of a two-lens afocal telescope, that is, the lenses are separated by the sum of their focal lengths. A minimum of two lenses are required to image (relay) the output plane while preserving the uniform phase profile. This can be seen from the fact that an afocal telescope produces a collimated output beam when the input beam is collimated. Another approach is to observe that each lens produces the Fourier transform of the field at its front focal plane at the back focal plane. The result is the Four transform of a Fourier transform, which is effectively a Fourier transform followed by an inverse Fourier transform with the coordinates reversed (15). The output beam is symmetric about the relayed target plane, producing a greater distance where the beam maintains the desired tangential dimensions.

IV. DESIGN CONSIDERATIONS

A primary design advantage of this lossless beam shaping technique is that the designer can start with a desired target spot quality and determine the optical system required, rather than designing multiple optical systems in the

hope of producing an acceptable output. This is because the dimensionless quantity β of Eq. (19) completely determines the quality of the spot at the target plane. In other words, different optical configurations and wavelengths will produce the same target spot quality if they share the same value of β. Low values of β produce target spots with more rounded sides and wider skirt regions, while higher values of β more closely approach the geometric ideal of a uniform intensity profile with infinitely steep sides. As Eq. (19) implies, the cost of increasing β involves either increasing the size of the Gaussian beam at the phase element, enlarging the target spot, shortening the focal length of the transform lens, or reducing the wavelength. By considering the application and consulting plots of target quality versus β, the designer can determine the minimum β that will satisfy the needs, and design the most economical system.

Figure 7 shows a standard layout for a beam shaping optical system. For most design situations, the size of the target spot and the wavelength will be determined by the application. The focal length of the transform lens may also be determined by standoff or other considerations, although a minimum focal length will maximize β. The final variable is the Gaussian beam radius at the shaping element. To achieve the desired β, the beam size should be expanded by an afocal telescope, as shown in the figure. With the optical system designed for one target geometry, there are two methods to produce additional target geometries. The first is to change the phase element. With the same expansion and focusing optics, a system could thus produce circular and rectangular beams of several sizes. It should be noted, however, that different target geometries will vary β, and hence spot quality, as determined by Eq. (19). The second method involves changing the focusing, or transform, lens while leaving the telescope and phase element fixed. This change can vary only the target size, not the geometry, but it has the advantage of maintaining a constant target spot quality. The variation in the focal length changes the spot size proportionally, and thus β remains constant.

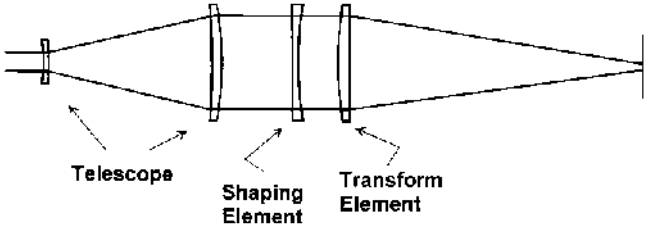


Figure 7 System optical layout. (From Ref. 13)

Once a target spot quality is determined, the required phase profile imparted on the beam by the phase element is then found by multiplying the phase function of Eqs. (17) or (18) by β. This multiplication scales the phase function to the particular geometry of the application. A telescope is then designed to expand the beam to the required value. A transform lens of the required focal length completes the system. The remainder of this section treats the design considerations in more detail, and discusses additional system configurations.

A. Target Spot Quality

Since β determines target spot quality independently of the circumstances of the design, graphs of the beam shape versus particular values of β are useful. The following simulations include system effects such as beam truncation and lens aberrations. They were calculated for a CO_2 laser system ($\lambda = 10.6\,\mu m$) with an $f/42$ to $f/21$ plano-convex lens. Aperture radii were truncated at $2r_0$ in these simulations, where r_0 was the $1/e^2$ radius of the beam. Figure 8a shows a square target spot with $\beta = 4$. The profile is fairly rounded. Figure 8b shows the square target spot with $\beta = 8$. The skirts of this spot have narrowed considerably. Figure 8c shows the square target spot with $\beta = 16$. The skirts of this spot have narrowed further. This system design is beginning to approach the geometric ideal of a uniform profile with infinitely steep sides. The square spot is a special case of the rectangular spot. With a circular input beam, a rectangular output can be produced by varying the β for each axis. For the case of a circular, uniform target spot,

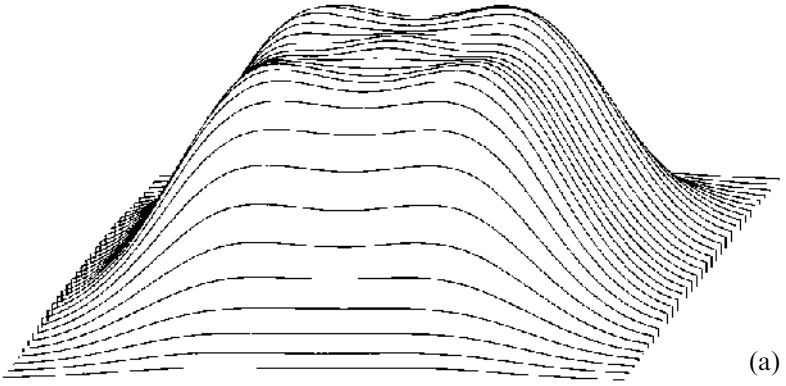
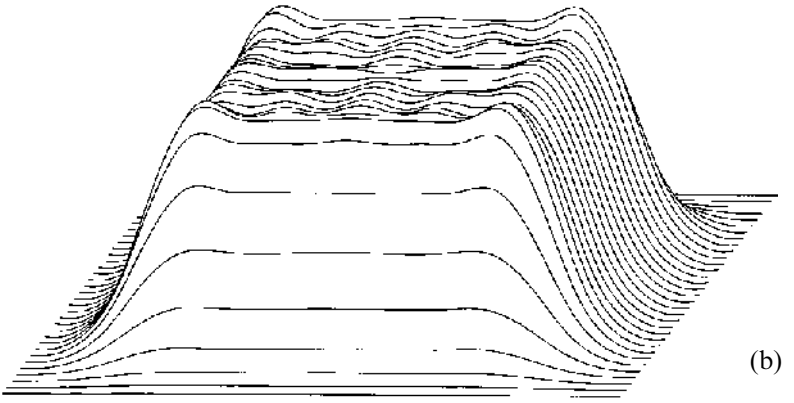
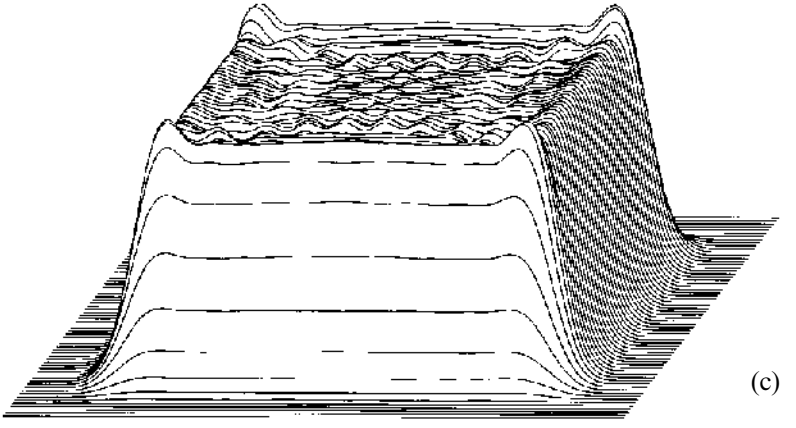


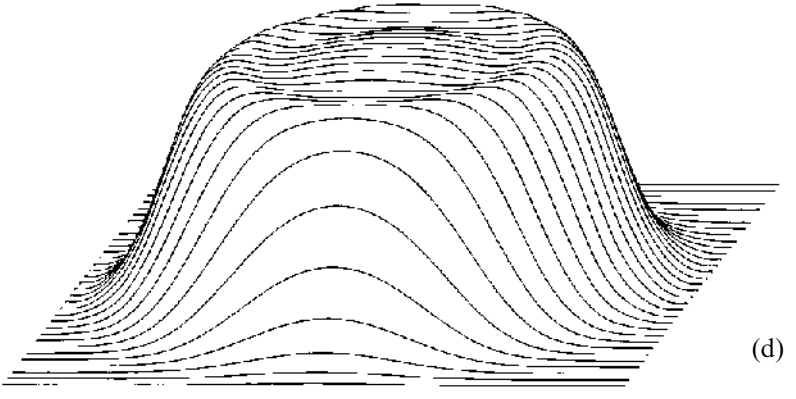
Figure 8 Profiles of square and circular spot geometries. (a) Square profile, $\beta = 4$. (b) Square profile, $\beta = 8$. (c) Square profile, $\beta = 16$. (d) Round profile, $\beta = 8$. (From Ref. 13)



(b)



(c)



(d)

Fig. 8d illustrates the profile for $\beta = 8$ and $3r_0$ truncation. This spot behaves similarly to the square case as β changes. Unlike the square case, however, the circular case exhibits noticeable ripple on the profile as the beam is truncated to $2r_0$.

B. Modeling System Performance

As was shown in Fig. 7, the optical system consists of the phase element and three lenses, each of which can contribute aberrations. A complete way to model system performance is thus desirable. Several optical design packages now offer the ability to input surfaces with general aspheric profiles as polynomial functions of x and y coordinates. For small values of β, the phase profile can be well approximated by varying the thickness of the element. In other words, the phase element acts as a "thin" element for small values of β (15). Some packages allow polynomial phase profiles to be input directly. In either case, the phase functions are fit to a polynomial with appropriate mathematical software, the polynomial multiplied by β, and the phase element inserted into the design package along with the other elements. After tracing a spot diagram with Gaussian apodization of the ray weights, the package then calculates a diffraction-based point spread function for the system, which uses the ray map in the exit pupil. The point spread function provides the diffraction response of the system for a point object. A distant point object produces a planar input wavefront characteristic of a Gaussian beam at its waist. Curvature in the input beam wavefront can be modeled by moving the object point to the appropriate distance. Wavelength variations in the input beam can be modeled with a polychromatic spot diagram and point spread function.

A sufficiently robust design package can model effects due to lens aberrations, beam truncation by optics, beam curvature, and alignment and scaling errors. To avoid aliasing in the point spread function, the spot diagram must sufficiently sample the exit pupil. For βs on the order of 16 or less, the phase profile varies fairly slowly, and most programs can sample it sufficiently. As β increases, however, the phase profile varies more rapidly, and sampling becomes more problematic. However, it should be remembered that β is a measure of how well the system approaches the geometric ideal. For high values of β, therefore, the weighted geometric spot diagram sufficiently models the system performance.

C. Telescope Considerations

As Sec. II.F discusses, displacement of the input Gaussian beam waist from the phase element produces a shift in the location of the output plane and a

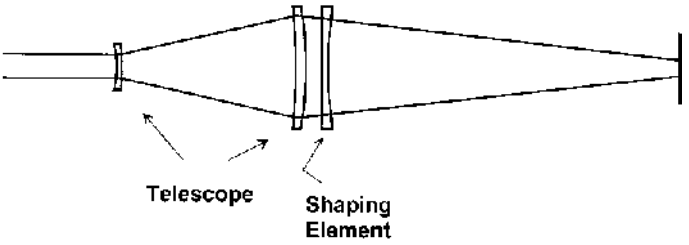


Figure 9 Combined expansion and transform functions in the telescope. (From Ref. 13)

change in scale of the target spot. The telescope can thus be adjusted to shift the target plane to a different location. The telescope can also compensate for curvature in the input Gaussian beam. In this case it is adjusted to place a beam waist at the phase element, with a corresponding slight change in input beam size. Both adjustments should generally result in negligible effects on β and target spot geometry.

Phase elements can be located either before or after the transform element (15). This allows the expansion and transform functions to be combined in the telescope, as shown in Fig. 9. In some situations, it may be necessary to compensate for tolerances in the incoming laser beam diameter. The phase element may then be located behind the transform element and moved along the beam axis until the beam size matches the design size. This movement will scale the target spot size, but β will remain constant.

D. Truncation Effects

For standard optical systems, the effects of truncation on Gaussian beams have been reported in the literature (27). The truncation of the input Gaussian beam by the circular apertures in a beam shaping system will also affect the target profiles. For the square spot, no noticeable degradation is seen for truncation down to $2r_0$. As aperture sizes decrease, however, further ripple becomes apparent. Figure 10 illustrates the effects of $1.5r_0$ truncation on the square spot. For the circular spot, ripple becomes apparent at $2r_0$ truncation, as is shown in Fig. 11. It is interesting to observe why a $2r_0$ circular aperture will affect a circular spot more than the square spot for the same β. For the circular spot, the edge-wave disturbances created by the aperture are symmetric through the system and interfere constructively at the target. For the square spot, however, the disturbances created by a circular aperture are altered by the phase element in a non-symmetric fashion, and thus do not all constructively interfere. This situation is related

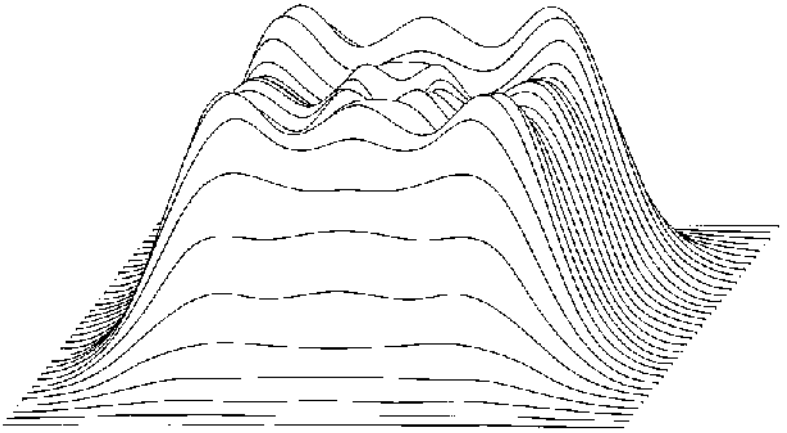


Figure 10 Square profile, $\beta = 8$, apertures $= 1.5r_0$. (From Ref. 13)

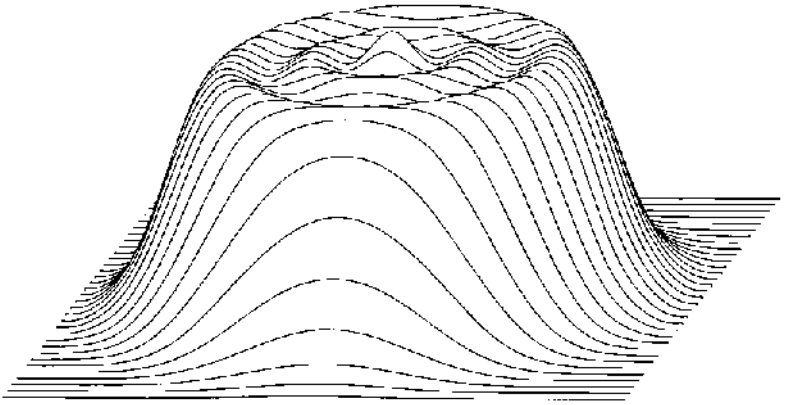


Figure 11 Round profile, $\beta = 8$, apertures $= 2r_0$. (From Ref. 13)

to diffraction by circular versus other apertures (27). For designs producing square or rectangular spots, system apertures of twice the Gaussian beam radius should provide good performance. For designs producing circular spots, system apertures of three times the Gaussian beam radius will be necessary to avoid ripple effects.

E. Positive and Negative Phase Functions

As discussed in Sec. II.E, the phase function has two solutions, positive and negative, for a given configuration. With reference to Eq. (1), ϕ is the phase

delay suffered by a wave in passing through the phase element. This situation is analogous to the phase delays introduced by thin lenses (15). For a positive phase function ϕ, Eq. (17) and Eq. (18) show that the phase delay will vary from zero at the optical axis to increasingly positive values as we move away from the axis. This situation is the same as that for a negative thin lens, and additional insight into the beam shaping system can be gained by viewing the shaping element geometrically, as is illustrated in Fig. 12. The target plane is the focal plane of the transform, or focusing, element and this is where the desired target spot appears. For an element with a positive phase function (a), the beam continues to decrease in size after the target plane. Geometrically, the element has reduced the power of the optical system and behaves much as an aberrated negative lens. For an element with a negative phase function (b), the beam passes through a minimum diameter before reaching the target plane. The element thus geometrically behaves as an aberrated positive lens. In both cases, the spot at the target plane will be identical. The positive phase function (a) has an advantage in depth of field, since the wavefront through the system is closer to planar than for the negative phase function (b). An analogous geometric explanation is that the marginal rays for (a) have smaller angles than (b), thus allowing a larger depth of field. Defocus of the target plane leads to deviations in spot uniformity, and it is treated in the next section. Particular applications may demand that the minimum beam size occur either before

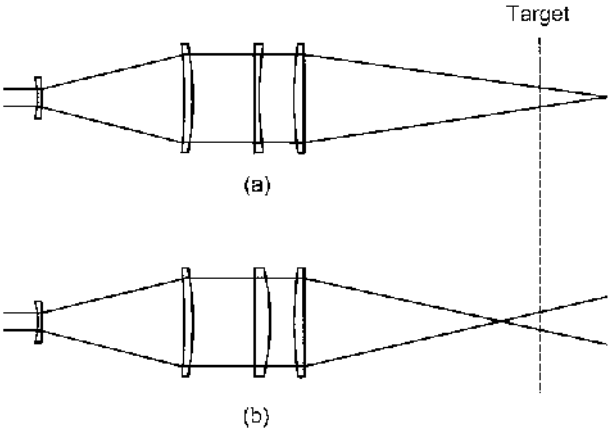


Figure 12 (a) Element with positive phase function acts geometrically as an aberrated negative lens. (b) Element with negative phase function acts geometrically as an aberrated positive lens. The target plane is the focal plane of the final (transform) lens. (From Ref. 13)

or after the target plane. If there is a choice, however, the positive phase function features the least sensitivity to defocus errors.

V. ALIGNMENT AND SCALING ERRORS

Unlike methods based on multifaceted integrators, this lossless beam shaping method is sensitive to alignment errors and variations in the input beam size. Figure 13 shows the effect of decentering the Gaussian beam on the phase element by 0.1 r_0 along one of the element axes. For other elements in the system, decentration is most important where it would move the beam on the phase element. For example, decentration of the negative lens in Fig. 7 would decenter the beam on the phase element and thus produce the effects shown in Fig. 13. On the other hand, decentration of the transform element would have a relatively small impact on the target spot quality. Spot quality is also fairly insensitive to tilt of the phase element, which acts much as a thin plate in this case.

Since the shaping element is designed for a particular input beam size, which in part determines the scaling factor β, it stands to reason that deviations from the design input beam size will affect the target spot. The following cases show the degradation in the square target spot for the $\beta = 8$ design of Fig. 8b, with a positive phase function. In Fig. 14a, the input Gaussian beam size is 10% larger than the size used in the design. The target spot

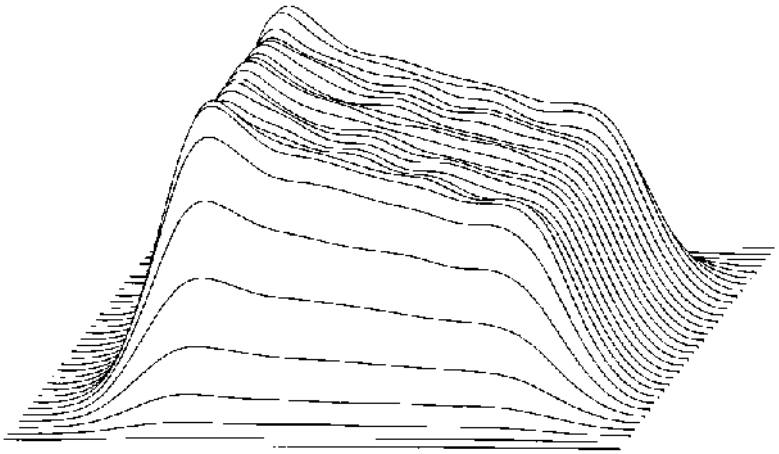


Figure 13 Input beam decentered along element axis by $0.1r_0, \beta = 8$ case. (From Ref. 13)

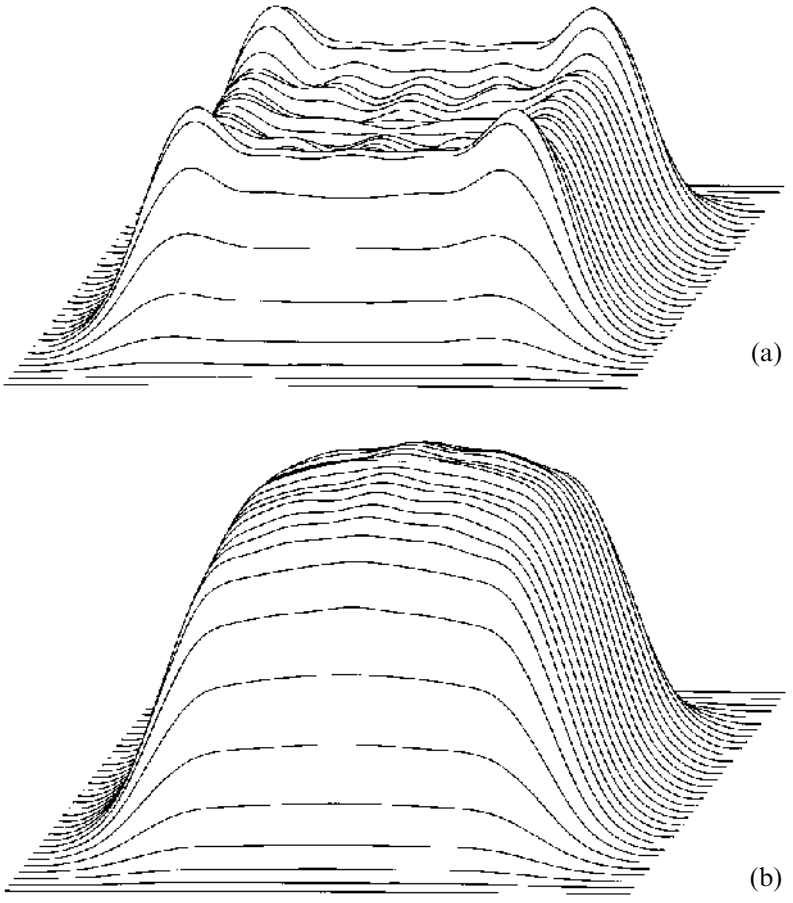


Figure 14 Effects of deviations in input beam size from design values: (a) Input beam 10% larger than design size. (b) Input beam 10% smaller than design size. (From Ref. 13)

shows significantly raised edges. Figure 14b shows the target spot for an input beam size that is 10% smaller than that used in the design. The edges of the spot have rounded off. For many applications, this rounding effect is less detrimental than that caused by the raised edges. Thus, if variation in the input beam size is anticipated for systems with a positive phase function, the element should be designed for a beam somewhere near the upper limit of the size range.

Target plane defocus also affects the quality of the target spot for the following reasons. The beam shaping system uses a lens to transform the

input beam plus the phase function to the desired shape. The output spot exists at the transform plane of the lens, which is also its focal point. In the derivation of the phase element, the problem was to minimize the difference between the desired irradiance at the target plane and that produced by the system. There were no constraints on the phase of the beam at the target or on the beam irradiance outside the transform plane. Since the phase is generally not uniform at the target plane, the shaped beam will not display the symmetry about the target plane characteristic of Gaussian beams at their waist, see Sec. III. The irradiance of the beam will thus deviate from the desired shape when the target plane is moved away from the focal point

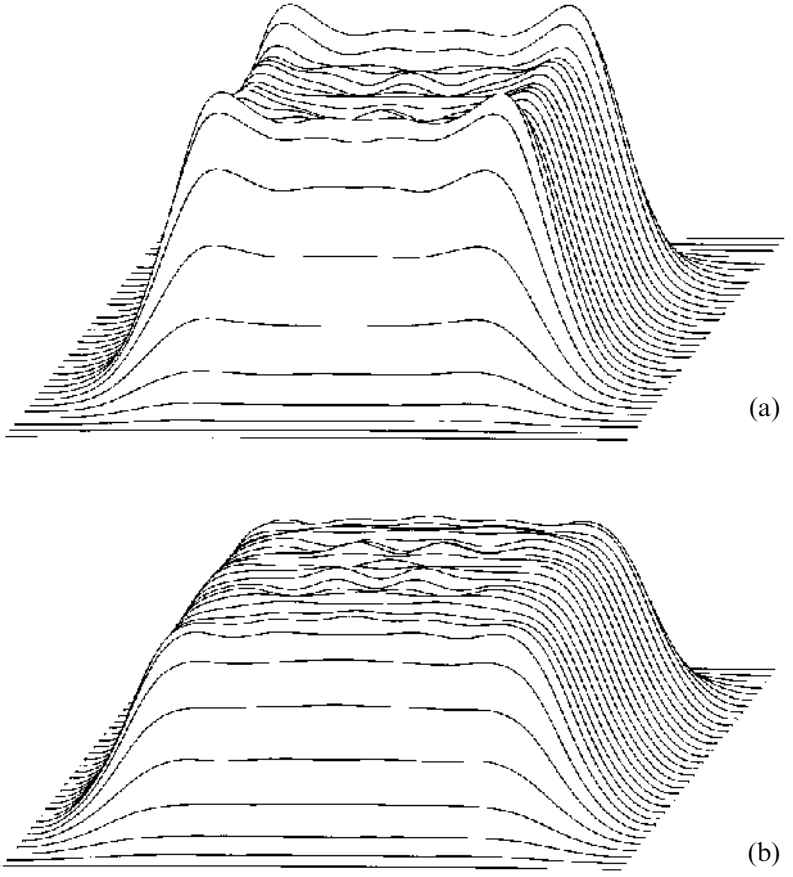


Figure 15 Defocus effects: (a) Target plane defocused by $+f/50$. (b) Target plane defocused by $-f/50$. (From Ref. 13)

of the transform lens. The following cases apply to the $\beta = 8$ design of Fig. 8b, with a positive phase function. Figure 15a shows the effect of moving the target plane away from the transform lens by $f/50$, where f is the focal length of the lens. The spot decreases in size, increases in average irradiance, and the edge areas rise relative to the center. Figure 15b shows the effect of moving the target plane toward the transform lens by $f/50$. Here the spot increases in size and decreases in average irradiance. The spot uniformity remains fairly good, however. Thus, if defocus is anticipated in systems with a positive phase function, the system should be designed for the upper part of the focus range.

As was discussed before, the beam's phase at the target is unconstrained. Thus, in the system illustrated in Fig. 7, the target spot can not simply be collimated with a negative lens to propagate as a flat-top beam. The phase at the target, however, can be computed. A conjugate phase plate placed at the target would cancel these phase differences, and the uniform profile would propagate as a collimated beam, subject to diffraction. On the other hand, if the target spot is simply desired at another location or scale, it can be re-imaged with a conventional afocal telescope as discussed in Sec. III.B.

VI. METHOD OF DESIGN

As stated before, the quality of the target spot can be selected to suit the application, and the necessary optical system parameters calculated directly. In most cases, the size of the Gaussian beam at the phase element will be the free variable that determines β. If the phase element and the optical system are to be studied with an optical design program, the phase function will need to be expressed as a polynomial. With the optical design program, the response of the system to tolerances in beam scaling, beam position, element position, element tilt, and target defocus can be studied. Beam truncation effects can also be modeled if necessary. If tolerances in input beam size or target position are expected in systems with a positive phase function, the target spot will degrade most gracefully if the element is designed for a slightly larger beam than expected with a target plane slightly further away than expected.

In order to facilitate modeling, the phase functions for rectangular and circular spots have been fitted to 10th-order polynomials. The fits are good to $\xi = 3\sqrt{2}$, which is $3r_0$ at the phase element. The form for the rectangular and circular cross section is

$$\phi(\xi) = a_2\xi^2 + a_4\xi^4 + a_6\xi^6 + a_8\xi^8 + a_{10}\xi^{10}, \tag{30}$$

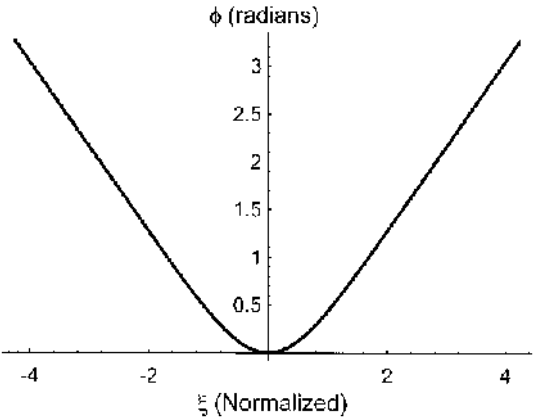
where

Rectangular spot	Circular spot
$a_2 = 4.73974 \times 10^{-1}$	$a_2 = 4.31128 \times 10^{-1}$
$a_4 = -5.50034 \times 10^{-2}$	$a_4 = -4.36550 \times 10^{-2}$
$a_6 = 4.99298 \times 10^{-3}$	$a_6 = 3.65204 \times 10^{-3}$
$a_8 = -2.37191 \times 10^{-4}$	$a_8 = -1.65025 \times 10^{-4}$
$a_{10} = 4.41478 \times 10^{-6}$	$a_{10} = 2.97368 \times 10^{-6}$.

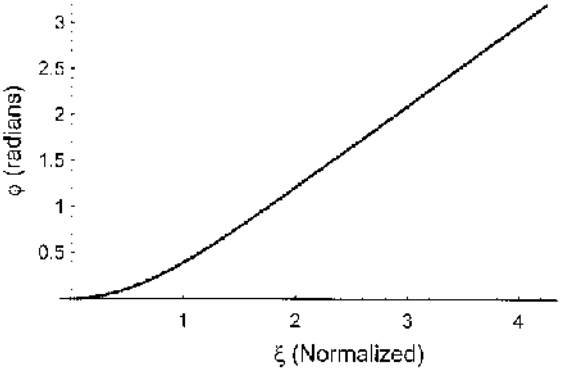
For the rectangular spot, Fig. 16a shows the quality of fit to the original function. No difference between the curves is visible, and they have an r.m.s. variation of 0.0046 radians. As discussed before, the phase function for each axis is multiplied by β to scale it to the desired geometry, and the dimensionless quantities ξ are replaced by actual coordinates according to Eq. (17). The scaled phase functions for each axis are then summed to define the complete phase function. For the circular spot, Fig. 16b shows the quality of fit. The r.m.s. variation between the two curves is 0.0025 radians. ξ is a radial coordinate in this case, and the phase function is radially symmetric.

The following example illustrates the use of the technique to solve an actual problem. Consider the case where a rectangular spot is desired 400 mm away from an optical system. The target spot dimensions are 2 mm (x axis) by 4 mm (y axis). A 10.6 μm laser produces a Gaussian beam with a $1/e^2$ radius of 3 mm, and the optical train is composed of ZnSe ($n = 2.403$). We choose a system layout as shown in Fig. 7, with an $f = 400$ mm focusing lens. From graphs of target spot quality versus β, we choose $\beta = 8$ as a minimum acceptable value. From Eq. (19), we see that the beam radius at the phase element, r_0, is the only unconstrained variable. Also, the x axis will require the most expansion to produce the required β, since its target dimension is smaller. For $\beta_x = 8$, we obtain $r_0 = 6.76$ mm at the phase element, for an expansion ratio of 2.25 from the telescope. We could anamorphically expand the beam to produce the same β for the y axis, but we choose a standard radially symmetric telescope for $\beta_y = 16$.

We choose to generate our phase profile by varying the thickness of the phase element. We wish to develop a polynomial that yields the phase element sagitta, or deviations from a plane at the surface vertex and the surface, as a function of distances from the optic axis. We must thus multiply the coefficients of Eq. (16) by β_x or β_y and convert them to produce sagitta as a function of element coordinates. The following equation gives



(a)



(b)

Figure 16 (a) Rectangular phase function and 10th-order fit. (b) Circular phase function and 10th-order fit. (From Ref. 13)

the sagitta of the phase surface:

$$\begin{aligned}\mathrm{Sag}(x,y) = {} & m_2x^2 + n_2y^2 + m_4x^4 + n_4y^4 + m_6x^6 + n_6y^6 + m_8x^8 \\ & + n_8y^8 + m_{10}x^{10} + n_{10}y^{10},\end{aligned} \tag{31}$$

where

$$m_i = \frac{a_i\lambda\beta_x\left(\frac{\sqrt{2}}{r_0}\right)^i}{2\pi(n-1)} \quad \text{and} \quad n_i = \frac{a_i\lambda\beta_y\left(\frac{\sqrt{2}}{r_0}\right)^i}{2\pi(n-1)}. \tag{32}$$

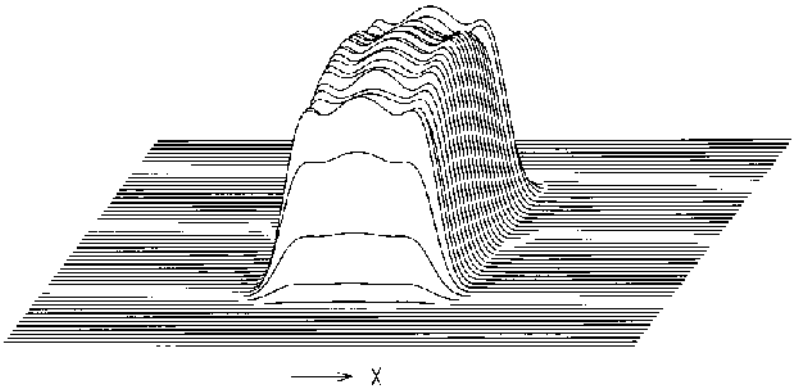


Figure 17 Simulation of output beam produced by sample problem. $\beta_x = 8$ and $\beta_y = 16$. The patch dimensions are 8×8 mm.

The coefficients in this polynomial would be

$$m_2 = 1.994 \times 10^{-4} \qquad n_2 = 3.988 \times 10^{-4}$$

$$m_4 = -1.0127 \times 10^{-6} \qquad n_4 = -2.0255 \times 10^{-6}$$

$$m_6 = 4.0235 \times 10^{-9} \qquad n_6 = 8.0471 \times 10^{-9}$$

$$m_8 = -8.3652 \times 10^{-12} \qquad n_8 = -1.673 \times 10^{-11}$$

$$m_{10} = 6.8144 \times 10^{-15} \qquad n_{10} = 1.3629 \times 10^{-14}.$$

If we choose to build an element with a positive phase function, the sign convention on the sagitta would be such that the phase surface had a concave shape. A simulation of the output spot from this example problem appears in Fig. 17.

VII. EXPERIMENTAL EVALUATION

A beam shaping system was developed for an application that required a long working distance, limited beam sizes, and operation at 10.6 μm. This resulted in a maximum target spot quality given by $\beta = 4.8$. The phase element was fabricated in ZnSe as a 16-level diffractive optic. Figure 18 shows the element profile along the x axis as measured by a stylus profilometer. The desired profile in terms of element thickness is shown by the smooth curve overlaid on the measured profile. In order to match the 2π phase shifts of the diffractive optic, 2π phase shifts (or thickness shifts of

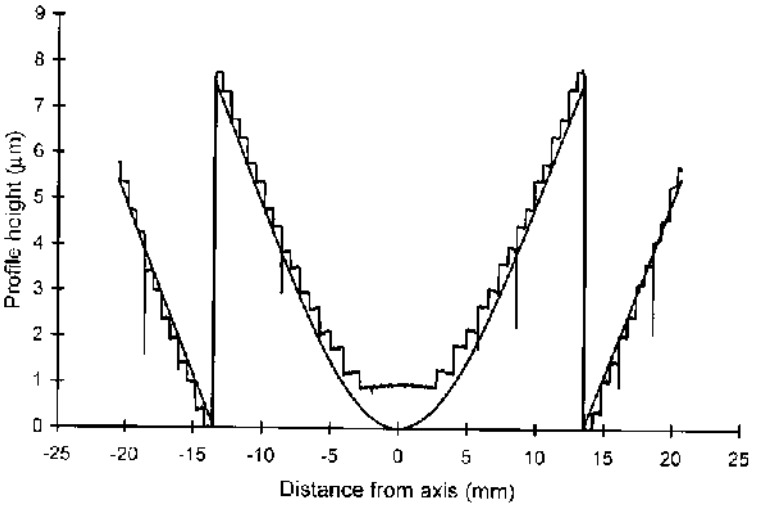


Figure 18 Measured profile of shaping element along x axis. It is a 16-level diffractive approximation to the desired profile shown by the smooth curve. (From Ref. 13)

$\lambda/(n-1)$) were applied to the graph of desired profile as well. Overall, the measured element profile was in reasonable agreement with the desired profile, with the exception of a displacement near the center of the element.

The phase element was tested on a different laser than that for which it was designed, although the measured beam dimensions were within the design goals. The laser beam passed through the beam shaping optical system and on to a target plane. A lens beyond the target plane re-imaged and magnified the target spot onto a pyroelectric array camera. The focus and magnification of the re-imaging system were set by placing a calibrated pinhole at the target plane and adjusting the lens for a sharp image on the camera. The laser beam dimensions at the input of the optical system were determined with orthogonal scanning knife-edges in conjunction with an automated focusing system. This device computed the internal beam waist size and location as well as the beam divergence. It then computed the same parameters for the external laser beam. These quantities then determined the initial beam size at the telescope system.

Before presenting the experimental results, it is instructive to see the predicted spot geometry for the actual beam input parameters. The system was modeled using the computed beam radii for the x and y axes at the first telescope lens. There was a difference of roughly 5% in the computed radius of curvature for the x and y axes, but this was ignored in the modeling. In

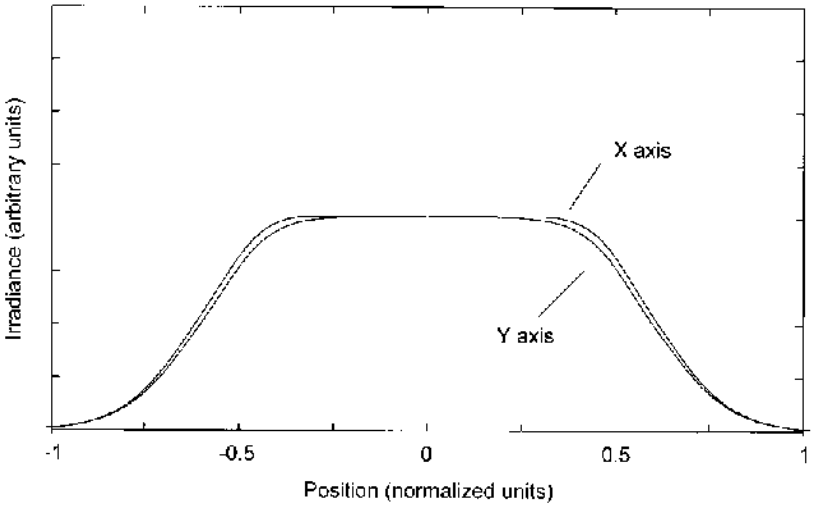
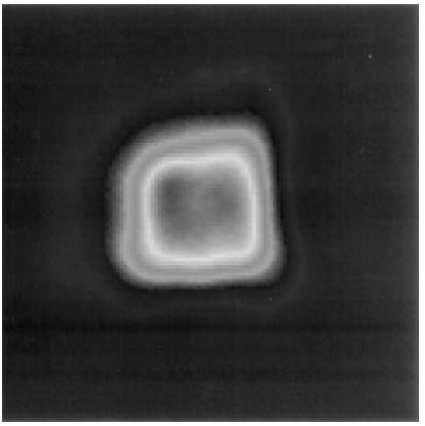


Figure 19 Predicted target spot profiles using the measured beam radii at the input of the optical system. Units are normalized. (From Ref. 13)

Fig. 19, predicted target spot profiles for the x and y axes are shown, scaled to normalized position units. The optical system was initially aligned using a visible reference beam. Final alignment of the beam shaping element was accomplished by viewing the target image with the pyroelectric array camera. Figure 20a shows an image of the target spot when the system was aligned. Each change of shade going in toward the center corresponds to an increase of irradiance. The square appearance is evident. Figure 20b shows a contour plot of this same target spot. Profiles of this image were extracted and are plotted in Fig. 21. These profiles use the same normalized position units as the predicted profiles of Fig. 19. The measured profiles are smaller than predicted, and they deviate somewhat from the desired uniform irradiance. Nevertheless, they show a general agreement with the predicted uniform profiles.

The size difference is partly a function of difficulties in establishing the best target plane. Distance measurements from the transform lens were somewhat inexact, and compensation was necessary for the curvature of the input beam. Also, the re-imaging lens and the camera were mounted independently, so that it was difficult to move the target plane once re-imaging focus was set. Most likely, the re-imaging magnification and focus varied during the alignment process. These uncertainties, coupled with alignment issues for the shaping element, made for difficulties with several independent adjustments during system alignment. It would have

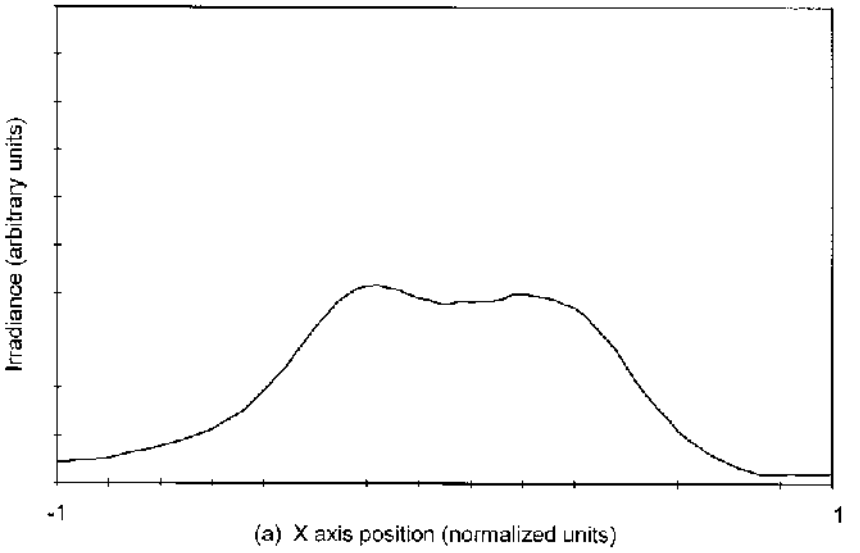


(a)

(b)

Figure 20 (a) Image of the target spot for a $\beta = 4.8$ system. (b) Contour plot of the target spot. Contours have equal intervals of 3.2% of the maximum value. (From Ref. 13)

been best to be able to mount the camera, re-imaging lens, and pinhole together on a common structure, with the pinhole mounted kinematically. When the camera system was used to find the best target plane, the pinhole could be replaced to mark it exactly. Also, a good approximation to the target plane position could be found by removing the shaping element from the system. The focused beam waist would occur very close to the system focal point, and would account for curvature in the input beam. Telescope separation could be adjusted to put the target plane at the desired location, with fewer subsequent adjustments to make during alignment.



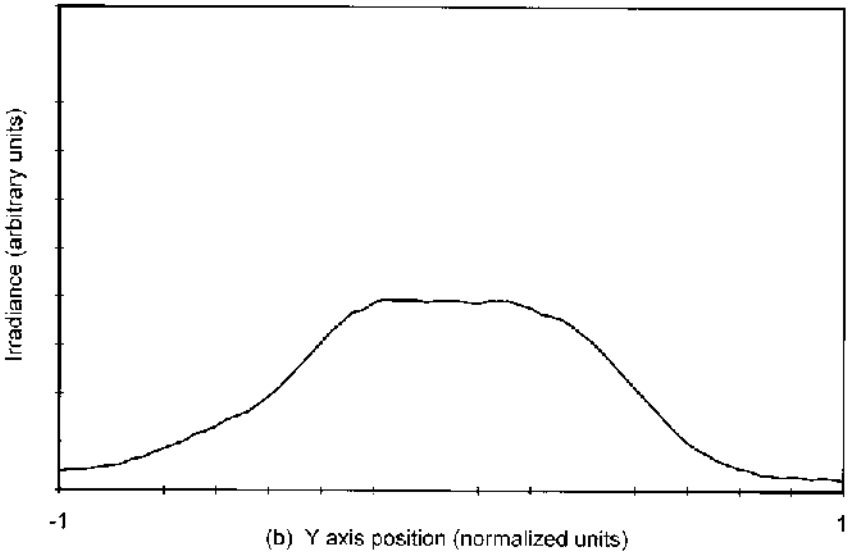


Figure 21 Measured profiles of the target spot, normalized to the same units as the predicted spot. (a) *X* axis. (b) *Y* axis. (From Ref. 13)

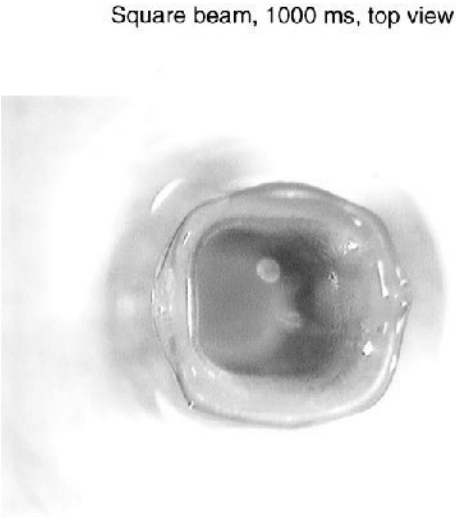


Figure 22 Top view of PMMA exposed to square-shaped beam for 1 s.

Differences in the measured profile uniformity are at least partly the result of two factors. First, the test laser displayed near-field deviations from a Gaussian shape due to its unstable resonator configuration. These near-field deviations tend to be masked in far-field beam measurements due to diffraction effects from the focusing lens. Second, the measured element profile shown in Fig. 18 differed somewhat from the desired profile, leading to a difference in the phase delay applied to the beam.

A common application of beam shaping is for material processing, so this square-shaped target spot was used to burn PMMA material in comparison with a standard Gaussian target spot. Figure 22 shows a top view of PMMA material exposed to the square-shaped beam for 1 s. The square shaped of the removed material is evident. Figure 23 shows side views of PMMA material exposed to both a Gaussian beam and a square-shaped beam for 50 ms. These views were made by cutting the material through the center of the burned spots. The square target beam would be advantageous for situations where a burn needs to be made to a uniform depth. As the figure shows, this target spot could be scanned over the material, and an even burn profile made.

A beam shaping system that produces a circular flat-top profile with $\beta = 20$ has also been fabricated (28). In this design, the focusing and shaping functions were combined in a single diffractive phase element designed for $\lambda = 0.633\,\mu m$. Combining the focusing function implies the addition of a

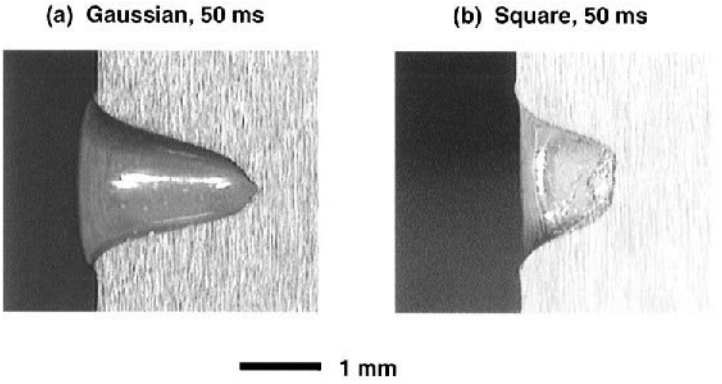


Figure 23 (a) Cut-away view of PMMA exposed to a Gaussian beam for 50 ms. (b) Cut-away view of PMMA exposed to square-shaped beam for 50 ms.

quadratic phase factor to the solution of Eq. (18). The design was fabricated using laser-assisted chemical etching, which produced a smooth profile with 2π phase discontinuities. The addition of the focusing function increased the slope of the phase function to the point that many more phase jumps were needed than in a shaping-only design. Figure 24 illustrates a cross section of the measured target spot for this system. the target spot shows fairly steep skirts and a uniformity error of less than $\pm 5\%$. One likely explanation for the uniformity error is interference caused by the phase discontinuities in the element.

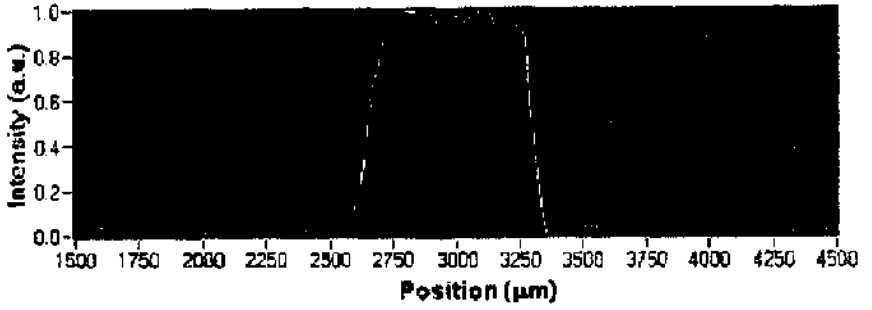


Figure 24 Experimental results for a combined shaping and focusing element. The system was designed to transform a single-mode Gaussian beam into a uniform, circular irradiance profile at the target with $\beta = 20$. This figure shows a cross section through the center of the irradiance profile. (From Ref. 28)

VIII. SUMMARY

Single-mode Gaussian beams can be transformed into circular or rectangular beams with approximately uniform irradiance profiles in a lossless manner by the introduction of an appropriate phase element in conjunction with a Fourier transform lens. This chapter presents a diffraction-theory-based solution for the phase delay obtained using the method of stationary phase. The quality of the target spot was shown to depend on a parameter β that is a function of the input beam size, the target spot size, the focal length of the transform lens, and the wavelength. This dimensionless parameter accounts for diffraction, independently of the particular system. In addition to being a result of applying the method of stationary phase to the evaluation of the diffraction integrals, β was shown to be directly applicable to the general beam shaping problem using the classical uncertainty principle. The use of the β parameter allows the designer to determine the system parameters necessary for a desired target spot quality, rather than iterate through several designs. Once the system parameters are known, either the circular or rectangular phase function can be scaled appropriately. The phase element function can be approximated by a polynomial, which allows standard optics modeling software to predict the effects of system aberrations and tolerances.

Techniques for collimating the shaped beam by using a conjugate phase element are outlined. The phase plate produces a beam with both uniform phase and amplitude, giving a greater depth of field. Bounds on the generalized Rayleigh range for uniform amplitude and phase beams were derived using the uncertainty principle.

Beam shaping system design techniques, based on the theory, were discussed and several numerical examples were presented to illustrate the range of solutions. The sensitivity of system performance to errors in alignment was discussed and illustrated by numerical simulations.

A particular design was implemented in hardware and tested. Experimental results show that the technique produced a square target spot that was close to the predicted profile. The application to of this system to material drilling and ablation is discussed.

REFERENCES

1. S Ream. A convex beam integrator. Laser Focus 15(11): 68–71, 1979.
2. D Dagenais, J Woodroffe, I Itzkan. Optical beam shaping of a high power laser for uniform target illumination. Applied Optics 24(5): 671–675, 1985.
3. VJ Doherty. Design of mirrors with segmented conical surfaces tangent to a discontinuous aspheric base. Proc. SPIE 399: 263–271, 1983.

4. FM Dickey and BD O'Neil. Multifaceted laser beam integrators, general formulation and design concepts. Opt Eng 27(11): 999–1007, 1988.
5. BR Frieden. Lossless conversion of a plane laser wave to a plane wave of uniform irradiance. Applied Optics 4(11): 1400–1403, 1965.
6. W Lee. Method for converting a Gaussian laser beam into a uniform beam. Opt. Comm. 36(6): 469–471, 1981.
7. WB Veldkamp. Laser beam profile shaping with binary diffraction gratings. Opt. Comm. 38(5,6): 381–386, 1981.
8. WB Veldkamp. Laser beam profile shaping with interlaced binary diffraction gratings. Applied Optics 21(17): 3209–3212, 1982.
9. CC Aleksoff, KK Ellis, BD Neagle. Holographic conversion of a Gaussian beam to a near-field uniform beam. Opt Eng 30(5): 537–543, 1991.
10. T Kosoburd, J Kedmi. Beam shaping with diffractive optical elements. Proc SPIE 1971: 390–399, 1993.
11. MT Eismann, AM Tai, JN Cederquist. Iterative design of a holographic beamformer. Applied Optics 28(13): 2641–2650, 1989.
12. MA Golub, M Duparré, EB Kley, R Dowarschik, B Lüdge, W Rockstroh, HJ Fuchs. New diffractive beam shaper generated with the aid of e-beam lithography. Opt Eng 35(5): 1400–1406, 1996.
13. FM Dickey, SC Holswade. Gaussian laser beam profile shaping. Opt Eng 35(1): 3285–3295, 1996.
14. SC Holswade, FM Dickey. Gaussian laser beam shaping: test and evaluation. Proc SPIE 2863: 237–245, 1996.
15. JW Goodman. Introduction to Fourier Optics. New York: McGraw-Hill, Chapt 5, 1968.
16. LE Franks. Signal Theory. Englewood Cliffs, NJ: Prentice-Hall, 1969.
17. RN Bracewell. The Fourier Transform and its Applications. New York: McGraw-Hill, 1978.
18. N Bleistein, RA Handelsman. Asymptotic Expansion of Integrals. New York: Dover, 1986.
19. N Bleistein. Mathematical Methods for Wave Phenomena. Orlando: Academic Press, 1984.
20. JJ Stamnes. Waves in Focal Regions. Bristol: IOP Publishing, 1986.
21. A Walther. The Ray and Wave Theory of Lenses. Cambridge: Cambridge University Press, 1995.
22. X Deng, S Liang, Z Chen, W Yu, R Ma. Uniform illumination of large targets using a lens array. Applied Optics 25(3): 377–381, 1986.
23. H-J Kahlert, U Sarbach, B Burghardt, B Klimt. Excimer laser illumination and imaging optics for controlled microstructure generation. Proc SPIE 1835: 110–118, 1992.
24. RO Harger. Synthetic Aperture Radar Systems: Theory and Design. Orlando: Academic Press, 1970.
25. LA Romero, FM Dickey. Lossless laser beam shaping. J Opt Soc Am: A 13(4): 751–760, 1996.
26. BEA Saleh, MC Teich. Fundamentals of Photonics. New York: John Wiley and Sons, 1991.

27. AE Siegman. Lasers. Mill Valley, CA: University Science Books, Section 18.4, 1986.
28. XG Huang, MR Wang, and C Yu. High-efficiency flat-top beam shaper fabricated by a nonlithographic technique. Opt. Eng. 38(2):208–213, 1999.

APPENDIX A

A simple illustration of the method of stationary phase is given by the problem of mapping a uniform amplitude beam into a uniform irradiance beam. Consider the problem of determining the phase element in Fig. 2 that maps a uniform amplitude and phase beam into a uniform irradiance beam at the focal plane of the transform lens. If the phase of the input beam is not uniform it can be corrected by the phase element.

For simplicity, we treat the problem in one dimension. The field at the focal plane of the Fourier transform lens is given by

$$E(\omega) = \int_{\infty}^{\infty} \text{rect}\left(\frac{x}{\alpha}\right) e^{ik[\phi(x/\alpha)-\omega x/k]}\, dx, \tag{A-1}$$

where α is the input beam width and k is an arbitrary parameter (not the wave number). Letting $\omega/k = \beta$, we can write Eq. (A-1) as

$$E(\omega) = \int_{\infty}^{\infty} \text{rect}\left(\frac{x}{\alpha}\right) e^{ik[\phi(x/\alpha)-\beta x]}\, dx. \tag{A-2}$$

Equation (A-2) can be approximated by the method of stationary phase, giving good results for large k. It is desired that the intensity of the field at the focal plane approximate a rect function. The stationary phase formula

$$I_c(k) \sim \mathrm{e}^{i\{k\psi(c)+\mu\pi/4\}} f(c) \left[\frac{2\pi}{k|\psi''(c)|}\right]^{1/2}, \tag{A-3}$$

where $\mu = \text{sign}\, \psi''(c)$ and c is determined by $\psi'(c) = 0$, gives

$$|E(\omega)|^2 = \text{rect}(c)\, \frac{2\pi}{k\phi''(c)}. \tag{A-4}$$

In obtaining Eq. (A-4) we have used

$$\psi(x) = \phi\left(\frac{x}{\alpha}\right) - \beta x. \tag{A-5}$$

Eq. (A-4) implies that $\phi''(x) =$ a constant. Let

$$\phi\left(\frac{x}{\alpha}\right) = \left(\frac{x}{\alpha}\right)^2 \tag{A-6}$$

and determine c by setting $d\psi(c)/dc = 0$, which gives $c = \beta\alpha^2/2$.

The field at the focal plane is then approximated by

$$E \approx \sqrt{\frac{\pi}{k}}\, e^{-i((\omega^2\alpha^2/4k)-\pi/4)}\, \mathrm{rect}\left(\frac{\omega\alpha}{2k}\right), \tag{A-7}$$

where $\omega = 2\pi x_f/\lambda f$ and the output spot width, W, is determined by $\pi W\alpha/k\lambda f = 1$, giving

$$k = \frac{\pi W\alpha}{\lambda f}. \tag{A-8}$$

The parameter k is similar to β in Sec. II.D, in that the larger the value of k the better the solution. Note that k is a dimensionless constant.

As in the Gaussian to flat-top solution, the phase element that accomplishes the mapping is $k\phi(x)$ with $\phi(x)$ determined by Eq. (A-6) and k determined by (A-8). It is interesting to note that the phase element is a simple (thin) lens in this case. Within the range of Fourier optics the solution is independent of the spacing between the thin lens phase element and the Fourier transform. If a small array of these elements were placed before the Fourier transform lens, the uniform patterns for each phase element would be superimposed in the focal plane of the Fourier transform lens. This is the diffraction theory basis for the "fly's eye lens" beam integration system.

A schematic diagram of the "fly's eye lens" beam integration system is shown in Fig. 25. In the figure, each lenslet array element is focused to a common region by a focusing (Fourier transform lens), effecting beam integration (averaging). Each lens in the array samples the input beam. If the size of the elemental lenses in the array is small enough, the irradiance of each sample will be approximately uniform and, as designed, the output irradiance for each lens element will be approximately uniform. The net output of the beam integration system is a summation of the output from each array element, which should be approximately uniform. However, there will be a fine-structure interference component in the output that will depend on the degree of coherence of the laser system. Beam integration systems are typically used with multimode laser systems. Multi-element imaging systems with two lenslet arrays can be designed to eliminate diffraction altogether (23). A detailed optical analysis of multi-faceted beam integration systems is presented in Chapter 7.

APPENDIX B

Assuming thescalar wave equation, it can be shown that the r.m.s. beam width is a quadratic function of distance along the optical axis. Equations

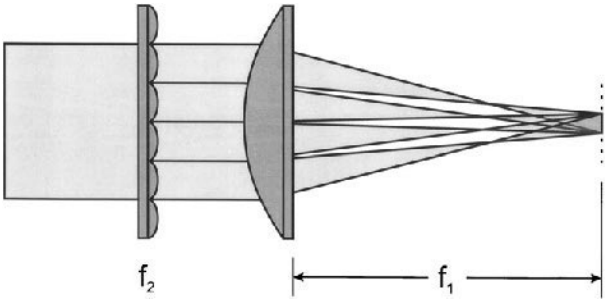


Figure 25 Optical schematic of the "fly's eye lens" beam integration system.

(28) and (29) in Sec. III.A can be obtained as follow. For the case of a monochromatic field, the scalar wave equation reduces to the Helmoltz equation

$$(\nabla^2 + k^2)u(x, y, z) = 0, \tag{B-1}$$

where

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2},$$

and $k = 2\pi/\lambda$. In the Fourier transform domain, the solution to Eq. (B-1) is given (20) by

$$U(f_x, f_x, z) = U(f_x, f_x, 0)e^{iz\sqrt{k^2-(2\pi f_x)^2-(2\pi f_y)^2}}, \tag{B-2}$$

where upper-case letters denote Fourier transformed quantities, f_x, f_y are Fourier transform variables, and $U(f_x, f_y, 0)$ is the Fourier transform of the aperture field (at $z = 0$). The solution of the wave equation given in Eq. (B-2) neglects evanescent waves requiring that $(2\pi f_x)^2 + (2\pi f_x)^2 \leq k^2$.

The mean-square beam width in the coordinate for a beam with its centroid at the coordinate origin is defined by

$$\begin{aligned}(\Delta\rho)^2 &= \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2|u(x, y, z)|^2\, dx\, dy}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|u(x, y, z)|^2\, dx\, dy} \\ &= \frac{\left(\frac{1}{2\pi}\right)^2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left|\frac{\partial U(f_x, f_y, z)}{\partial f_x}\right|^2 df_x\, df_y}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|U(f_x, f_y, z)|^2\, df_x\, df_y}.\end{aligned} \tag{B-3}$$

The integrand in the numerator of the expression after the last equality sign in Eq. (B-3) can be expanded using Eq. (B-2) to obtain

$$\left|\frac{\partial U(f_x, f_x, z)}{\partial f_x}\right|^2 = \left|\frac{\partial U(f_x, f_y, 0)}{\partial f_x}\right|^2 - 2\,\mathrm{Im}\left[\frac{(2\pi)^2 z f_x}{\sqrt{k^2 - (2\pi f_x)^2 - (2\pi f_y)^2}} \times U(f_x, f_y, 0)\ \frac{\partial U^*(f_x, f_y, 0)}{\partial f_x}\right] + \frac{(2\pi)^4 z^2 f_x^2}{k^2 - (2\pi f_x)^2 - (2\pi f_y)^2}\,|U(f_x, f_y, 0)|^2. \tag{B-4}$$

Substituting this result in Eq. (B-3) gives

$$(\Delta\rho)^2 = \frac{\left(\frac{1}{2\pi}\right)^2 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left|\frac{\partial U(f_x, f_y, 0)}{\partial f_x}\right|^2 df_x\, df_y}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |U(f_x, f_y, 0)|^2\, df_x\, df_y} - 2\,\frac{z\left(\frac{1}{2\pi}\right)^2 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{Im}\left[\frac{(2\pi)^2 f_x}{\sqrt{k^2 - (2\pi f_x)^2 - (2\pi f_y)^2}} \times U(f_x, f_x, 0)\frac{\partial U^*(f_x, f_y, 0)}{\partial f_x}\right] df_x\, df_y}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |U(f_x, f_y, 0)|^2\, df_x\, df_y} + \frac{z^2\left(\frac{1}{2\pi}\right)^2 \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{(2\pi)^4 f_x^2}{k^2 - (2\pi f_x)^2 - (2\pi f_y)^2} |U(f_x, f_y, 0)|^2\, df_x\, df_y}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |U(f_x, f_y, 0)|^2\, df_x\, df_y}. \tag{B-5}$$

Equation (B-5) can be written in the simple form

$$(\Delta\rho)^2 = a - bz + cz^2, \tag{B-6}$$

where a, b, and c are defined as the coefficients of the corresponding powers of z. in Eq. (B-5). The main result expressed in Eq. (B-5), or equivalently Eq. (B-6), that the mean-square beam width is a quadratic function z.

Further, if $b = 0$ in Eq. (B-6) the minimum beam width will be in the plane $z = 0$, and Eq. (B-6) becomes

$$(\Delta\rho)^2 = a + cz^2. \tag{B-7}$$

In Eq. (B-7) a is, by definition, the minimum mean-square beam radius. That is, $a = (\Delta\rho_{\min})^2$. Determining the conditions that give $b = 0$ is not a trivial matter. A sufficient, but not necessary condition, is that the term in brackets in the numerator integral defining b is real. This is achieved if $U(f_x, f_x, 0)$ is a real function with the possibility of multiplication by a complex constant. This condition includes beams that have purely even or odd amplitude functions and a uniform phase. It would also include, for example, beams modulated by suitably symmetric square waves.

To minimize the spread of the beam it is desirable to have c as small as possible. However, a and c are related by the uncertainty principle. To apply the uncertainty principle, we can again use the condition $(2\pi f_x)^2 + (2\pi f_x)^2 \leq k^2$ in the integral defining c to obtain

$$\frac{ck^2}{(2\pi)^2} \geq \frac{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_x^2 |U(F_x, f_y, 0)|^2 \, df_x \, df_y}{\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |U(f_x, f_y, 0)|^2 \, df_x \, df_y}. \tag{B-8}$$

Applying the uncertainty principle (see Sec. II.C) one can obtain

$$ac \geq \frac{\lambda^2}{4(2\pi)^2}, \tag{B-9}$$

or equivalently,

$$\sqrt{a}\sqrt{c} \geq \frac{\lambda}{4\pi}. \tag{B-10}$$

One can define a *Rayleigh range* by the condition $a + cz_0^2 = 2a$, giving

$$z_0 = \frac{\sqrt{a}}{\sqrt{c}}. \tag{B-11}$$

This result can be readily checked by comparing it to the Rayleigh range for Gaussian beams. It is well known that a Gaussian function is a minimum uncertainty function (16), that is, equality obtains in the uncertainty principle. Thus, substituting Eq. (B-10) in Eq. (B-11) gives the Rayleigh range for a Gaussian beam as

$$z_0 = \frac{4\pi a}{\lambda} = \frac{4\pi(\Delta\rho_{\min})^2}{\lambda}. \tag{B-12}$$

Noting that $\Delta\rho_{\min} = W_0/2$, where W_0 is the $1/e^2$ beam radius, we obtain $z_0 = \pi W_0/\lambda$, which is the standard result for Gaussian beams (26). Finally, the uncertainty principle gives the inequality

$$z_0 \leq \frac{4\pi a}{\lambda} = \frac{4\pi(\Delta\rho_{\min})^2}{\lambda}, \tag{B-13}$$

as a bound on the Rayleigh range. The components of the vector wave equation obey the scalar wave equation. Hence, these results can readily be extended to solutions of the vector wave equation.

4
Geometrical Methods

David L. Shealy
University of Alabama at Birmingham, Birmingham, Alabama

I. INTRODUCTION

Using geometrical methods to shape a laser beam profile involves application of geometrical optics to solve the optical design problem. Specifically, the laws of reflection and refraction are used along with ray tracing, conservation of energy within a bundle of rays, and the constant optical path length condition to design laser beam profile shaping optical systems. Interference or diffraction effects are not considered as part of the design process in this chapter. That is, only lenses and mirrors are used for the optical components of the laser beam profile shaping systems discussed in this chapter.

There are many diverse applications of lasers in science and technology (1–4). These applications use a variety of the unique properties of lasers, such as the high intensity, coherent, monochromatic light of lasers. For illumination applications, such as in medicine, holography, optical processing, spectroscopy, photography, materials processing, and laser fusion, it is very important for the laser beam to uniformly illuminate the target surface. Truncating a Gaussian beam directly with an aperture or attenuating the beam with a neutral density filter or electrooptics device with an appropriate transversal transmittance profile (5) are straightforward ways to more uniformly illuminate the target surface with a laser beam. However, these intensity apodization techniques result in substantial lose of energy. It is also important to use a simple and efficient beam profile shaping system. Both reflective (6–10) and refractive (11–16) optical systems can be used to shape laser beam profiles as well as holographic and diffraction elements (17–22).

McDermit and Horton (6) presented a method for designing a rotationally symmetric reflective optical system for illuminating a receiver surface in a prescribed manner using a non-uniform input beam profile. Using their method, two mirrors were designed to allow a laser to uniformly heat a flat surface as part of a material testing procedure. Malyak (10) has designed a two-mirror laser profile shaping system where the second mirror is decentered relative to the first mirror to eliminate the central obscuration present in the axially symmetric design. A set of equations is presented for the mirror surface figure for the non-rotationally symmetric laser profile shaping optical system.

Kreuzer (11) has patented a coherent-light optical system to yield an output beam of desired intensity distribution and wavefront shape. Two aspherical optical surfaces are used to achieve this intensity profile shaping. Glass et al. (16) have designed both reflective and refractive laser profile shaping systems for use in laser fusion research via the principal surface of the optical system. Rhodes and Shealy (12,13) derived a set of differential equations using intensity mapping to calculate the shape of two aspherical surfaces of a lens system that expands and converts a Gaussian laser beam profile into a collimated, uniform irradiance output beam. Using their method, two plano-aspherical lenses have been designed, fabricated and used in a holographic projection system (23–25). While seeking to use spherical surfaces for each optical element, Wang and Shealy (26) presented a method for designing an expanded, uniform irradiance profile laser beam using two axial gradient-index lenses. Their design procedure yields the gradient-index profiles as well as the curvatures and separation of the lens surfaces. Cornwell (46,47) introduced non-projective transformations for use in designing reflective and refractive laser beam profile shaping systems and presented a systematic seven-step procedure for designing laser profile shaping systems with either rectangular or polar symmetry. A number of illustrative examples of using non-projective transformations to design both reflective and refractive laser profile shaping systems is presented in Refs. 46 and 47.

A general theory of designing laser profile shaping optical systems as applied to a two-lens optical system is presented in Sec. II. Specific attention is devoted to the application of conservation of energy and the constant optical path length condition as constraints on the optical surface figure of laser profile shaping systems. A summary of a seven–step procedure for using non-projective transformation to design laser profile shaping systems is presented in Sec. III. Applications of these geometrical methods for designing refractive and reflective laser profile shaping optical systems are presented in Secs IV and V. Limitations of geometrical methods for designing laser profile shaping systems are discussed in Secs IV.A.1 and IV.A.2.

II. THEORY OF LASER BEAM PROFILE SHAPING

The concepts of rays, wavefront, and energy propagation are fundamental to understanding and using geometrical optics for shaping laser beam profiles. A brief overview of these concepts is presented in this section. Then, geometrical optics is used to set up several constraint equations which are used to determine the reflective or refractive surface shape or gradient index profile as part of the optical design of laser beam profile shaping optical systems. There are many discussions of geometrical optics in the literature (27–30). In order to determine or optimize the illumination within an optical system, the optical field must be determined throughout the system. The optical field is a local plane wave solution of Maxwell's equations for an isotropic, non-conducting, charge-free medium and is a solution of the scalar wave equation (31, 32)

$$\left(\nabla^2 + n^2 k_0^2\right) u(\mathbf{r}) = 0 \tag{1}$$

where $u(\mathbf{r})$ represents the components of the electric field at any point $\mathbf{r}$, n is the index of refraction at $\mathbf{r}$, $k_0 = \omega/c = 2\pi/\lambda_0$ is the wave number in free space, ω is the frequency of the wave, c is the speed of light, and λ_0 is the wavelength of light. Assume that a solution to Eq. (1) can be written as

$$u(\mathbf{r}) = u_0(\mathbf{r}) \exp[ik_0 S(\mathbf{r})] \tag{2}$$

where $u_0(\mathbf{r})$ and $S(\mathbf{r})$ are unknown functions of $\mathbf{r}$. Equation (1) leads to the following conditions which must be satisfied by $u_0(\mathbf{r})$ and $S(\mathbf{r})$:

$$(\nabla S)^2 = n^2 \tag{3}$$

$$2u_0 \nabla S \cdot \nabla u_0 + u_0^2 \nabla^2 S = 0 \tag{4}$$

where the term proportional to $\left(1/k_0^2\right)$ has been neglected. Equation (3) is known as the eikonal equation and is a basic equation of geometrical optics. The surfaces

$$S(x, y, z) = \text{const.} \tag{5}$$

are constant phase fronts of the optical field and are known as the geometrical wavefronts. For isotropic media, rays are normal to the wavefront. A unit vector normal to the wavefront and along a ray at the point $\mathbf{r}$ is given by

$$\mathbf{a}(\mathbf{r}) = \frac{\nabla S(\mathbf{r})}{|\nabla S(\mathbf{r})|} = \frac{\nabla S(\mathbf{r})}{n(\mathbf{r})} \tag{6}$$

where the eikonal equation was used to simplify this expression for the ray vector $\mathbf{a}$. Additional properties of rays are given below.

Equation (4) is equivalent to conservation of radiant energy within a bundle of rays and leads to the geometrical optics intensity law for propagation of a bundle of rays as illustrated in Ref. 30. Using the vector identity

$$\nabla \cdot (f\,\mathbf{v}) = f\ \nabla \cdot \mathbf{v} + \mathbf{v} \cdot \nabla f, \tag{7}$$

Eq. (4) can be rewritten as

$$\nabla \cdot \left(u_0^2\, \nabla S\right) = \nabla \cdot \left(u_0^2 n \mathbf{a}\right) = 0. \tag{8}$$

Recognizing that the energy density of a field is proportional to the square of the field amplitude u_0^2 and that the intensity I is equal to energy density of the field times the speed of propagation within medium, then Eq. (8) can be written as

$$\nabla \cdot (I\,\mathbf{a}) = 0. \tag{9}$$

Multiplying Eq. (8) by the constant $(c/4\pi)$ for c.g.s. units gives the correct dimensions for intensity (33). Equation (9) expresses conservation of radiant energy for non-conducting medium. Integrating Eq. (9) over a tube surrounding a bundle of rays (34) as illustrated in Fig. 1 gives, after application of Gauss' theorem,

$$I_1\, dA_1 = I_2\, dA_2. \tag{10}$$

Equation (10) expresses conservation of energy along a ray bundle between any two surfaces intersecting the beam and is a basic equation used to design laser beam profile shaping optical systems.

According to geometrical optics, the phase and amplitude of the optical field are evaluated independently. First, the ray paths are evaluated throughout the optical system, which enables computing the phase in terms of the optical path length of rays passing thought the system. Next, the amplitude of the optical field is determined by monitoring the intensity variations along each ray (35–37). This approach for evaluating the phase and amplitude of the optical field is in contrast to the more rigorous wave

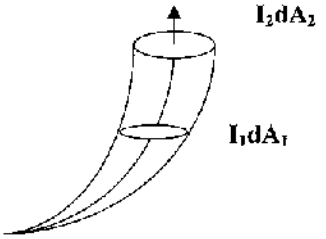


Figure 1 Conservation of energy within a bundle of rays.

optics or electromagnetic theory approach, which involves solving coupled partial differential equations for the complex electromagnetic fields where the phase and amplitudes are interdependent.

Rays generally characterize the direction of the flow or propagation of radiant energy, except near foci or the edge of a shadow where interference and diffraction takes place. Thus, a ray is a mathematical construct rather than a physical entity. Snell's law relates the direction of incident and refracted rays at an interface between media of different indices of refraction, which can be written in vector form (38):

$$n(\mathbf{a} \times \hat{\mathbf{n}}) = n'(\mathbf{A} \times \hat{\mathbf{n}}) \tag{11}$$

where $\mathbf{a}$ and $\mathbf{A}$ are unit vectors along the incident and refracted rays, $\hat{\mathbf{n}}$ is a unit vector along the normal to the interface surface with the general orientation of the incident ray, and (n, n') are the indices of refraction of the incident and refracting media. The vector nature of this equation insures coplanarity of rays and interface surface normal, as required by electromagnetic theory. For ray tracing it is convenient to vector multiply Eq. (11) by $\hat{\mathbf{n}}$ and simplify the resulting triple vector product into the form

$$n'\mathbf{A} = n\mathbf{a} + \left[n' \cos i' - n \cos i\right] \hat{\mathbf{n}} \tag{12}$$

where

$$\cos i' = \mathbf{A} \cdot \hat{\mathbf{n}} \quad \text{and} \quad \cos i = \mathbf{a} \cdot \hat{\mathbf{n}} \tag{13}$$

and (i, i') are the angles of incidence and refraction. When mirrors are involved, the refraction ray equations can be used for reflection by setting $n' = -n$ and using the optics sign convention (39). Explicitly, a unit vector $\mathbf{A}$ along a reflected ray is given by

$$\mathbf{A} = \mathbf{a} - 2\hat{\mathbf{n}}(\mathbf{a} \cdot \hat{\mathbf{n}}) \tag{14}$$

where $\hat{\mathbf{n}}$ in a unit normal vector at the point of reflections, and $\mathbf{a}$ is a unit vector along the incident ray.

Rays may also be defined as lines normal to the geometrical wavefront. Wave propagation is commonly described by wavefronts. A wavefront is a surface of constant phase of the wave or optical path length from the source or reference surface. In electromagnetic theory the direction of radiant energy propagation is given by the Poynting vector or cross product of the electric and magnetic fields.

Each ray generally follows the path of shortest time through the optical system according to Fermat's principle which states that a ray from points P to Q is the curve C connecting these two points such that the integral

$$\text{optical path length, } \mathrm{OPL}(C) = \int_C n(x,\ y,\ z)\, ds \tag{15}$$

is an extremum (maximum, minimum, or stationary). The quantity $n(x, y, z)$ is the index of refraction of the medium, and ds is the infinitesimal arc length of the curve. For a homogeneous medium, the optical path length between P and Q is the geometrical path length between the two points times the index of refraction of the medium. In general, the optical path length divided by the speed of light in free space, c, gives the time for light to travel from point P to Q along the ray path C. The ray path C can be determined using the calculus of variations (40). It can be shown (32) that when the index of refraction, $n(\mathbf{r})$, is a smooth function, the ray path C satisfies the differential equation

$$\frac{d}{ds}\left(n(\mathbf{r})\,\frac{d\mathbf{r}}{ds}\right) = \nabla n(\mathbf{r}) \tag{16}$$

where $\mathbf{r}$ is the position vector of any point on the ray. Equation (16) is known as the ray equation and is difficult to solve in many cases. For homogeneous medium ($n = \text{const.}$), the ray path is represented by a straight line

$$\mathbf{r}(s) = \mathbf{a}s + \mathbf{b} \tag{17}$$

where $\mathbf{a}$ and $\mathbf{b}$ are constant vectors, and s is the ray path length. Constant index of refraction materials are used in many optical systems. When the index of refraction is a function of position, such as the axial distance, z, the ray paths are curved. Section IV.B will illustrate how laser profile shaping systems are designed using materials with a gradient index of refraction, $n(z)$. Now, the ray trace equations, conservation of energy along ray bundle, and the constant optical path length condition are used to design optical systems for shaping laser beam profiles.

A. Optical Design and Energy Balance Condition

For the class of applications for which geometrical methods are used to design a laser beam profile shaping system, it is desirable to transform an input laser beam with a Gaussian intensity profile into an output beam with uniform intensity profile while retaining the input wavefront shape. Further, it is often desirable to expand the laser beam diameter. To achieve these design goals, the optical surface shapes or index of refraction profiles may be used as design variables. Consider the geometrical configuration of a refracting laser beam profile shaping system shown in Fig. 2. For applications using monochromatic laser beams, there are no chromatic aberrations present, and it is satisfactory to use the same material (index of refraction) for all lenses. The two curved surfaces (reflecting or refracting, depending on the

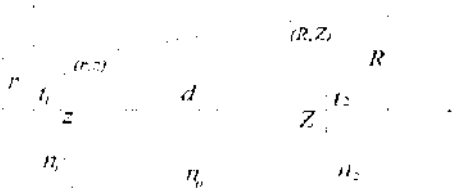


Figure 2 Geometrical configuration of laser expander. (From Ref. 23)

application) are used to satisfy the design conditions for shaping a laser beam profile.

Consider an incoming bundle of rays with an energy density $\sigma(r)$ incident upon a lens system at a radial distance r from the optical axis. Assume the incident beam is a laser in the fundamental, Gaussian TEM_{00} mode with central intensity normalized to unity. Then,

$$\sigma(r) = \exp[-2(r/r_0)^2] \tag{18}$$

where r_0 is the radius of the beam. The conventional definition (41) of the diameter of a laser beam or waist is the location at which the field amplitude is $1/e$ of its peak value. At the beam waist the beam intensity is $1/e^2$ of its axial value. For a Gaussian beam, 86.5% of the total energy of the beam is contained within the waist. This beam leaves the optical system at a radial distance R from the optical axis with an energy density of $\Sigma(R) = \text{const}$.

Integrating Eq. (10) over the incident and exit apertures gives

$$\int_0^{2\pi} d\theta \int_0^r \sigma(r) r \, dr = \int_0^{2\pi} d\theta \int_0^R \Sigma(R) \, R \, dR \tag{19}$$

where reflection and absorption losses have not been considered in Eq. (19). Since $\Sigma(R)$ is a constant, the radius of beam in the exit aperture R can be evaluated by carrying out the integration using Eq. (18) for $\sigma(r)$ to obtain

$$R = \left\{ \frac{r_0^2}{2\Sigma} [1 - \exp(-2r^2/r_0^2)] \right\}^{1/2} \tag{20}$$

where

$$\Sigma = \frac{r_0^2}{2R_{\max}^2} [1 - \exp(-2r_{\max}^2/r_0^2)]. \tag{21}$$

In Eq. (21), $r_{\max}$ is the working aperture of the first lens, and $R_{\max}$ is the corresponding point on the second lens. Next, the shape of the first lens surface s must be determined such that it will refract rays to the second lens

surface S at the radial point R given by Eq. (20), and the shape of the second lens surface S must be determined so that rays will refract to be parallel to the optical axis as the beam leaves the optical system.

Rays are refracted at surface s according to Snell's law, and the direction of refracted ray **A** traveling from the point (r, z) on surface s to the point (R, Z) on surface S is given by Eq. (12) where the incident rays are along the optical axis, that is, $\mathbf{a} = \hat{\mathbf{k}}$. Explicitly, a unit vector along the refracted rays over surface s is given by

$$\mathbf{A} = \gamma \hat{\mathbf{k}} + \Omega \hat{\mathbf{n}} \tag{22}$$

where

$$\Omega = \frac{-\gamma + \{1 + (z')^2(1 - \gamma^2)\}^{1/2}}{[1 + (z')^2]^{1/2}} \tag{23}$$

$$\hat{\mathbf{n}} = (-z'\hat{\mathbf{r}} + \hat{\mathbf{k}})/[1 + (z')^2]^{1/2}, \tag{24}$$

$$\gamma = n/n_0, \tag{25}$$

$$z' = dz(r)/dr. \tag{26}$$

The unit vectors $(\hat{\mathbf{r}}, \hat{\mathbf{k}})$ are along the $r - z$ directions.

The ray path connecting (r, z) and (R, Z) is a straight line according to Eq. (17) with slope given by Eq. (22) and can be written as

$$(R - r)\, A_z = (Z - z) A_r \tag{27}$$

where A_r and A_z are the (r, z) components of **A** given by Eq. (22). Combining Eqs. (22) and (27) yields, after squaring and collecting terms in powers of z':

$$\begin{aligned}(z')^4[\gamma^2(R - r)^2 + (\gamma^2 - 1)(Z - z)^2] - (z')^3(R - r)(Z - z) \\ - (z')^2(1 - \gamma^2)[(R - r)^2 + (Z - z)^2] - 2z'(R - r)(Z - z) \\ - (R - r)^2 = 0\end{aligned} \tag{28}$$

which can be factored and reduced to yield

$$\begin{aligned}[(1 - \gamma^2)(Z - z)^2 - \gamma^2(R - r)^2](z')^2 + 2(R - r)(Z - z)z' \\ + (R - r)^2 = 0.\end{aligned} \tag{29}$$

In Eqs. (28) and (29), R is given by Eq. (20), and z is the solution to the differential equation, Eq. (29), as a function of the entrance aperture coordinate r. Z in Eqs. (28) and (29) is not yet known. For beam profile shaping applications which only need to illuminate a specific surface S with uniform intensity, the equation of the surface S expressed as $Z = Z(R)$ will be

adequate to solve the differential equation (29) for the shape of the surface s. For other beam profile shaping applications which seek also to control the shape of the output wavefront, the constant optical path length condition is used for rays passing through this optical system to determine the shape of surface S. The constant optical length condition yields a functional relationship between Z and z so that the differential equation (29) can be solved. The next section describes how the constant optical path length condition can be used to determine the shapes of the surfaces s and S, when solved simultaneously with Eq. (29).

B. Constant Optical Path Length Condition

In order for the output wavefront to have the same shape as the input plane wavefront, it is necessary for all rays passing through the optical system to have the same optical path length, which is defined in Eq. (15). For uniform index media, the optical path length of a ray passing through an optical system is the sum of the geometrical path length of a specific ray times the index of refraction of the component of the system. For the two-lens beam shaping system shown in Fig. 2, the optical path length $(OPL)_0$ of a ray passing along the optical axis is

$$(OPL)_0 = nt_1 + n_0 d + nt_2. \tag{30}$$

For an arbitrary ray of height r, the optical path length is

$$(OPL)_r = nz + n_0[(R-r)^2 + (Z-z)^2]^{1/2} + n(t_1 + d + t_2 - Z). \tag{31}$$

Recall that the origin of the z-axis remains fixed at the vertex of the first optical surface of the system. The constant optical path length condition is satisfied for this optical system by requiring

$$(OPL)_0 = (OPL)_r. \tag{32}$$

Combining Eqs. (30)–(32) leads to

$$n_0[(R-r)^2 + (Z-z)^2]^{1/2} = n(Z-z) - d(n-n_0), \tag{33}$$

which can be viewed as a quadratic equation for $(Z-z)$ as a function of entrance pupil aperture radius r. After squaring Eq. (33) and collecting terms, the solution of the resulting quadratic equation for $(Z-z)$can be written as

$$(Z-z) = \frac{n(n-n_0)d + [n_0^2(n-1)^2 d^2 + (n^2 - n_0^2)(R-r)^2]^{1/2}}{n^2 - n_0^2} \tag{34}$$

where the positive sign of the radical has been used so that the solution reduces to the appropriate value of $(Z - z) = d$ when $r = R = 0$. Several approaches for solving the differential equation (29) and Eq. (34) are presented in Sec. II.C.

C. Solution of the Differential Equations

The goal of the optical design of a laser beam profile shaping system is to define the optical components adequately so that the system can be analyzed, fabricated, and tested. This generally requires specification of the shapes of and spacing between the optical surfaces as well as the index of refraction of all the media. For the refracting beam expander system illustrated in Fig. 2, the shape of surfaces s and S must be determined. Equations (20), (29), and (34) can be solved numerically for $R(r)$, $z(r)$, $Z(r)$ when n, d, t_1, t_2 are given. In contrast to conventional optical design, which optimizes the design parameters to minimize a merit function, the present method of solving differential equations defines the optical surfaces s and S by producing tables of numerical data $(r,\ z)$ for surface s and $(R,\ Z)$ for surface S. This resulting surface data can be fit to equations for the surfaces s and S, which will be discussed in more details in Sec. II.D. It does not seem possible to solve these differential equation analytically for $z(r)$ and $Z(R)$.

Specifically, solving Eq. (29) as a quadratic equation for z'gives

$$z' = \frac{-(R-r)(Z-z) \pm n(R-r)\sqrt{(Z-z)^2 - (R-r)^2}}{(1-n^2)(Z-z)^2 - n^2(R-r)^2} \tag{35}$$

where the positive solution for z' is used for the laser shaping lens configuration shown in Fig. 2 so that the first lens is divergent. For a specific lens system, such as defined in Table 1, the energy density Σ of the output beam is computed from Eq. (21); the height of a ray R at the second lens for each ray in the entrance pupil with height r is computed from Eq. (20); $(Z - z)$ is computed from Eq. (34); and z' is computed from Eq. (35). Then, z is determined by numerically integrating z' from Eq. (35), and Z is evaluated from Eq. (34). A FORTRAN program for computing (r, z) and (R, Z) is given in Appendix A of Ref. 24. In these studies, typical numerical errors are less than 10^{-15}.

It is interesting to note that combining Eqs. (29) and (34) permits z' to be expressed as a function of r, thus, enabling $z(r)$ to be evaluated by an integration. Then, the shape of S is computed from Eq. (34). Cornwell (46) and Malyak (10) reach a similar conclusion. Namely, the shape of the first

Table 1 Parameters of Laser Shaping System

Wavelength of laser, $\lambda = 411.57\,\text{nm}$ (HeCd laser)
Radius of the working aperture of the primary lens, $r_1 = 8\,\text{mm}$
Radius of the working aperture of the secondary lens, $R_1 = 12.5\,\text{mm}$
Distance between the lenses, $d = 150\,\text{mm}$
Surrounding medium is air, $n_0 = 1.0$
Index of refraction of the lenses, $n = 1.43916$ (CaF_2)
Thickness of the lenses, $t_1 = t_2 = 10\,\text{mm}$

element can written as

$$z(r) = \int f(r)\,dr + C \tag{36}$$

where C is a constant, and $f(r)$ is a known function. The shape of the second surface can be computed from the following expression:

$$Z(r) = z(r) + g(r) \tag{37}$$

where $g(r)$ is another known function. According to Ref. 46, the optical surfaces of both reflective and refractive laser profile shaping systems with collimated input and output beams satisfy equations in the form of Eqs. (36) and (37).

Explicitly for the refractive laser profile shaping system shown in Fig. 2 and discussed in Secs II.A and II.B, it follows from Eqs. (20) and (34) that

$$Z = z + \frac{n(n-n_0)d + \left[n_0^2(n-1)^2d^2 + (n^2-n_0^2) \times\left\{\sqrt{\frac{r_0^2}{2\Sigma}\left[1-\exp\left(-2\,\frac{r^2}{r_0^2}\right)\right]} - r\right\}^2\right]^{1/2}}{n^2-n_0^2}. \tag{38}$$

Comparing the last two equations gives

$$g(r) = \frac{n(n-n_0)d + \left[n_0^2(n-1)^2d^2 + (n^2-n_0^2) \times\left\{\sqrt{\frac{r_0^2}{2\Sigma}\left[1-\exp\left(-2\,\frac{r^2}{r_0^2}\right)\right]} - r\right\}^2\right]^{1/2}}{n^2-n_0^2}. \tag{39}$$

Similarly, Eq. (29) can be written as

$$a(r)z'^2 + b(r)z' + c^2(r) = 0 \tag{40}$$

where

$$c(r) = (R - r) = \left(\sqrt{\frac{r_0^2}{2\Sigma}\left[1 - \exp\left(-2\,\frac{r^2}{r_0^2}\right)\right]} - r\right), \tag{41}$$

$$b(r) = 2c(r)g(r), \tag{42}$$

$$a(r) = (1 - \gamma^2)g^2(r) - \gamma^2 c^2(r). \tag{43}$$

Applying the quadratic formula to Eq. (40) gives

$$z' = \frac{-b(r) \pm \sqrt{b^2(r) - 4a(r)c^2(r)}}{2a(r)} \tag{44}$$

where the positve solution for z' is used for the laser shaping lens configuration shown in Fig. 2. Integrating Eq. (44) leads to Eq. (36) where

$$f(r) = \frac{-b(r) \pm \sqrt{b^2(r) - 4a(r)c^2(r)}}{2a(r)}. \tag{45}$$

D. Analysis of Optical System

After solving the differential equations as described in Sec. II.C, the aspherical surfaces of the lens system are represented by sets of data $[(r, z)$ and $(R, Z)]$. It is convenient for analysis and fabrication of optical elements, if the aspherical surfaces are expressed in terms of analytical functions. The conventional optics surface equation (42) of a conic term plus rotationally symmetric aspherical deformation terms has been used to fit the curved surfaces of a laser shaping system. The optics surface equation can be written in the form

$$z = \frac{cr^2}{1 + \sqrt{1 - (1 + \kappa)c^2r^2}} + \sum_{i=1}^{N} A_{2i}r^{2i} \tag{46}$$

where c (vertex curvature), κ (conic constant), and A_{2i} (coefficients of the polynomial deformation terms) are surface parameters that are determined by the fitting process for each surface. A nonlinear least squares fitting program based on the simplex method (43) has been successfully used (23–25) to represent the optical surfaces of a laser profile shaping system.

Optical surfaces can be planar, spherical, conic, or general aspheric. Equation (46) describes a general aspheric surface with the sag z being an

even function of the surface radial distance from the optical axis. The first term in Eq. (46) represents the conic term. If κ and A_{2i} are zero, the surface will be spherical with radius equal to $1/c$. If κ is not zero, but all A_{2i} are zero, then the surface is a conic section, such as an ellipse ($\kappa > 0$, or $-1 < \kappa < 0$), parabola ($\kappa = -1$), and hyperbola ($\kappa < -1$). For a general aspheric surface, κ is not equal to zero and some polynomial terms are needed to obtain a fit with least squared errors less than 10^{-15}. If a surface shape is close to a conic, then only a few deformation terms will be needed to obtain a very accurate fit of Eq. (46) to the numerical data representing the curved surfaces of the laser beam profile shaping optical system. The surface curvature is determined from the vertex surface data. However, κ and A_{2i} can have many different solutions which are dependent upon each other. In other words, there is not a unique solution for this nonlinear fitting process. If κ is changed, then A_{2i} will change in order for Eq. (46) to represent the same surface. From a numerical point of view, the solution with the smallest least squared fitting error represents the best solution. However, from a fabrication point of view, it is desirable to use a small number of aspherical terms. Therefore, it is desirable to identify simple surface representations which will be easier to fabricate. More information about this fitting process will be presented in Sec. IV.A.

When the surface equations, d-spacing and indices of refraction are known for all elements of a laser beam profile shaping optical system, as shown in Fig. 2, conventional optical design software packages (Code V, Oslo, or Zemax) can be used to ray trace and analyze its performance. Typically, these standard optical design and analysis software packages compute aberrations-plots, MTF-plots, spot diagrams, and optical path difference (OPD) of systems. The variation of the OPD over the aperture describes the variation in shape of the wavefront as it propagates through the lens. Unfortunately, standard optical design software packages do not provide a direct way to compute the irradiance over the exit pupil of a laser beam profile shaping system. The flux flow equation (36) has been used to compute the irradiance along a ray as it propagates through the optical system (12,13). Alternatively, radiometric calculations for optical systems can be performed by using "unit sphere" methods (35) or by writing a macro for an optical design software package to compute the irradiance over an output surface of the laser shaping system (44). For a laser shaping system with parameters given in Table 1, typical plots of the irradiance over the input and output surfaces of laser beam profile shaping systems are illustrated in Figs 3 and 4 respectively. Specific results for computing the irradiance over an output surface of a laser shaping system will be presented and discussed in Sec. IIV.A. In Sec. IV.C, the geometrical methods for designing laser profile shaping systems, which have been discussed in Sec.

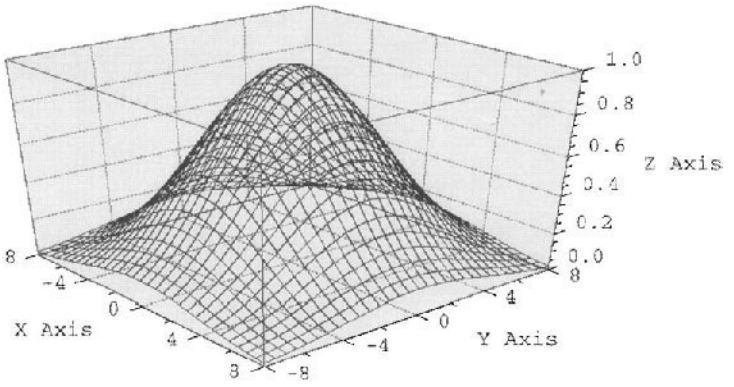


Figure 3 Irradiance profile on input laser beam. (From Ref. 24)

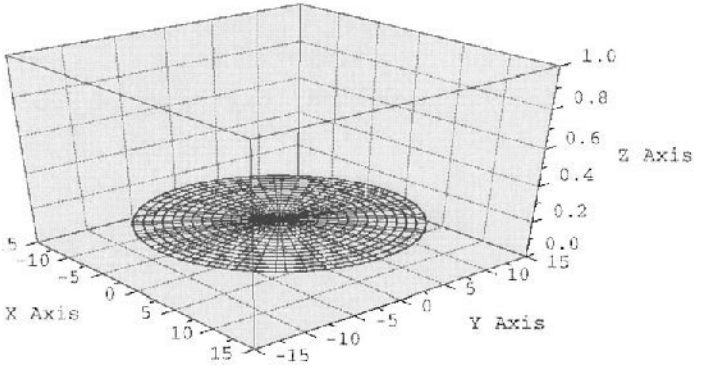


Figure 4 Irradiance profile of output beam. (From Ref. 24)

IV.B, will be related to a seven-step recipe for designing two-element optical systems which perform non-projective transformations as discussed by Cornwell (46,47).

III. NON-PROJECTIVE TRANSFORMATIONS

For a projective transformation (45) in optics, a point in image space can be expressed as a linear function of the coordinates of the object point. Perfect imaging systems, such as Maxwell's "fish-eye" lens or stigmatic imaging of

surfaces, are examples of projective transformations in optics. In practice, aberrations are present in many optical systems, and point-to-point imaging is not possible, except to the first-order or paraxial approximation. Cornwell (46,47) notes that all real optical systems perform non-projective transformations to some extent. That is, there is a non-linear dependence between input (or object) and output (or image) coordinates.

The redistribution of rays leading to the beam shaping illustrated in Fig. 4 requires a non-linear relationship between the input and output aperture coordinates as given by Eqs. (20) and (38). Therefore, the geometrical methods of Sec. II for designing a laser profile shaping system are an example of a non-projective transformation in optics. Cornwell notes that the first element of a laser beam profile shaping (non-projective transform) system creates sufficient aberrations in the wavefront to restructure the intensity of the beam after propagation of the wavefront over a specified distance. Then, the second element of a laser beam profile shaping system has suitable contour to restore the original wavefront shape of the beam. If the purpose of a laser beam profile shaping system is to uniformly illuminate a surface, then the second element is not needed. Symbolically, a laser beam profile shaping system may be considered to be a "black box" which transforms an input laser beam (plane wave) with a Gaussian intensity distribution into an output beam (plane wave) with uniform intensity distribution. The input and output beams have radii r and R, respectively, as shown in Fig. 2. Reference 46 presents extensive discussion of many types of laser beam profile shaping systems and draws some interesting general conclusions.

In particular, Cornwell provides a seven-step recipe for designing two-element systems, which perform non-projective transformations, such as, laser profile shaping systems. Since the contents of Ref. 46 are not widely available in the optics literature to the knowledge of this author, these seven steps are summarized below:

1. Write out differential power expressions for the intensity distributions over the input and output planes.

   Rectangular coordinates

   $$I_{\text{in}}(x, y)\, dx\, dy = I_{\text{out}}(X, Y)\, dX\, dY \tag{47}$$

   Polar coordinates

   $$I_{\text{in}}(r) r\, dr = I_{\text{out}}(R) R\, dR \tag{48}$$

2. Use the conservation of energy to relate the input and output beam parameters.

Rectangular coordinates

$$\int_{\text{Input Aperture}} I_{\text{in}}(x, y)\, dx\, dy = \int_{\text{Output Aperture}} I_{\text{out}}(X, Y)\, dX\, dY \tag{49}$$

Polar coordinates

$$\int_{\text{Input Aperture}} I_{\text{in}}(r) r\, dr = \int_{\text{Output Aperture}} I_{\text{out}}(R) R\, dR \tag{50}$$

3. Determine the magnification relating the input and output ray heights.

Rectangular coordinates
Assume the intensity functions are separable:

$$I_{\text{in}}(x, y) = a_x(x) a_y(y), \tag{51}$$

$$I_{\text{out}}(X, Y) = A_X(X) A_Y(Y). \tag{52}$$

Allowing for non-uniform shaping of a laser beam profile in two orthogonal directions, $X = m_x(x)x$ and $Y = m_y(y)y$, the rectangular magnifications follow from combining Eqs. (47), (51), and (52):

$$m_x(x) = \frac{1}{x}\left[C_1 \int_0^x \frac{a_x(u)\, du}{A_X(u m_x(u))} + C_2\right] \tag{53}$$

$$m_y(y) = \frac{1}{y}\left[\frac{1}{C_1} \int_0^y \frac{a_y(v)\, dv}{A_Y(v m_y(v))} + C_3\right] \tag{54}$$

where C_i are constants determined by boundary conditions, such as the magnification for a rim ray, as done in Eq. (20).

Polar coordinates

$$R = m(r) r \tag{55}$$

$$m(r) = \frac{1}{r}\left[2 \int_0^r \frac{I_{\text{in}}(r) r\, dr}{I_{\text{out}}(m(r) r)} + C\right] \tag{56}$$

where C is a constant determined from the boundary conditions.

4. Express the optical path length (OPL) between input and output reference surfaces of an arbitrary ray in terms of the OPL of a reference ray, such as Eq. (32).
5. Determine the sag $z(r)$ of the first element.
6. Determine the inverse magnification relating the ray coordinates at the first and second elements.
7. Determine the sag $Z(R)$ of the second element.

IV. REFRACTIVE INTENSITY PROFILE SHAPING SYSTEMS

Laser beam profile shaping optics are well suited for applications whose overall efficiency increases when the irradiance over the detector (or substrate) is uniform, such as in compact holographic projector systems (48–50). These compact holographic projection systems have been reported to offer a practical way to make a highly corrected mesh or grid pattern over curved surfaces where the pattern can range in size from sub-micron to multi-micron. The laser profile shaping optics within a holographic projection system enables uniform features to be written over substrates of several centimeters in diameter (25).

To understand this increase in system efficiency when using laser profile shaping optics, note that for a Gaussian beam with irradiance $\sigma(r)$ given by Eq. (18), the intensity of the beam decreases to $1/e^2 \cong 13.5\%$ of its axial value at the beam radius. The effect of this variation in beam intensity over a Gaussian beam is illustrated in Fig. 5: (A) shows significant variation in pattern densities at the center and edge of beam for the same substrate (film) and exposure time when laser profile shaping optics is not part of the system, and (B) shows almost uniform pattern densities at the center and edge of beam when laser profile shaping optics is part of the system.

Therefore, when beam shaping optics are introduced into a holographic projection processing system (25) as illustrated in Fig. 6, the detector substrate will be uniformly illuminated, and photochemical reactions take place at the same rate over the entire substrate area, thus enabling the full beam diameter to be available for material processing. Introducing laser shaping optics into holographic projection processing systems have led to a significant increase in quality of micro-optics fabricated over the substrate.

In Sec. VI.A, a detailed discussion of the design, analysis, fabrication, and testing of a two-lens beam shaping system will be presented. Section IV.B will discuss using axial gradient-index materials to design a laser beam profile shaping optical system, and Sec. V will discuss reflective laser beam profile shaping systems.

A. Two Plano-convex Lens Systems

Section II.C describes how to use the geometrical methods of this chapter to determine the shape* of the curved surfaces of the plano-convex lens system

* The shape of a refracting surface is defined by arrays of data (r, z) and (R, Z) defining the sag and radial heights of surfaces s and S.

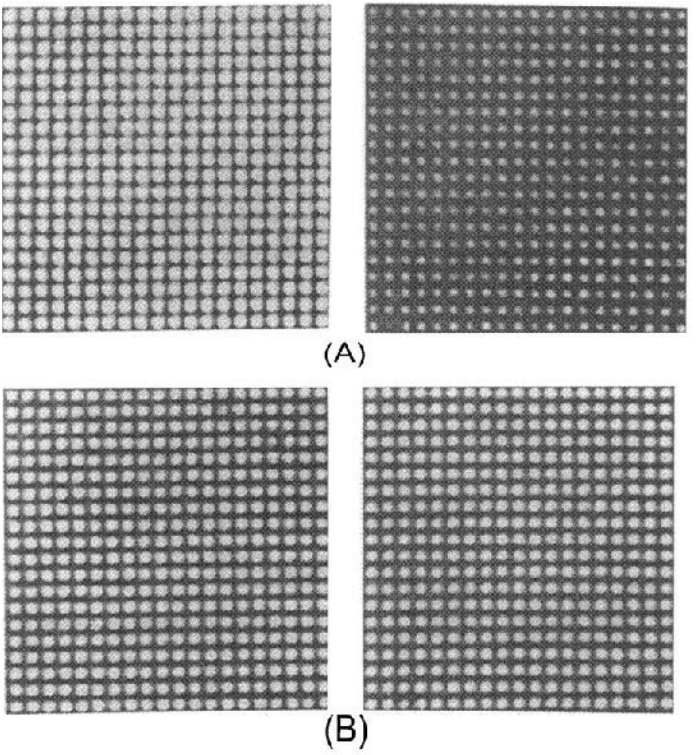


Figure 5 Interference patterns produced by a four-beam holographic projection processing system when illuminated with a Gaussian beam. The image on the left side of the figure was taken near the center of the beam, and the image on the right side of the figure was taken near the edge (waist) of the beam. The images (A) were taken when the laser beam profile optics was not part of the projection system, and the images (B) were taken when laser beam profile shaping optics was part of the projection system. (From Ref. 24)

shown in Fig. 2. Table 1 summarizes the design parameters of a prototype laser profile shaping system, which has been designed, fabricated and tested (23). The lens thickness is not involved in the design differential equations, but is an important factor to be considered since it is related to the energy absorption by the optics. The lens thickness of each element has been assumed to be 10 mm for this two-lens system. Another system parameter, the distance between the two lenses is important for optical design, fabrication, and testing of the system. The larger the spacing between the two elements, the smaller the surface curvature of each element required to

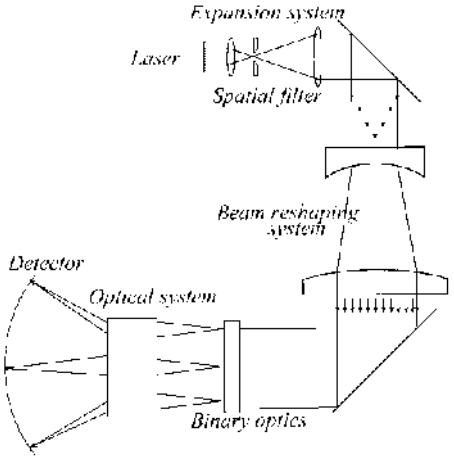


Figure 6 Holographic projection system. (From Ref. 25)

satisfy the design conditions. However, if the element spacing is too large, the system will be difficult to assemble and test. The distance between these elements has been arbitrarily chosen for this system to be five times the diameter of the lens elements.

The index of refraction of the lenses affects the shape of refracting surfaces s and S. For this application the lens elements must have a high transmission for the wavelength of the laser being used [441.57 nm for a helium cadmium (HeCd) laser] with this profile shaping system. There are several materials (51) with good transmission properties at 441.57 nm: acrylic (PMMA) plastic, calcium fluoride (CaF_2), crown glass (BK7), lithium fluoride (LiF), fused quartz, and fused silica. Acrylic plastic is difficult to polish. Fused quartz, Fused silica, and crown glass are difficult to machine with a single-point diamond lathe as required in making these aspheric surfaces. Lithium fluoride contains defects and is difficult to work with. Calcium fluoride has good transmission properties ($\sim 95\%$) within the range of 150–9000 nm, can be machined with a single-point diamond lathe, and is also the least expensive of this group of materials. Therefore, CaF_2 has been selected as the lens material for this laser beam profile shaping system.

The index of refraction of CaF_2 is a function (52) of the wavelength of the light being used:

$$n^2 = 1 + \sum_{i=1}^{3} \frac{A_i \lambda^2}{\lambda^2 - \lambda_i^2} \tag{57}$$

where λ is measured in units of microns (μm) and

$$\lambda_1 = 0.050263605\,\mu\text{m}; \quad \lambda_2 = 0.1003909\,\mu\text{m}; \quad \lambda_3 = 34.649040\,\mu\text{m};$$

$$A_1 = 0.5675888; \quad A_2 = 0.4710914; \quad A_3 = 3.848723.$$

Using Eq. (57), the index of refraction of CaF_2 has been calculated to be 1.43916 at the HeCd laser wavelength of 441.57 nm.

For this laser profile shaping system, the laser beam is expanded from 16 mm to 25 mm in diameter while the beam intensity profile is flattened. After solving the differential equations, the nonlinear least squares fitting process of Sec. II.D was used to fit the lens surface data to the conventional optics surface Eq. (46). The data for surfaces s and S obtained from solving the differential equations has been fit to many different expressions for the optics surface equation. More aspheric terms will generally give a better fit with smaller least squared errors. However, it is easier to fabricate and test surfaces with a non-zero conic term and a small number of polynomial deformation terms. The data in Table 2 with non-zero conic constant plus five deformation terms represents a compromise between fitting accuracy, optical modeling performance, and ease of fabrication.

1. Analysis of Lens Performance

Using the analytical representation of the primary (s) and secondary (S) lens surfaces from Table 2 for the laser profile shaping system shown in Fig. 2, the optical performance of this system has been modeled and compared to the design conditions of the laser profile shaping system described in Sec. II.

Table 2 Surface parameters of an HeCd (441.57 nm) laser profile shaping system where the distance between the primary and secondary lens is 150 mm

Lens surface parameters	Primary	Secondary
Diameter (mm)	30.0	30.0
Vertex radius (mm)	47.861445	113.64905
Index of refraction	1.43916 (CaF_2)	1.43916 (CaF_2)
Thickness (mm)	10.0	10.0
Conic constant, κ	-1.1143607	-1.4877144
A_4 (mm^{-3})	$-7.1532887 \times 10^{-5}$	$-2.6322455 \times 10^{-6}$
A_6 (mm^{-5})	3.3729843×10^{-7}	9.4058758×10^{-9}
A_8 (mm^{-7})	$-1.4916816 \times 10^{-9}$	$-2.3096843 \times 10^{-10}$
A_{10} (mm^{-9}	$5.9836543 \times 10^{-12}$	$1.5839557 \times 10^{-12}$
A_{12} (mm^{-11})	$-1.5166511 \times 10^{-14}$	$-4.8438745 \times 10^{-15}$

Applying the flux flow equation (36) to this lens system, the intensity of a cross-section of the output beam is shown in Fig. 7. A careful computation of the area under the surfaces generated by rotating around the optical axis the input and output intensity profiles given in Fig. 7 shows that the total power is conserved for this laser beam profile shaping system.

The optical path difference (OPD) of the output beam for the laser profile shaping system defined in Table 2 has also been evaluated over the aperture. The maximum OPD for this system is 0.0017λ, which corresponds to the absolute OPD of 0.75 nm for the HeCd laser. This demonstrates that the shape of the output wavefront has the same shape as the input wavefront, as required by the constant OPL condition discussed in Sec. II.B.

It is interesting to compare the output intensity profiles of the designed and fitted system more carefully. Using the flux flow equation, the output intensity profile of the designed and fitted systems are shown Fig. 8.

The label "designed system" means that the numerical data for the refracting surfaces, which was obtained from solving the differential equations, was used directly to compute via the flux flow equation this intensity profile. The label "fitted system" means that the system described in Table 2 was used to compute the output intensity profile. It is clear from Fig. 8 that the intensity profile of the fitted system oscillates at the edge of the aperture, which is a common problem associated with fitting polynomial functions near the edge of a data set. This fitting error is less than 0.2%, which can

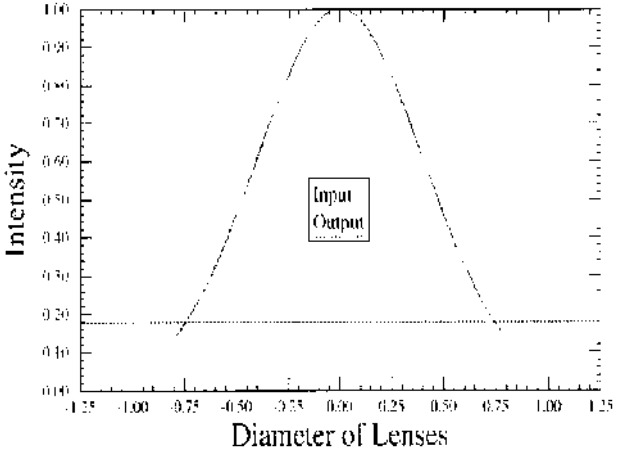


Figure 7 Intensity profile of a meridional cross-section of the input Gaussian beam and the output (uniform) beam for the laser profile shaping system given in Table 2. (From Ref. 23)

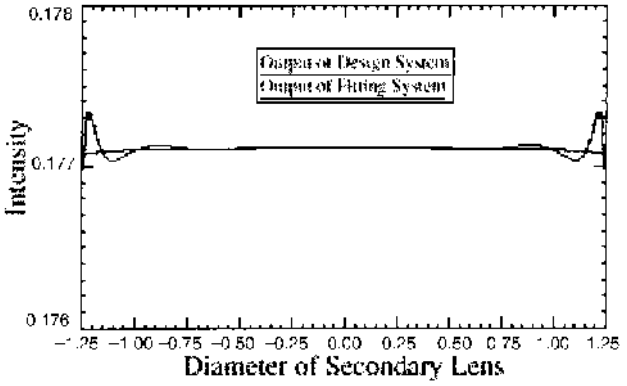


Figure 8 Comparison of the output intensity profile for the designed and fitted laser profile shaping system defined in Table 1. (From Ref. 23)

explain the error in OPD of the output beam. This edge fitting problem can be reduced by designing initially a system with a larger input aperture than will be used in practice.

It is also interesting to consider how this laser beam profile shaping system would perform if it were used at a different laser wavelength, λ'. Since the index of refraction is a function of wavelength as given by Eq. (57) and since these optics have been designed for a specific $n(\lambda)$, it is important to determine whether a set of laser beam profile shaping optics can be used for multiple λ. For example, if the optics defined in Table 2 were used with a Helium Neon (HeNe) laser with wavelength 632.8 nm, the index of refraction of the lenses would be reduced to 1.43289. Then, according to Snell's law, the light leaving the primary lens would not be refracted as strongly as the HeCd light. The separation between the two lenses would need to be increased to accommodate this decrease in index of these lenses to insure that the light is incident upon the secondary at the appropriate height so that it will be refracted parallel to optical axis as illustrated in Fig. 9.

Now, we will derive an expression for the lens separation for a particular ray so that this ray will be incident upon surface S at a suitable point such that after refraction, this ray will leave the secondary lens at the appropriate height R to insure conservation of energy and to be parallel to the optical axis. The new distance d can be calculated from Eq. (29), since r, R, and z' are known. Equation (29) can be written as a quadratic equation of $(Z - z)$

$$(1 - n^2)z'^2(Z - z)^2 + 2(R - r)z'(Z - z) + (R - r)^2(1 - n^2z'^2) = 0 \quad (58)$$

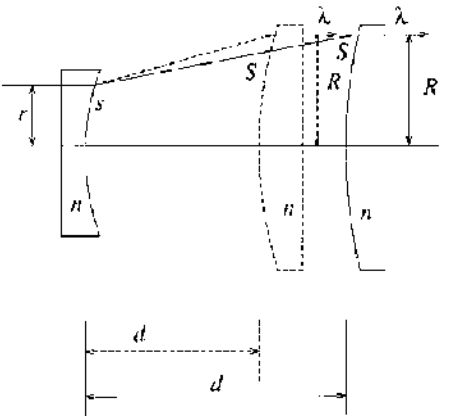


Figure 9 Illustrates the relationship between the wavelength and the lens spacing for a laser profile shaping system. (From Ref. 24)

which has the physical solution

$$Z - z = \frac{(R - r)[1 + n\sqrt{1 + (1 - n^2)z'2}]}{(n^2 - 1)z'}. \tag{59}$$

Once $(Z - z)$ is known, then the corresponding value of d can be determined by solving Eq. (34) as a quadratic equation for d:

$$(n^2 - 1)(n - 1)^2 d^2 - 2n(n - 1)(n^2 - 1)(Z - z)d + (n^2 - 1)^2(Z - z)^2 - (n^2 - 1)(R - r)^2 = 0 \tag{60}$$

with the physical solution

$$d = \frac{n(Z - z) - \sqrt{(Z - z)^2 + (R - r)^2}}{n - 1}. \tag{61}$$

If the lenses are separated by the vertex spacing d of Eq. (61), then the ray leaves the secondary lens parallel to the optical axis. However, every ray in the entrance pupil will require a different lens spacing d. Calculations have shown that the optical system defined in Table 1 will need a lens spacing of approximately $d = 152.18$ mm when used to shape the 632.8 nm HeNe laser beam profile. The maximum difference between the lens spacing for all rays is less than 10 μm, which is smaller than alignment errors. Therefore, it seems reasonable to use this laser beam profile shaping system at multiple wavelengths.

2. Experimental Results

The two plano-aspherical lenses of Table 2 were fabricated by Janos Technology, Inc. of Townshend, Vermont, using a single point diamond lathe; CaF_2 was used as the lens material. A scanning video system shown in Fig. 10 was used to measure the input and output beam profiles.

A Panasonic TV camera (Model WV-1800) was used to sample the laser beam before and after passing through the laser beam profile shaping optics. The camera was mounted on a translation stage behind a pinhole. By scanning across the beam, it was possible to use the same region of the detector for measuring the intensity of all parts of input and output beam. The image processing software – NIH Image v. 1.44 (53) – was used to acquire, display, edit, enhance, analyze, and print images. Reference 24 provides a full discussion of this testing procedure as well as tolerance analysis and other results not summarized in this section. The input and output beam profile using the HeCd laser are shown in Fig. 11.

The open diamond symbols are the measured intensity of input beam, and the solid curve is a Gaussian profile fitted to the input beam data. The solid diamond symbols are the measured intensity of the output beam, and the horizontal solid line is the designed output beam intensity. These results clearly show that the input beam has been transformed into a more uniform beam. However, manufacturing and alignment errors can be linked to the variations of the output beam profile from its theoretical value.

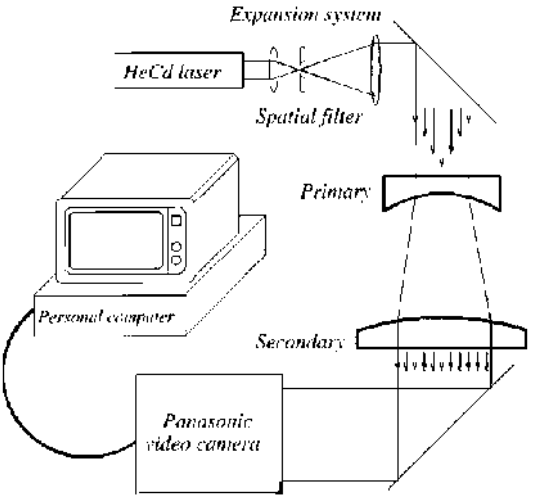


Figure 10 Testing system for laser profile shaping optics. (From Ref. 24)

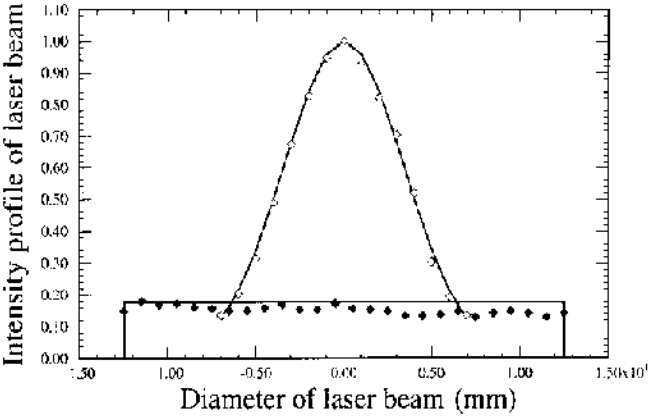


Figure 11 The input and output intensity profile of an HeCd laser shaping system. The open diamond symbols are measured data points of the input beam, which was fit to a Gaussian curve shown as a solid curve. The solid diamond symbols are measured data points of the output beam. (From Refs. 23 and 24)

The output wavefront shape of this two plano-aspherical lens laser beam profile shaping system has also been evaluated. Imposing the constant optical path length condition, as described in Sec. II.B, requires the output wavefront to have the same geometrical shaping as the input wavefront, which is assumed to be planar. The collimated radiation of the input laser beam has a very small divergence, which means that the beam diameter does not increase or decrease with propagation within limits of geometrical optics. For propagation of the output wavefront of this laser profile shaping system around the laboratory, say 10 m, the output beam diameter remained constant. To quantify this result, Ronchigrams (54) of the input and output beams were recorded using the video system described in this section for recording the beam intensity profile (55). Analysis of these Ronchigrams indicates that the output beam wavefront has the same shape as the input wavefront.

In addition, a HeNe laser was used to illuminate these profile shaping optics, and the predictions of Sec. IV.A.1 have also been confirmed. When using a HeNe laser with 632.8 nm radiation, the lens spacing was increased to 152.2 mm according to the predictions of Eq. (61). The input and output beam profiles are shown in Fig. 12.

The open diamond symbols represent the input beam, and the solid diamond symbols represent the output beam. These predictions and experimental results confirm the geometrical optics theory of laser profile shaping systems presented in this chapter.

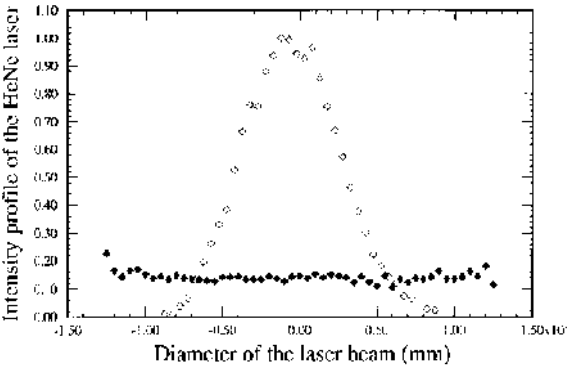


Figure 12 The input and output intensity profile of an HeNe laser use with HeCd laser profile shaping system after increasing the lens spacing to 152.2 mm. The open diamond symbols are measured data points of the input beam. The solid diamond symbols are measured data points of the output beam. (From Ref. 24)

B. Axial GRIN Lens Systems

Gradient-index (GRIN) glasses have been shown to be able to provide additional degrees of freedom for designing optical systems. Sands (56) has shown that the contributions of an axial GRIN to the third-order aberrations of an optical system are equivalent to those of an aspherical surface. This suggests that the aspherical surfaces of the laser beam profile shaping section of Sec. IV.A can be replaced by axial GRIN lenses with spherical surfaces. Wang and Shealy (57) have demonstrated, without taking into explicit account the functional dependence of the index of refraction on the wavelength of light, that it is possible to design axial GRIN laser beam profile shaping systems with realistic materials and spherical surfaces. See Ref. 57 for a more detailed discussion of the results presented in this section.

1. Mathematical Developments

The ray equations for propagation of light through GRIN materials follow from Fermat's principle and Eq. (16). For an axial-GRIN material with the symmetry axis along the z-axis, the ray equations can be written in the following form (58):

$$x = x_0 + K_0 \int_{z_0}^{z} \frac{dz}{M}, \tag{62}$$

$$y = y_0 + L_0 \int_{z_0}^{z} \frac{dz}{M}, \tag{63}$$

with

$$M = [n^2(z) - K^2 - L^2]^{1/2} = [n^2(z) - K_0^2 - L_0^2]^{1/2}$$

where K, L, M are the three optical direction cosines; K_0, L_0, M_0 are initial values of the three optical direction cosines of ray within the GRIN material; and x_0, y_0, z_0 are the initial coordinates of the ray within the GRIN material.

The geometrical configuration of a two-lens GRIN laser beam profile shaping system is shown in Fig. 13.

The optical axis is also the symmetry axis of both GRIN lenses. The input laser beam will not be deflected by the plano-surface of the first lens, but will diverge from the spherical surface S_1. The GRIN profile of the first lens will cause the rays at different heights to refract in such a way as to convert the input Gaussian intensity profile into a uniform intensity profile at the second lens. The spherical surface S_2 of the second lens will refract the rays so that the output beam will be parallel to the optical axis.

Following the discussion presented in Sec. II.A for input Gaussian laser beam with intensity given by Eq. (18), the energy collected within a circle of radius r_1 is given by

$$E(r_1) = 2\pi \int_0^{r_1} \sigma(r')r'\,dr' = \frac{\pi r_0^2}{2}\,[1 - \exp(-2r_1^2/r_0^2)]. \tag{64}$$

If the beam reaches a uniform intensity Σ within a circle of radius r_2 on the second lens, then applying conservation of energy as expressed by Eq. (19) gives

$$r_2 = \left\{ \frac{r_0^2}{2\Sigma}[1 - \exp(-2r_1^2/r_0^2)] \right\}^{1/2}, \tag{65}$$

where Σ is given by Eq. (21).

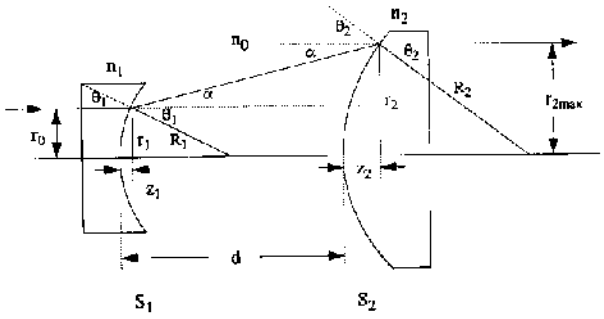


Figure 13 Geometrical configuration of a two-GRIN-lens laser profile shaping system. (From Ref. 26)

As shown in Fig. 13, the surfaces S_1 and S_2 are two spherical surfaces. From the surface equations*, the surface sag z_1 and z_2 can be written as

$$z_1 = \frac{r_1^2/R_1}{1 + \sqrt{1 - r_1^2/R_1^2}}, \tag{66}$$

or

$$r_1^2 = 2z_1R_1 - z_1^2, \tag{67}$$

and

$$z_2 = \frac{r_2^2/R_2}{1 + \sqrt{1 - r_2^2/R_2^2}}, \tag{68}$$

and

$$r_2^2 = 2z_2R_2 - z_2^2. \tag{69}$$

The geometrical relations shown in Fig. 13 justify the following expressions:

$$\tan\alpha = \frac{r_2 - r_1}{d - z_1 + z_2}, \tag{70}$$

$$\cot\theta_1 = \frac{R_1 - z_1}{r_1} = \frac{R_1 - z_1}{\sqrt{2R_1z_1 - z_1^2}}, \tag{71}$$

$$\cot\theta_2 = \frac{R_2 - z_2}{r_2} = \frac{R_2 - z_2}{\sqrt{2R_2z_2 - z_2^2}}. \tag{72}$$

Applying Snell's law at surfaces S_1 and S_2 gives

$$n_1(z_1)\sin\theta_1 = n_0\sin(\theta_1 + \alpha), \tag{73}$$

$$n_2(z_2)\sin\theta_2 = n_0\sin(\theta_2 + \alpha). \tag{74}$$

The GRIN profiles of these lenses can be determined from Eqs. (73) and (74). Applying the sum of two angles trigonometric formula, these equations can be written as

$$n_1(z_1) = n_0(\cot\theta_1\sin\alpha + \cos\alpha), \tag{75}$$

$$n_2(z_2) = n_0(\cot\theta_2\sin\alpha + \cos\alpha). \tag{76}$$

* From Eq. (46), where $\kappa = 0 = A_{2i}, c = 1/R$, and R_1, R_2 are the radii of curvature of the spherical lens surfaces.

To evaluate the GRIN profile of the first lens, the right-hand sign of Eq. (75) needs to be expressed as a function of z_1. Combining Eqs. (65) and (67) gives

$$r_2 = \left(\frac{r_0^2}{2\Sigma} \{1 - \exp[-2(2R_1 z_1 - z_1^2)/r_0^2]\} \right)^{1/2}. \tag{77}$$

Solving Eqs. (77) and (68) for z_2 as a function of z_1 leads to the following:

$$z_2 = \frac{\left(\frac{r_0^2}{2\Sigma} \{1 - \exp[-2((2R_1 z_1 - z_1^2)/r_1^2)]\} \right) \Big/ R_2}{1 + \left[1 - \left(\frac{r_0^2}{2\Sigma} \{1 - \exp[-2((2R_1 z_1 - z_1^2)/r_0^2)]\} \right) \Big/ R_2^2 \right]^{1/2}}. \tag{78}$$

The resulting expression for GRIN function of the first lens is

$$n_1(z_1) = n_0 \left\{ \left[\frac{R_1 - z_1}{(2R_1 z_1 - z_1^2)^{1/2}} \right] \sin \alpha(z_1) + \cos \alpha(z_1) \right\}. \tag{79}$$

In similar fashion, the GRIN function for the second lens can be written in terms of z_2 to give

$$n_2(z_2) = n_0 \left\{ \left[\frac{R_2 - z_2}{(2R_2 z_2 - z_2^2)^{1/2}} \right] \sin \alpha(z_2) + \cos \alpha(z_2) \right\}. \tag{80}$$

Equations (79) and (80) are formulas for the GRIN profiles of the two plano-spherical lenses in a refractive laser profile shaping system. These results are based on geometrical optics, energy conversion along a tube of rays, and the constant optical path length condition. Now, these results will be used to design two GRIN laser beam profile shaper systems.

2. Using GRIN Lenses

The use of GRIN materials in optical systems has been limited by fabrication capabilities of these materials. Considerable progress has been made towards better controlling the GRIN profile while also increasing the change of the index of refraction and the depth of the gradient of the index (59–63) within the material. Until the recent development of the GRADIUM^{TM} GSF glass family by LightPath Technologies (63), it has been difficult to obtain an overall index change larger than 0.08 and a depth of the gradient greater than 5 mm. It is possible to obtain linear and near-parabolic GRIN profiles. For the designs presented in this section (57), these constraints have been used – maximum overall index of refraction change is 0.08 and the depth of the GRIN gradient is 5 mm. As GRIN technology improves, optical designs for laser profile shaping systems will be able to use a broader spectrum of materials to reduce size and cost of

laser beam profile shaping systems. For example, the GRADIUM^{TM} GSF glass family from LightPath Technologies (64) has an overall index of refraction change ranging from 0.04 to 0.14 for a thickness of the gradient ranging from 6 to 26 mm. With this expanded range of parameters, more versatile GRIN laser beam profile shaping systems can be developed.

Using Eqs. (79) and (80) the GRIN profiles of a two-lens laser beam profile shaping system shown in Fig. 13 can be computed for any set of layout parameters including the lens spacing d and the spherical lens surface vertex radii R_1 and R_2. However, only GRIN materials with realistic gradient-index profiles can be used when building a laser beam profile shaping optical system. For a particular case, when the input and output beam radii, r_0 and $r_{2\max}$, are given, the depth of the gradient Δz is completely determined by the vertex radius of a spherical lens. The relationship between the depth of the index gradient and the vertex radius is illustrated in Fig. 14.

In summary, smaller the vertex radii yield larger depth of the index gradient. From the equation of a spherical surface S_1, Eq. (67), it follows that if the depth of the index gradient Δz_1 has a specific value, then the lower limit on choosing R_1 will be given by

$$R_1 \geq \frac{r_0^2 + (\Delta z_1)^2}{2(\Delta z_1)}. \tag{81}$$

It is interesting to note that the refraction of rays at different heights from the optical axis are determined by the vertex radius and the gradient-index distribution across the beam diameter. The more planar a spherical surface is (larger vertex radius), then the greater the gradient of the index across the

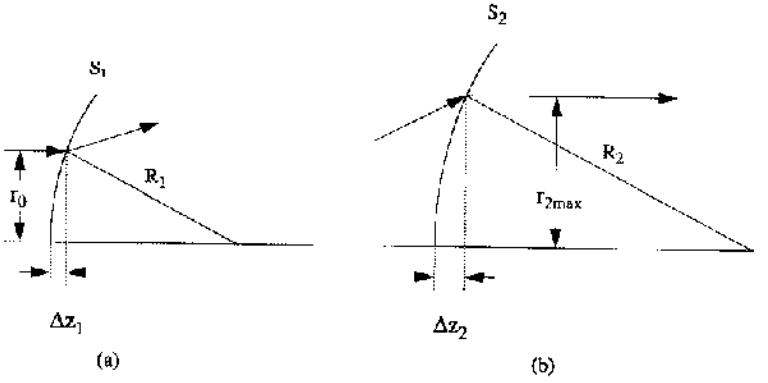


Figure 14 Relations between vertex radii and depth of the gradient-index material. (From Ref. 26)

surfaces S_1 and S_2 will be for achieving the same deflection of the rays. That is, the overall index change Δn, which can be fabricated, sets the upper limit for choosing the vertex radii R_1 and R_2.

The spacing d between the two lenses also affects the gradient-index profile. Shorter systems require stronger index gradients to achieve the same redistribution of the laser beam profile. Selection of the layout parameters R_1, R_2, and d need to be guided by the current gradient-index fabrication technology. The relationship between GRIN characteristics (Δn and Δz_1) and system layout parameters (R_1 and d) of the primary lens for the GRIN laser beam profile shaping system are illustrated in Fig. 15.

A similar plot for the secondary lens can be constructed. Then, for a given beam waist r_0 and manufacturing specifications for the GRIN material (Δn and Δz_1), the system layout parameters (R_1, R_2, and d) can be determined. Two specific laser profile shaping optical systems using GRIN lenses will be discussed in more details in this section. The first system transforms the input Gaussian beam profile into a uniform output beam profile of the same diameter as the input beam. The second system expands the input beam by a factor of two while also transforming the input Gaussian beam to a uniform output beam profile.

Consider a laser profile shaping system with layout parameters given in Table 3 and illustrated in Fig. 16.

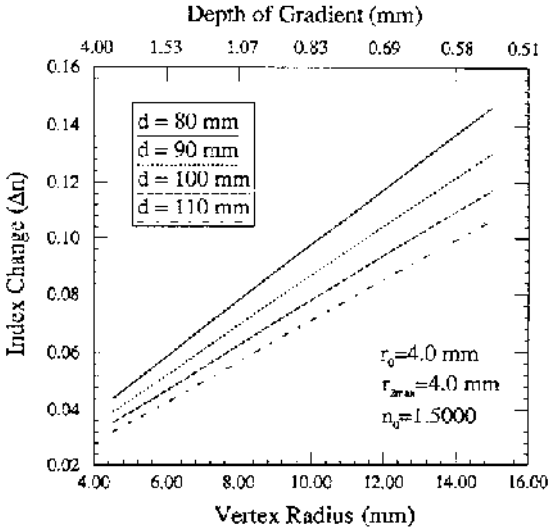


Figure 15 Relationship between GRIN characteristics and laser profile shaping optical system design parameters. (From Ref. 26)

Table 3 Layout parameter for a laser profile shaping system with no expansion of the diameter of the input beam

System variables	Values
Primary lens vertex radius, R_1	5 mm
Secondary lens vertex radius, R_2	5 mm
Spacing between lenses, d	100 mm
Incident beam waist (radius), r_0	114.0 mm
Exiting beam radius, $r_{2\,max}$	4.0 mm
Index of connector, n_0	1.5

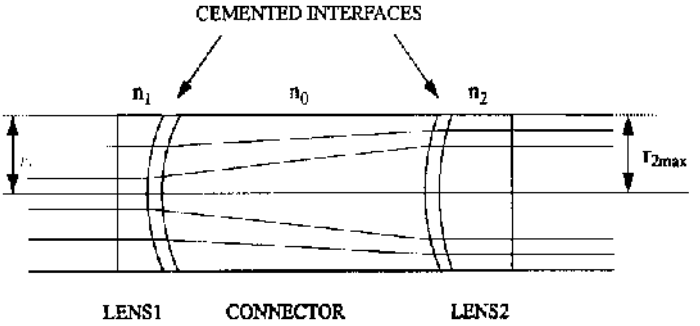


Figure 16 Layout of a GRIN laser profile shaping system. (From Ref. 26)

For this system, the input and output beam have the same diameter, which means that the marginal rays (displaced a distance r_0 from the optical axis) must not be deflected by the optical components. According to Snell's law, a ray will not be deflected by a surface when the index of refraction is the same on both sides of the interface. This means that there must be a dense material of index n_0 connecting the two lenses. The connector in this design is a glass bar with the same index of refraction as that of the base glasses used to fabricate the GRIN lenses. Using the layout parameters given in Table 3, the GRIN profiles of the primary and secondary lenses have been computed from Eqs. (79) and (80) as a function of the radial distance from the optical axis. These results are shown in Figs. 17 and Fig. 18 for the primary and secondary lens materials, respectively.

Fitting with a least-squares technique the GRIN profiles as a function of the sag z of each surface gives the following empirical expressions for $n_1(z_1)$ and $n_2(z_2)$:

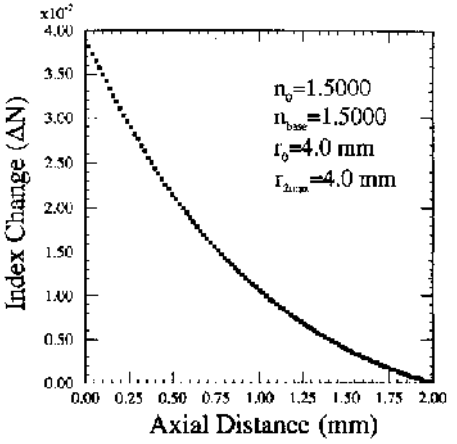


Figure 17 GRIN profile of the primary lens. (From Ref. 26)

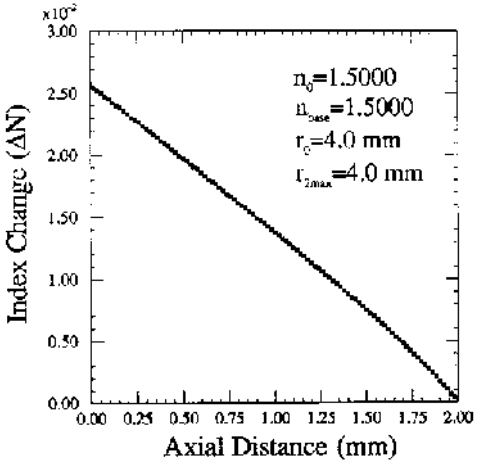


Figure 18 GRIN profile of the secondary lens. (From Ref. 26)

$$n_1 = 1.537910 - 0.036171z_1 + 0.008827z_1^2, \tag{82}$$

$$n_2 = 1.525456 - 0.010882z_2 - 0.000801z_2^2. \tag{83}$$

The self-consistency of this design of a GRIN laser beam profile shaping system has been checked by doing a ray trace to compute the intensity of the output beam. A grid of Gaussian distribution over the entrance pupil was

used so that the number of rays per unit area represents the intensity of the beam as it passes through this optical system. The intensity distribution of the input and output beams are shown in Fig. 19.

It is clear from these results that the input Gaussian beam has been transformed into a uniform intensity output beam.

Now consider a 2× laser beam profile shaping system with layout parameters given in Table 4 and illustrated in Fig. 20.

When compared to a non-expanding laser profile shaping system, beam expanders deflect rays to a greater extent. Therefore, it is important to choose carefully the system layout parameters (R_1, R_2, d) so that the resulting GRIN profile can be fabricated. Unless a dense medium connects

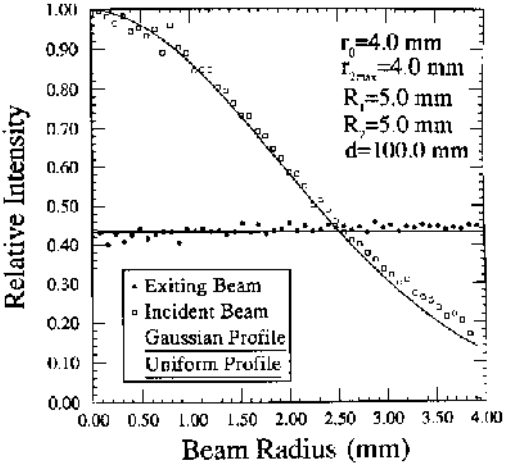


Figure 19 Computed intensity distribution for the input and output beams. (From Ref. 26)

Table 4 Layout parameter for a 2× laser profile shaping system

System variables	Values
Primary lens vertex radius, R_1	5 mm
Secondary lens vertex radius, R_2	10 mm
Spacing between lenses, d	150 mm
Incident beam waist (radius), r_0	4.0 mm
Exiting beam radius, $r_{2\,\max}$	8.0 mm
Index of connector, n_0	1.5

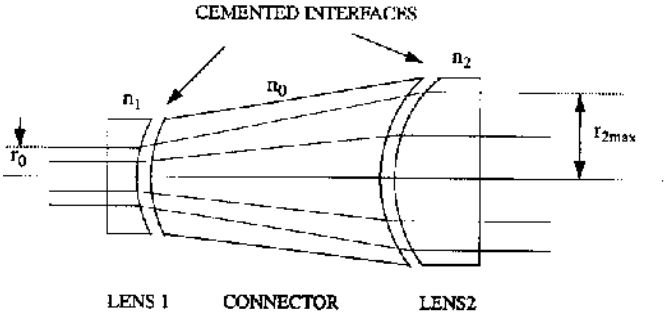


Figure 20 Layout of a GRIN laser expander and profile shaping system. (From Ref. 26)

the two lenses, the overall index change will be too large for current material fabrication technologies. Using the layout parameters given in Table 4, the GRIN profiles of the primary and secondary lenses have been computed from Eqs. (79) and (80) as a function of the radial distance from the optical axis. These results are shown in Fig. 21 and Fig. 22 for the primary and secondary lens materials, respectively.

Fitting with a least-squares technique the GRIN profiles as a function of the sag z of each surface gives the following empirical expressions for $n_1(z_1)$ and $n_2(z_2)$:

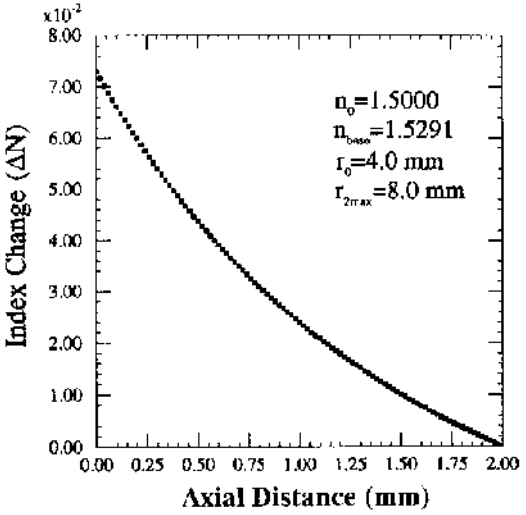


Figure 21 GRIN profile of the primary lens. (From Ref. 26)

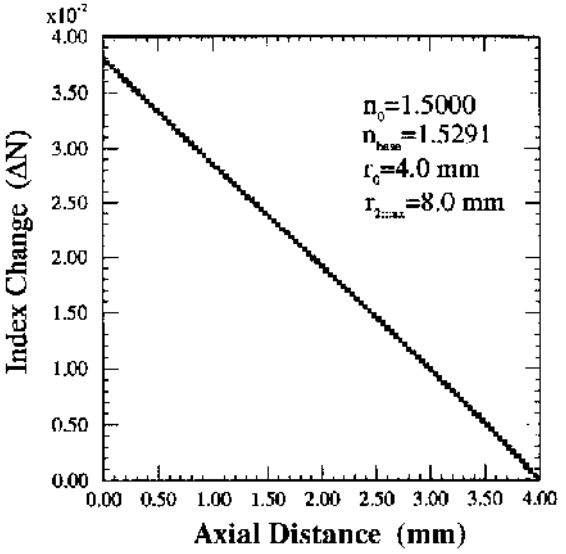


Figure 22 GRIN profile of the secondary lens. (From Ref. 26)

$$n_1(z_1) = 1.600350 - 0.059840z_1 + 0.012423z_1^2, \tag{84}$$

$$n_2(z_2) = 1.567088 - 0.009410z_2. \tag{85}$$

Similarly, the self-consistency of this design of a GRIN laser beam profile shaping system has been checked by doing a ray trace to compute the intensity of the output beam. A grid of Gaussian distribution over the entrance pupil was used so that the number of rays per unit area represents the intensity of the beam as it passes through this optical system. The intensity distribution of the input and output beams are shown in Fig. 23. It is evident that the Gaussian input beam has been transformed into a uniform output beam.

3. Summary of GRIN Profile Shaping System

The theory and design procedures for using axial GRIN plano-convex lenses with spherical surfaces has been presented. This is in contrast to the aspherical surfaces required for constant index materials, as presented in Sec. IV.A. Two axial GRIN laser beam profile shaping system – 2× expander and non-expander – have been designed, analyzed, and shown via simulations to transform a Gaussian input beam into a uniform intensity profile output beam. Linear least-squares fitting techniques have shown that

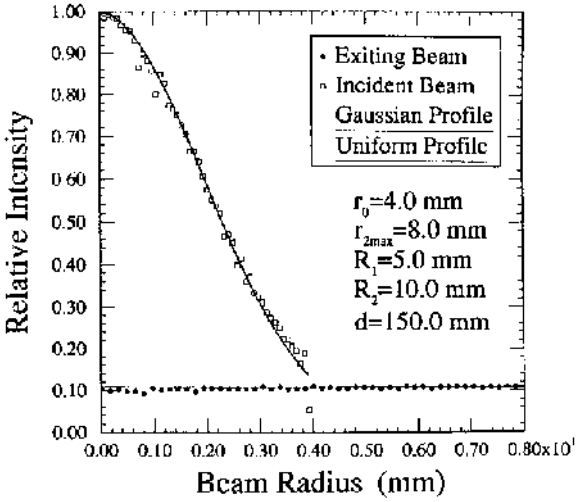


Figure 23 Computed intensity distribution for the input and output laser beams. (From Ref. 26)

the required GRIN materials to be linear or near-parabolic functions of the axial distance where the depth of the gradient and the overall index change are within current GRIN fabrication techniques. As GRIN fabrication technologies improve as illustrated by the LightPath Technologies (63,64) development of the GRADIUM^{TM} GSF glass family, there will be new opportunities for building compact and versatile laser profile shaping systems.

V. REFLECTIVE INTENSITY PROFILE SHAPING SYSTEMS

The geometrical methods of Secs. II and III can also be used to design reflective laser beam profile shaping systems, which are well suited for applications involving high-power lasers or using radiation for which the transmission properties of materials is low. These reflective systems may or may not have a central obscuration. One- and two-mirror systems with central obscuration have been used (6–9) for shaping the irradiance profile of laser beams. McDermit and Horton (6,7) have reported differential equations for one- and two-mirror systems which transform collimated sources of known irradiance profiles into a prescribed spatial irradiation over a known receiver surface. Cornwell (46,47) has developed a general approach using non-projective transformations to design two-mirror laser beam

profile shaping systems with either rectangular or polar symmetry. Malyak (10) has designed a two-mirror unobsured optical system using rotationally symmetric aspheres to convert an input Gaussian beam into a uniform intensity output beam. The geometrical methods used to design one- and two-mirror reflective laser profile shaping systems will be presented in this section.

In Sec. V.A, the differential equation approach of McDermit and Horton for designing a one-mirror system to transform a collimated input beam profile into prescribed illumination of a receiver surface will be summarized. Next, in Sec. V.B, the non-projective transformations of Cornwell (46,47) and the differential equation approach of Malyak (10) have been used to describe the design of two-mirror laser profile shaping systems with either rectangular or polar symmetry.

A. One-mirror Profile Shaping Systems

Consider the geometrical configuration of a one-mirror laser beam profile shaping system shown in Fig. 24.

The input radiation is collimated (parallel to optical axis) with a known intensity profile. The receiving surface is illuminated with a prescribed intensity distribution while the output beam is not collimated. Unit vectors along the input and output beams are given by

$$\mathbf{a} = \hat{\mathbf{k}} \tag{86}$$

$$\mathbf{A} = \mathbf{a} - 2\hat{\mathbf{n}}_1(\mathbf{a} \cdot \hat{\mathbf{n}}_1) = \frac{2z'\hat{\mathbf{r}} - (1 - z'^2)\hat{\mathbf{k}}}{(1 + z'^2)} \tag{87}$$

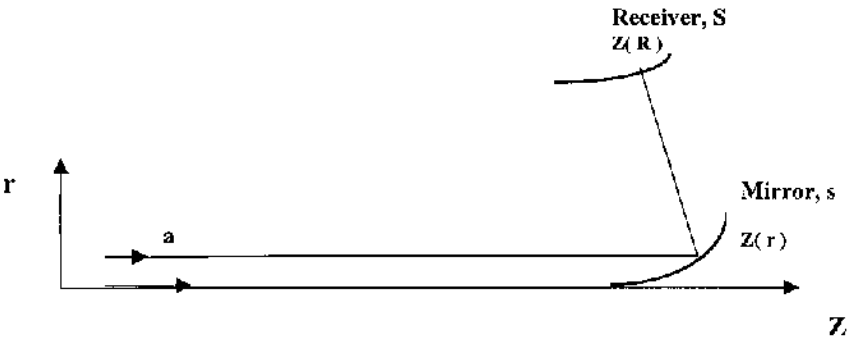


Figure 24 Geometrical configuration of a one-mirror laser profile shaping system.

where Eqs. (14) and (24) were used. The ray trace equation connecting the mirror surface s with the receiving surface S in the r–z plane is given by

$$\frac{R-r}{Z(R)-z(r)}=\frac{A_r}{A_z}=\frac{2z'(r)}{-[1-z'^2(r)]} \tag{88}$$

where $z=z(r)$ represents the unknown equation of the mirror surface s, and $Z=Z(R)$ represents the known equation of the receiving surface S. Equation (88) can be written as

$$-(R-r)z'^2+2(Z-z)z'+(R-r)=0. \tag{89}$$

Applying the differential energy balance Eq. (10) to this problem gives

$$I_1(r)2\pi r\,dr=I_2(R)2\pi R[dR^2+dZ^2]^{1/2}. \tag{90}$$

where $I_1(r)$ is the beam intensity incident upon the first mirror surface, and $I_2(R)$ is the intensity incident upon the second mirror surface. Equation (90) can be rearranged into the form

$$\frac{dZ}{dr}=\frac{I_1(r)}{I_2(R)}\frac{r}{R}\frac{1}{\left[1+\left(\dfrac{dR}{dZ}\right)^2\right]^{1/2}}. \tag{91}$$

Recall that $I_1(r)$, $I_2(R)$, and $Z(R)$ are known functions of their respective variables, and $z(r)$ is an unknown function at this point of the analysis. Also, note that the ray trace Eq. (88) expresses a mapping between surfaces s and S:

$$(r,z)\Rightarrow(R,Z) \tag{92}$$

which implies that R is a function of r, $R(r)$. From the chain rule for differentiation of a function of function, the term (dZ/dr) in Eq. (91) can be written as

$$\frac{dZ}{dr}=\frac{dZ}{dR}\frac{dR}{dr}=Z'(R)\frac{dR}{dr}. \tag{93}$$

where $Z'(R)\equiv dZ(R)/dR$ can be evaluated directly from the equation of the surface S, and dR/dr can be evaluated from the ray trace equation (89). Differentiating Eq. (89) with respect to r gives

$$-2z'z''(R-r)-z'^2\left(\frac{dR}{dr}-1\right)+2z''(Z-z)+2z'\left[\frac{dZ}{dr}-z'\right]+\left(\frac{dR}{dr}-1\right)=0. \tag{94}$$

Combining Eqs. (93) and (94) leads to

$$2z''[-z'(R-r)+(Z-z)]+\frac{dZ}{dr}\left[(1-z'^2)\left(\frac{dR}{dZ}\right)+2z'\right]-(1+z'^2)=0. \tag{95}$$

However, rewriting Eq. (89) gives the relationship

$$(Z-z)=\frac{(R-r)}{2z'}\,[z'^2-1], \tag{96}$$

which has been used to express Eq. (95) in the following form:

$$\frac{z''}{z'}=\frac{1}{(R-r)}\left\{\left(\frac{dZ}{dr}\right)\left[\left(\frac{dR}{dZ}\right)\frac{(1-z'^2)}{(1+z'^2)}+\frac{2z'}{(1+z'^2)}\right]-1\right\}. \tag{97}$$

Replacing the term (dZ/dr) in Eq. (97) with the right-hand-side of Eq. (91) gives the following differential equation for the mirror surface in terms of known functions

$$\frac{z''}{z'}=\frac{1}{(R-r)}\left\{\frac{I_1(r)}{I_2(R)}\left(\frac{r}{R}\right)\frac{\left[\left(\frac{dR}{dZ}\right)\left(\frac{1-z'^2}{1+z'^2}\right)+\frac{2z'}{(1+z'^2)}\right]}{\sqrt{1+\left(\frac{dR}{dZ}\right)^2}}-1\right\}. \tag{98}$$

Equation (98) is equivalent to Eq. 13 of Ref. 6 and Eq. 3.14 of Ref. 7. When appropriate boundary conditions are given, then Eq. (98) can be solved for the shape of the mirror surface which will illuminate the receiver surface S with a prescribed intensity $I_2(R)$ for a given source intensity profile $I_1(r)$. References 6 and 7 develop an extension of this analysis to two-mirror intensity profile shaping systems. A number of specific solutions for both one- and two-mirror systems are given in Refs. 6 and 7 including two laser beam profile shaping systems:

- uniform illumination of a plane perpendicular to the incident beam using a one-mirror system for an input Gaussian beam, Fig. 7.3 of Ref. 7, and
- uniform illumination of a plane on the optical axis with a two-mirror system for an input Gaussian beam, Fig. 7.9 of Ref. 7.

In the next section, the non-projective transformations* of Cornwell (46,47)

* See section 4.3 for definitions of projection and non-projection transformations.

will be presented as a geometrical method for designing two-mirror laser beam profile shaping system for either rectangular or polar symmetry.

B. Two-mirror Profile Shaping Systems

In this section, the design equations of a two-mirror intensity profile shaping systems will be developed. Results in both rectangular and polar coordinate systems will be presented. For more details and applications of these results, the interested reader is encouraged to see Refs 10 and 46. Reference 46 is the original source of the development of non-projective transformations in optics used to develop the material presented in this section. Development of the design equations for rectangular and polar coordinate systems will follow the seven-step recipe of Cornwell (46,47) for non-projective transformation summarized in Sec. 4.3. The geometrical configuration of non-projective transformations is illustrated in Fig. 25.

The non-projective transformation represent a mapping between the input plane and the output beam which takes into account conservation of energy, constant optical path length of wavefront between the input and output planes, and the ray trace equations. The input and output beams will be assumed to be collimated and parallel to the optical axis.

1. Systems with Rectangular Symmetry

The relationship between the element of areas on the input and output planes is illustrated in Fig. 26 and can be written in the following form:

$$I_{\text{in}}(x, y)\, dx\, dy = I_{\text{out}}(X, Y)\, dX\, dY. \tag{99}$$

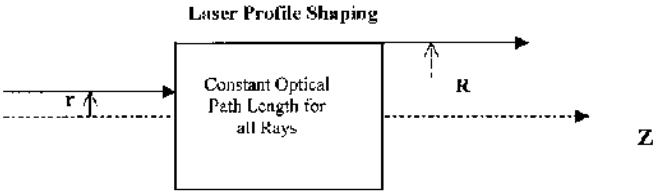


Figure 25 Geometry of the input plane and output plane of a non-projective transformation. Either rectangular or polar coordinate systems can be used, depending on the symmetry of the laser beam profiles and the optical system.

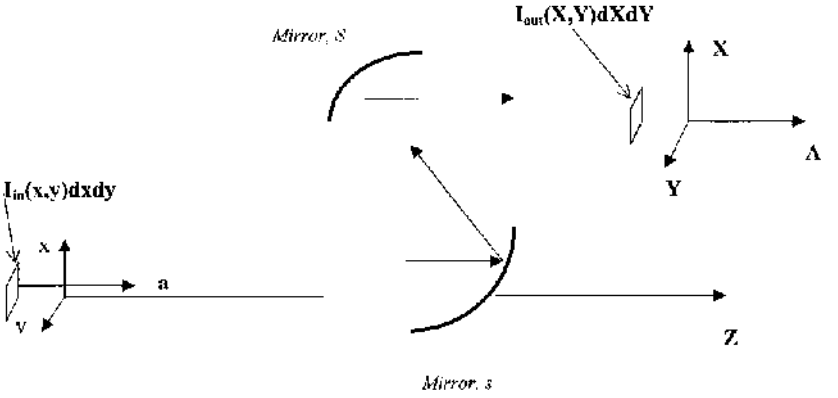


Figure 26 Symbolic relationship between the input and output element of area in rectangular coordinates. Conservation of energy within corresponding element of areas on the input and output planes is one principle used to determine the optical surface shapes of laser profile shaping systems. The constant optical path length of all rays passing from the input to the output plane is the second principle used to determine the optical surface shapes of laser profile shaping system. For two-mirror systems, the axial optical path length is the geometrical path length L.

where I_{in} and I_{out} are the input and output intensity profiles of the laser beam. The total energy must also be conserved which is represented by integrating Eq. (99) over the full aperture of the input and output planes

$$\int_{\text{Full Input Aperture}} I_{\text{in}}(x, y)\, dx\, dy = \int_{\text{Full Output Aperture}} I_{\text{out}}(X, Y)\, dX\, dY. \tag{100}$$

For laser beam profile shaping systems with rectangular symmetry, assume that the input and output intensity profiles can be separated into a product of one-dimensional amplitude functions, as illustrated in Sec. III:

$$I_{\text{in}}(x, y) = a_x(x)a_y(y), \tag{101}$$

$$I_{\text{out}}(X, Y) = A_X(X)A_Y(Y). \tag{102}$$

Allowing for non-uniform shaping of a laser beam profile in two orthogonal directions, assume that there is an independent and non-uniform magnification of the x and y ray coordinates between the input and output planes:

$$X = m_x(x)x, \tag{103}$$

$$Y = m_y(y)y. \tag{104}$$

The rectangular magnifications $m_x(x)$ and $m_y(y)$ can be determined by imposing the incremental expression of conservation of energy, Eq. (99), for the separated intensity functions, Eqs. (101) and (102):

$$a_x(x)a_y(y)\,dx\,dy = A_X(X)A_Y(Y)\left[\frac{\partial(xm_x(x))}{\partial x}\right]dx\left[\frac{\partial(ym_y(y))}{\partial y}\right]dy. \quad (105)$$

Rewriting Eq. (105) with terms depending on x on the left-hand side of the equation leads to separation of variables:

$$\frac{a_x(x)}{A_X(m_x(x))}\,\frac{1}{\left[\dfrac{\partial(xm_x(x))}{\partial x}\right]} = \frac{A_Y(m_y(y))}{a_y(y)}\left[\frac{\partial(ym_y(y))}{\partial y}\right] = \frac{1}{C_1},\ \text{const.}, \quad (106)$$

or,

$$\left[\frac{\partial(xm_x(x))}{\partial x}\right] = C_1\,\frac{a_x(x)}{A_X(m_x(x))}, \quad (107)$$

$$\left[\frac{\partial(ym_y(y))}{\partial y}\right] = \frac{1}{C_1}\,\frac{a_y(y)}{A_Y(m_y(y))}. \quad (108)$$

The constant C_1 is determined from the boundary conditions, such as the magnifications at the edge of the beam. Integrating Eqs. (107) and (108) gives

$$m_x(x) = \frac{1}{x}\left[C_1\int_0^x \frac{a_x(u)\,du}{A_X(um_x(u))} + C_2\right], \quad (109)$$

$$m_y(y) = \frac{1}{y}\left[\frac{1}{C_1}\int_0^y \frac{a_y(v)\,dv}{A_Y(vm_y(v))} + C_3\right] \quad (110)$$

where C_2 and C_3 are constants determined by boundary conditions, such as, the magnification of a rim ray, as done in Eqs. (20). Equations (109) and (110) are integral equations for the x–y ray magnifications. For many applications using laser beam profile shaping systems, the output intensity profile is uniform and solution to either the differential equations (107) and (108) or the integral equations (109) and (110) is straight forward.

Now, the ray trace equations connecting points on the input plane to the output plane will be developed. The geometrical configuration of a two-mirror laser beam profile shaping system is illustrated in Fig. 27.

The unit ray vector $\mathbf{A_{1\to 2}}$ connecting the two mirror surfaces $s[x, y, z(x, y)]$ and $S[X, Y, Z(X, Y)]$ along a ray path can be written in the following form:

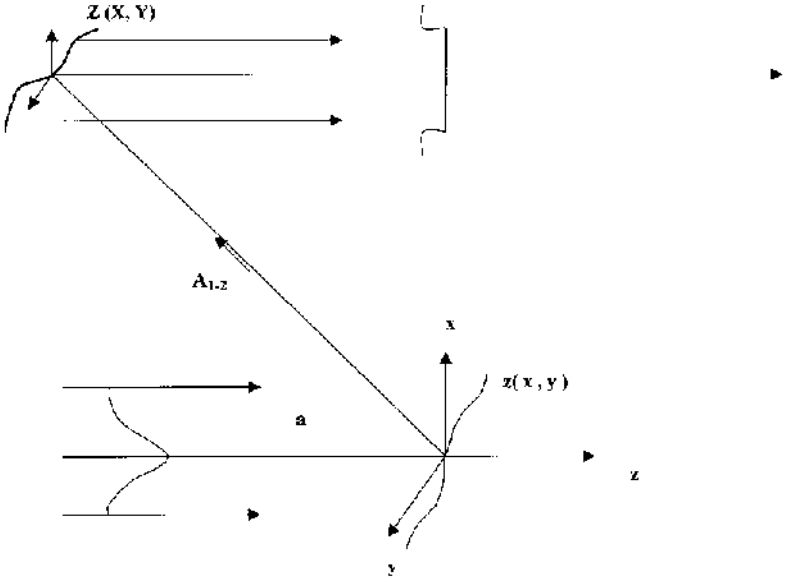


Figure 27 Geometrical configuration of a two-mirror laser profile shaping system.

$$\mathbf{A_{1\to 2}} = \frac{(X-x)\hat{\mathbf{i}} + (Y-y-h)\hat{\mathbf{j}} + (\mathrm{L}+Z-z)\hat{\mathbf{k}}}{\sqrt{(X-x)^2 + (Y-y)^2 + (\mathrm{L}+Z-z)^2}} \tag{111}$$

where L is the distance along the z-axis separating the local coordinate system on each mirror. The functions $z(x, y)$ and $Z(X, Y)$ are the sag of each mirror in the local coordinate system of each mirror. Using Eq. (14), the ray vector connecting each mirror may also be written in terms of the slope of the first mirror:

$$\mathbf{A_{1\to 2}} = \mathbf{a} - 2\hat{\mathbf{n}}(\mathbf{a}\cdot\hat{\mathbf{n}}) = \frac{-2z_x\hat{\mathbf{i}} - 2z_y\hat{\mathbf{j}} + (1-z_x^2-z_y^2)\hat{\mathbf{k}}}{[1+z_x^2+z_y^2]} \tag{112}$$

where

$$\mathbf{a} = \hat{\mathbf{k}}, \tag{113}$$

$$\hat{\mathbf{n}} = \frac{-z_x\hat{\mathbf{i}} - z_y\hat{\mathbf{j}} + \hat{\mathbf{k}}}{\sqrt{1+z_x^2+z_y^2}}, \tag{114}$$

$$z_x = \frac{\partial z(x,y)}{\partial x} \quad \text{and} \quad z_y = \frac{\partial z(x,y)}{\partial y}. \tag{115}$$

Equating the x–y and x–z components of the ray vector $\mathbf{A_{1\to 2}}$ from Eqs. (111) and (112) gives the following ray path equations:

$$\frac{z_x}{X-x}=\frac{z_y}{Y-y-h}, \tag{116}$$

$$\frac{-2z_x}{X-x}=\frac{1-z_x^2-z_y^2}{\mathrm{L}+Z-z}. \tag{117}$$

Solving Eqs. (116) and (117) z_x leads to a quadratic equation with the following solutions:

$$z_x=\frac{(X-x)((\mathrm{L}+Z(X,Y)-z(x,y))}{(X-x)^2+(Y-y-h)^2} \pm\frac{(X-x)\sqrt{((\mathrm{L}+Z(X,Y)-z(x,y))^2+(X-x)^2+(Y-y-h)^2}}{(X-x)^2+(Y-y-h)^2}. \tag{118}$$

Equation (118) is a partial differential equation for the unknown mirror surface functions $z(x,y)$ and $Z(X,Y)$.

The constant optical path length (OPL) condition provides another independent condition to be satisfied by the mirror surface functions $z(x,y)$ and $Z(X,Y)$. The OPL for an axial ray and general ray are given by

$$(OPL)_{\text{Axial Ray}}=\sqrt{\mathrm{L}^2+h^2}, \tag{119}$$

$$(OLP)_{\text{General Ray}} =\sqrt{(X-x)^2+(Y-y-h)^2+((\mathrm{L}+Z(X,Y)-z(x,y))^2} +Z(X,Y)-z(x,y). \tag{120}$$

However, the OPL is constant for all rays. Therefore, equating the right-hand side of Eqs. (119) and (120) leads to the following expression for $Z(X,Y)-z(x,y)$:

$$Z(X,Y)-z(x,y)=\frac{-(x-X)^2-(y-Y+h)^2+h^2}{2(\mathrm{L}+l_0)} \tag{121}$$

where $l_0=\sqrt{\mathrm{L}^2+h^2}$. Using the negative sign in Eq. (118) as the physically meaningful solution, it has been shown (10) that combining

Eqs. (121), (118), and (116) leads to the following expressions for $z_x(x, y)$ and $z_y(x, y)$:

$$z_x = \frac{x - X}{\mathrm{L} + l_0}, \tag{122}$$

$$z_y = \frac{y - Y + h}{\mathrm{L} + l_0}. \tag{123}$$

Assuming the sag of the first mirror can be written in the form

$$z(x, y) = \int z_x(x, y)\, dx + \int z_y(x, y)\, dy, \tag{124}$$

where $z_x(x, y)$ and $z_y(x, y)$ are given by Eqs. (122) and (123). Then, the sag of the first mirror can be written as

$$z(x, y) = \frac{1}{\mathrm{L} + l_0} \left\{ \int_0^x u[1 - m_x(u)]\, du + \frac{1}{\mathrm{L} + l_0} \int_0^y v[1 - m_y(v)]\, dv + C_4 \right\} \tag{125}$$

where C_4 is a constant of integration, and Eqs. (103) and (104) have been used. Expressions for m_x and m_y in Eq. (125) are found from Eqs. (107)–(110). Finally, an expression for the sag of the second mirror follows from Eq. (121):

$$Z(X, Y) = \left[z(x, y) + \frac{-(x - Y)^2 - (y - Y + h)^2 + h^2}{2(\mathrm{L} + l_0)} \right]_{\substack{x = m_x^{-1}(X)X \\ y = m_y^{-1}(Y)Y}} \tag{126}$$

where the x and y terms are eliminated from Eq. (126) by solving for the inverse of the magnifications.

A number of interesting applications of these results are presented in Refs. 10 and 46. These applications include transformation of a linear ramp beam profile of one slope to another slope with different offset distances from the x or y axes and development of an unobscured two-mirror laser profile shaping system.

2. Systems with Polar Symmetry

A general solution for the shapes of the two mirror surfaces of a rotationally symmetric laser profile shaping system is given in Ref. 46. In this section, the results developed in the previous section for rectangular symmetry will be used to obtain equations for the sag of the two mirror surfaces for a laser profile shaping system with polar symmetry.

Assuming both the input and output beam profiles have rotational symmetry, then a ray entering the system at a polar angle $\theta = \theta_0$ with respect

to the x–y coordinate system of the first mirror will leave the system at a polar angle $\Theta = \Theta_0$ with respect to the X–Y coordinate system of the second mirror. Converting to polar coordinates and setting $\theta = \Theta$ as a result of radial symmetry of the beams, then the partial derivatives of the surface sag of the first mirror are given by

$$g_r = g_x \frac{\partial x}{\partial r} + g_y \frac{\partial y}{\partial r} = \frac{1}{L + l_0} [r - R + hr \sin \theta], \tag{127}$$

$$g_\theta = g_x \frac{\partial x}{\partial \theta} + g_y \frac{\partial y}{\partial \theta} = \frac{1}{L + l_0} [hr \cos \theta] \tag{128}$$

where

$$x = r \cos \theta; \quad y = r \sin \theta; \quad X = R \cos \Theta; \quad Y = R \sin \Theta. \tag{129}$$

Integrating Eqs. (127) and (128) gives

$$g(r, \theta) = \frac{1}{L + l_0} \left\{ \int_0^r r'[1 - m(r')]\, dr' + hr \sin \theta \right\} \tag{130}$$

where $m(r)$ is given by Eq. (56). An equation for the sag of the second mirror, $G(R, \Theta)$, can be determined by substituting Eq. (130) into Eq. (126). Alternatively, an expression for $G(R, \Theta)$ follows from geometric considerations when the input and output beams are parallel to the z-axis. As a result of symmetry, the surface of the first mirror at the point (r, θ) is parallel to the surface of the second mirror at the point (R, Θ), where (r, θ) and (R, Θ) are connected by a ray. Thus, the partial derivatives of g and G are equal, that is, $G_R = g_r$ and $G_\Theta = g_\theta$. Following a similar derivation leading to Eq. (130) gives

$$G(R, \Theta) = \frac{1}{L + l_0} \left\{ \int_0^R R'[m^{-1}(R') - 1] dR' + hR \sin \Theta \right\}. \tag{131}$$

Recall that beam shaping places a constraint on the ray heights as expressed through the radial magnification given by Eqs. (55) and (56).

This completes the analysis of a two-mirror system. To design a specific two-mirror laser beam shaping system, the radial magnification function $m(r)$ must be determined. Then, the mirror surface sag functions are computed from Eqs (130) and (131). Malyak (10) presents specific examples of a Gaussian input beam being transformed into a uniform intensity output beam with a smaller diameter. Integral equations for the sag of the first and second mirror surfaces is given, and several numerical results for the mirror surfaces are also presented in Ref. 10.

VI. CONCLUSION

Geometrical methods have been used to design reflective and refractive optical systems that will transform a collimated input laser beam with a Gaussian intensity profile into a collimated (expanded or reduced) output beam with a uniform intensity profile. These laser beam shaping optical systems have been designed by varying the shapes of two optical surfaces until the design conditions (conservation of energy and constant optical path length between input and output beams) are satisfied. It has also been shown that the gradient-index profiles of two plano-spherical lenses can be varied to satisfy the design conditions of laser profile beam shaping systems. In this chapter, the optical design conditions have been expressed in terms of differential equations, which have been solved to determine the shapes of the optical surfaces.

For a two-element refractive system, a prototype laser beam profile shaping system has been designed, fabricated, and tested. Experimental results confirm validity of the geometrical methods presented in this chapter for designing laser profile shaping systems. When laser light of a different wavelength than used for design, a model has been developed for predicting the change in inter-lens spacing (d-spacing) which is necessary to shape a laser beam profile with a new wavelength. Experimental results confirm that a two-element refractive system, which was designed for use with HeCd (441.57 nm) will effectively shape the profile of a HeNe (632.8 nm) laser when the inter-lens spacing is adjusted.

The methods of McDermit and Horton (6,7), Cornwell (46,47) and Malyak (10) for designing one- or two-mirror laser profile shaping systems are presented and discussed in Sec. V.

REFERENCES

1. DR Herriot. Applications of laser light. Scientific American 219: 140–156, 1968.
2. L Levi. Applied Optics. New York: John Wiley and Sons, 1968, p. 322.
3. WB Veldkamp, CJ Kastner. Beam profile shaping for laser radars that use detector arrays. Applied Optics 21: 345–356, 1982.
4. HM Haskal. Laser recording with truncated Gaussian beams. Applied Optics 18: 2143–2146, 1979.
5. WW Simmons, GW Leppelmeir, BC Johnson. Optical beam shaping devices using polarization effects. Applied Optics 13: 1629–1632, 1974.
6. JH McDermit, TE Horton. Reflective optics for obtaining prescribed irradiative distributions from collimated sources. Applied Optics 13: 1444–1450, 1974.

7. JH McDermit. Curved reflective surfaces for obtaining prescribed irradiation distributions. PhD Dissertation, University of Mississippi, Oxford, MS, 1972.
8. JW Ogland. Mirror system for uniform beam transformation in high-power annular lasers. Applied Optics 17: 2917–2923, 1978.
9. PW Scott, WH Southwell. Reflective optics for irradiance redistribution of laser beams: design. Applied Optics 20: 1606–1610, 1981.
10. PW Malyak. Two-mirror unobscured optical system for reshaping the irradiance distribution of a laser beam. Applied Optics 31: 4377–4383, 1992.
11. JL Kreuzer. Coherent light optical system yielding an output beam of desired intensity distribution at a desired equiphase surface. US Patent 3,476,463, 4 November 1969.
12. PW Rhodes, DL Shealy. Refractive optical systems for irradiance redistribution of collimated radiation: their design and analysis. Applied Optics 19: 3545–3553, 1980.
13. PW Rhodes. Design and analysis of refractive optical systems for irradiance redistribution of collimated radiation. MS Thesis, The University of Alabama Birmingham AL, 1979.
14. D Shafer. Gaussian to flat-top intensity distributing lens. Optics & Laser Technology 14: 159–160, 1982.
15. Y Ozaki, K. Takamoto. Cylindrical fly's eye lens for intensity redistribution of an excimer laser beam. Applied Optics 28: 106–110, 1989.
16. AJ Glass, A Greenbaum, J Trenholme. In: JI Davis, W Clements, eds, Laser Program Annual Report UCRL-50021-74, Lawrence Livermore Laboratory, 234-242, 1974.
17. WB Veldkamp. Laser beam profile shaping with interlaced binary diffraction gratings. Applied Optics 21: 3209–3212, 1982.
18. CY Han, Y Ishii, K Murata. Reshaping collimated laser beams with Gaussian profile to uniform profiles. Applied Optics 22: 3644–3647, 1983.
19. CC Aleksoff, KK Ellis, BD Neagle. Holographic conversion of a Gaussian beam to a near-field uniform beam. Optical Engineering 30: 537–543, 1991.
20. U Levy. Diffraction optics spawns new products and a multimillion-dollar market. Photonics Spectra 135–140, May 1992.
21. J Cordingley. Application of binary diffractive optic for beam shaping in semiconductor processing by lasers. Applied Optics 32: 2538–2542, 1993.
22. MA Golub, M Duparre, E-B Kley, RKB Ludge, W Rockstroh, H-J Fuchs. New diffractive beam shaper generated with the aid of e-beam lithography. Optical Engineering 35: 1400–1406, 1996.
23. W Jiang, DL Shealy, JC Martin. Design and testing of a refractive reshaping system. Proc SPIE 2000: 64–75, 1993.
24. W Jiang. Application of a laser beam profile reshaper to enhance performance of holographic projection systems. PhD Dissertation, The University of Alabama, Birmingham AL, 1993.
25. W Jiang, DL Shealy, KM Baker. Optical design and testing of a holographic projection system. Proc SPIE 2152: 244–252, 1994.

26. C Wang, DL Shealy. Design of gradient-index lens systems for laser beam shaping. Applied Optics 32: 4763–4769, 1993.
27. P Mouroulis, J Macdonald. Geometrical Optics and Optical Design. New York: Oxford University Press, 1997.
28. M Born, E Wolf. Principles of Optics. 5th ed. New York: Pergamon Press, 1975.
29. WT Welford. Aberrations of the Symmetrical Optical System. New York: Academic Press, 1974.
30. GA Deschamps. Ray Techniques in Electromagnetics. Proc IEEE 60: 1022–1035, 1972.
31. S Solimeno, B Crosignani, P DiPorto. Guiding, Diffraction, and Confinement of Optical Radiation, Orlando, FL: Academic Press, 1984.
32. AK Ghatak, K Thyagarajan, Contemporary Optics. New York: Plenum Press, 1980, p. 24.
33. FA Jenkins, HE White. Fundamentals of Optics. New York: McGraw-Hill, 1957.
34. M Born, E Wolf. Principles of Optics, 5th ed. New York: Pergamon Press, 1975, p. 115.
35. DG Koch. Simplified Irradiance/Illuminance Calculations in Optical Systems. Presented at the International Symposium on Optical Systems Design, Berlin, Germany, 14 September 1992, and published in Proceedings of SPIE 1780:226–240, 1993.
36. DG Burkhard, DL Shealy. Simplified formula for the illuminance in an optical system. Applied Optics 20: 897–909, 1981.
37. DG Burkhard, DL Shealy. A different approach to lighting and imaging: formulas for dlux density, exact lens and mirror equations and caustic surfaces in terms of the differential geometry of surfaces. Proc. SPIE 692, 248–272, 1986.
38. ON Stavroudis. The Optics of Rays, Wavefronts, and Caustics. New York: Academic Press, 1972, pp. 82–84.
39. Military Standardization Handbook – Optical Design, section 6.8, volume MIL-HDBK-141, Defense Supply Agency, Washington 25, DC, 5 October 1962.
40. ON Stavroudis. The Optics of Rays, Wavefronts, and Caustics. New York: Academic Press, 1972, Chapter II.
41. Melles Griot, Optics Guide 5, Irvine, CA, 1989.
42. WJ Smith, Genesee Optics Software, Inc. Modern Lens Design. New York: McGraw-Hill, 1992, p. 51.
43. A Meddler, R Mead. A simplex method for function minimization. Computer Journal 7: 308, 1965.
44. NC Evans, DL Shealy. Design and optimization of an irradiance profile shaping system with a genetic algorithm method. Applied Optics 37: 5216–5221, 1998.
45. M Born, E Wolf. Principles of Optics, 5th ed. New York: Pergamon Press, 1975, Sec. 4.3.

46. DF Cornwell. Non-projective transformations in optics. PhD Dissertation, University of Miami, Coral Gables, FL, 1980.
47. DF Cornwell. Nonprojective transformations in optics. Proc SPIE 294: 62–72, 1981.
48. KM Baker, DL Shealy, W Jiang. Directional light filters: three-dimensional azo dye formed images within optical resins. Proc SPIE 2404: 144–158, 1995.
49. KM Baker. Highly corrected submicrometer grid patterning on curved surfaces. Applied Optics 38: 339–351, 1999.
50. KM Baker. Highly corrected close-packed microlens arrays and moth-eye structuring on curved surfaces. Applied Optics 38: 352–356, 1999.
51. Ealing Electro-Optics, Product Guide, Holliston, MA, 1989.
52. WL Wolf, GJ Zissis, (eds.). The Infrared Handbook, revised edition. Ann Arbor, MI: Environmental Research Institute of Michigan, 1989, Chapter 7.
53. W Rasband. NIH Image 1.44. This is a public domain image processing and analysis program for the Macintosh computer. It is available via anonymous FTP from zippy.nimh.nig.gov (128.231.98.32). The email address of the author of this software is wayne@helix.nih.gov.
54. D Malacara. Optical Shop Testing. New York: John Wiley and Sons, 1977, pp. 283–322.
55. These Ronchigrams were recorded by W. Jiang during her graduate work at The University of Alabama at Birmingham and have not been published.
56. PJ Sands. Inhomogeneous lenses. IV. Aberrations of lenses with axial index distributions. Journal of the Optical Society of America 61: 1086–1091, 1971.
57. C Wang, DL Shealy. Design of gradient-index lens systems for laser beam reshaping. Applied Optics 32: 4763–4769, 1993.
58. EW Marchand. Gradient Index Optics. New York: Academic Press, 1978, Chapter 5.
59. K Shingyouchi, S Konishi. Gradient-index doped silica rod lenses produced by a solgel method. Applied Optics 29: 4061–4063, 1990.
60. JE Samuels, DT Moore. Gradient-index profile control from mixed molten salt baths. Applied Optics 29: 40420-4050, 1990.
61. DS Kindred, J Bentley, DT Moore. Axial and radial gradient-index titania flint glasses. Applied Optics 29: 4036–4041, 1990.
62. Y Koike, H Hidaka, Y Ohtsuka. Plastic axial gradient-index lens. Applied Optics 24: 4321–4325, 1985.
63. LightPath Technologies, Inc., 6820 Academy Parkway, NE, Albuquerque, NM 87109. See http://www.lightpath.com .
64. See http://www.lightpath.com/GTHTML/GTGPGPP.HTM.

5

Optimization-based Techniques for Laser Shaping Optics

Neal C. Evans and David L. Shealy
University of Alabama at Birmingham, Birmingham, Alabama

I. INTRODUCTION

Recently, the application of machine learning techniques including neural networks and genetic algorithms (GAs) to optimization problems has blossomed. Such techniques hold great promise not only because of their extraordinary efficiency and flexibility but also because they potentially allow the solution of previously intractable problems. So long as a fitness landscape* can be well defined, a GA can be unleashed to roam this territory in an incessant search for the best solutions. The GA must not, however, be characterized as a mindless automaton that wanders aimlessly about this terrain. Indeed, the essence of its value lies in the fact that the "... genetic algorithm [can] yield computer-based complex adaptive systems that can evolve strategies that no human being ever devised" (1).

Though there are numerous variations of GAs, they all share a central theme: their search strategy borrows concepts from natural selection and genetics (2). Once presented with a specific optimization problem, the GA produces a set of potential solutions. These solutions are referred to as "organisms" and a set of organisms is a "generation." GAs typically start with a randomly distributed seed generation, $G(0)$. For each generation

* Or, using more conventional optics terminology, a multi-dimensional "merit function" which quantifies the worth of an optical system in terms of the various parameters which define the system.

$G(t)$, a new generation, $G(t+1)$ is produced based on the strengths and weaknesses of $G(t)$. Organisms are represented by a single string, or chromosome, which is built from the values of the parameters to be optimized for a particular problem (3). These techniques attribute to GAs several unique features, as described by Goldberg (4):

1. GAs work with a coding of the parameter set, not with the parameters themselves.
2. GAs search from a population of points, not a single point.
3. GAs use payoff... [merit function] information only, not derivatives or other auxiliary knowledge.
4. GAs use probabilistic transition rules, not deterministic rules.

This evokes the intriguing thought of employing GAs to find solutions to problems in optics and optical design where analytical methods are difficult to apply and other optimization techniques are extremely inefficient or fail to yield good solutions altogether. As a first step, one must develop a GA optimization method and apply it to several well-understood problems. The key to this "proof-of-principle" stage lies in the fact that these problems have been attacked from a number of different perspectives. Not only does this provide a basis for judging the efficiency of the GA relative to other optimization techniques, but also answers the fundamental question "Does this method work?" For the applications presented in this chapter, a GA method is developed that can be used to design laser beam-shaping systems that convert Gaussian input irradiance profiles to uniform output profiles.

The laser beam-shaping system problem can be solved by a number of different methods, some of which are numerical and some of which are analytical. Some may even employ a combination of both. Also, there are several different classes of beam-shaping systems, the most popular being those using diffractive elements (5–8) and those using refractive elements (9,10). Reflective systems have also been produced (11). Many of the systems and methods are discussed in other chapters in this book. The GA method can be used to optimize most systems of the above classes, which will be demonstrated in this chapter by the solution of three laser profile-shaping problems.

A. Scope of Applications

For the first problem, the GA determines the shape of one surface of a beam-shaping element such that the wavefront of a beam entering the system is modified to have a uniform irradiance profile on a surface some distance away (12). To increase the complexity of the problem a bit, the beam is shaped such that it is uniform on a *spherical* surface. Thus, the

system is diverging and the non-paraxial aspects of the system must be accounted for in subsequent irradiance calculations. A similar problem has recently been addressed in the literature using diffractive elements and a parametric optimization method (13).

For the second problem, the GA is given two aspherical surfaces to shape, where each respective surface is part of a separate shaping element. The GA must do this with the constraints that the outgoing beam is parallel to the optical (Z-) axis and that it has a specified radius. This problem is designed to mimic the system designed by Jiang, Shealy and Martin (14) presented in Chapter 4, Sec. IV.A. This should make for an interesting comparison of the efficiency of the two methods, in addition to demonstrating whether multiple solutions to the problem exist.

The final problem presented in this chapter is inspired by a gradient-index shaping system designed by Wang and Shealy (15). For this problem, the GA not only must determine shaping attributes such as element thickness and surface shape, but also must choose gradient glass types from a catalog. Since the glass type can only be chosen from a finite set of values, the parameters that describe the glass types are discrete. Many conventional optimization techniques work in a continuous parameter space, since they are often driven by first- and second-order derivatives. The ability to choose from discrete parameters is a particularly powerful feature of the GA, relative to other optimization codes. The GA method here is not a unique application in optics. Indeed, it should be noted that many other examples of GA-designed systems can be found in the optics literature (16,17). In the next sections, the fundamental principles governing GA optimization are introduced by describing its application to the problems above.

B. Computational Methods for Irradiance Calculations via Ray-trace Methods

Fundamental to laser beam shaping computations is a fast, accurate means of determining irradiance (energy per unit area per unit time) profiles at different locations in a system. To do this, one must start with first principles: energy must be conserved in a non-dissipative optical system. This principle is mathematically expressed in the form of the energy conservation law. The energy conservation law has broad application, from designing reflective beam shapers via analytical differential equation methods to the development of finite-element mesh methods for the design of beam shaping holograms (18–20). To employ energy conservation, one starts by describing the irradiance profile of a bundle of rays striking the input pupil of a beam profile-shaping system by a radially symmetric function $\sigma(\rho)$. These rays propagate through the beam profile-shaping system (the ''black box'') and

exit to strike the output surface. The irradiance distribution on the output surface is represented by the function $u(P)$. Assuming no energy is dissipated by the system, the energy conservation law (21) must be satisfied:

$$E = \int_I \sigma(\rho)(\hat{\mathbf{n}}^{\text{in}}(\rho) \cdot \mathbf{v}^{\text{in}}(\rho))\, da = \int_O u(P)(\hat{\mathbf{n}}^{\text{out}}(P) \cdot \mathbf{v}^{\text{out}}(P))\, dA, \tag{1}$$

where E is the total energy entering the system and I and O are the input plane and output surfaces, over which the respective integrations occur. Also, $\hat{\mathbf{n}}^{\text{in}}(\rho)$ and $\hat{\mathbf{n}}^{\text{out}}(P)$ are the normal vectors on the input and output surfaces, respectively. $\mathbf{v}^{\text{in}}(\rho)$ and $\mathbf{v}^{\text{out}}(P)$ are unit vectors along the direction of an individual ray (striking the input surface at radial height ρ, and the output surface at radial height P) at the input and output surfaces, respectively. See Fig. 1 for further elaboration of terms in Eq. (1). da and dA are derived below. Balancing the radiant energy striking differential ring da with the radiant energy exiting differential ring dA, as required by Eq. (1), one can see that

$$u(P) = \sigma(\rho)\, \frac{(\hat{\mathbf{n}}^{\text{in}}(P) \cdot \mathbf{v}^{\text{in}}(P))\, da}{(\hat{\mathbf{n}}^{\text{out}}(P) \cdot \mathbf{v}^{\text{out}}(P))\, dA}. \tag{2}$$

To determine the ratio da/dA in Eq. (2), N rays are traced through the system, where N is a reasonably large number, though not so large as to be computationally expensive. For this application, $N = 200$ is chosen, which gives adequate resolution for the input and output profiles. Each ray enters parallel to the optical (Z-) axis at a specified height, ρ_i, where

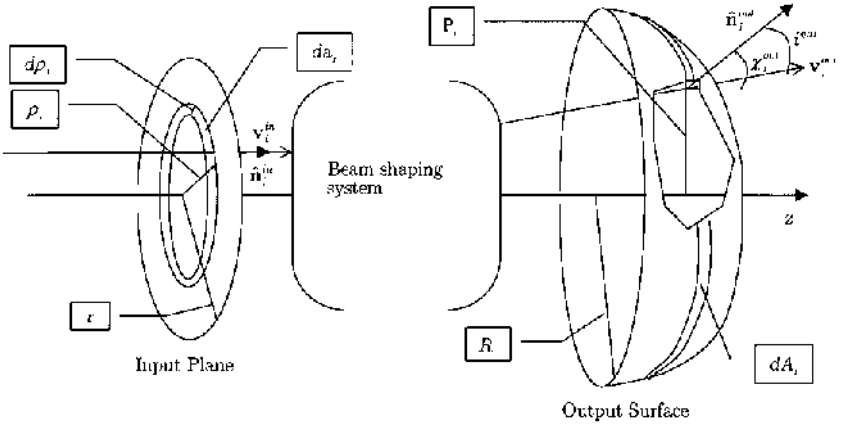


Figure 1 Beam expander with input plane and output surface. The beam profile is shaped to be uniform on the output surface. (From Ref. 12).

the set of ρ_i are distributed equally across the radius of the input plane according to the following function:

$$\rho_i = \left(\frac{r}{N}\right) i, \quad i = 0, \ldots, N. \tag{3}$$

Each ray will exit the system and strike a point on the output surface, as shown in Fig. 1. At the point where the ray strikes the output surface, P_i, the axial distance from the optical axis, and χ_i, the angle between the unit vector normal to the output surface at the intercept point and the optical (Z-) axis, are measured. Thus, an array with $3N$ members (three columns: ρ_i, P_i, χ_i, and N rows) is populated from ray trace data.

One can see in Fig. 1 that da_i is given by $2\pi\,\rho_i\,d\rho_i$, where

$$d\rho_i = \rho_i - \rho_{i-1}. \tag{4}$$

The definition of $d\rho_i$ in this manner is arbitrary; definitions such as $d\rho_i = \rho_{i+1} - \rho_i$ or $d\rho_i = \rho_{i+1} - \rho_{i-1}$ would be just as effective. Furthermore, the subscript i is introduced to emphasize the numerical nature of the solution to the now discrete function in Eq. (2). Calculation of dA_i is somewhat more complicated, since the output surface is a not necessarily flat like the input plane. In general, dA_i is given by $2\pi P_i\,dS_i$, where dS_i is found by referring to Fig. 2:

$$dS_i = \frac{dP_i}{\cos\chi_i^{\text{out}}}. \tag{5}$$

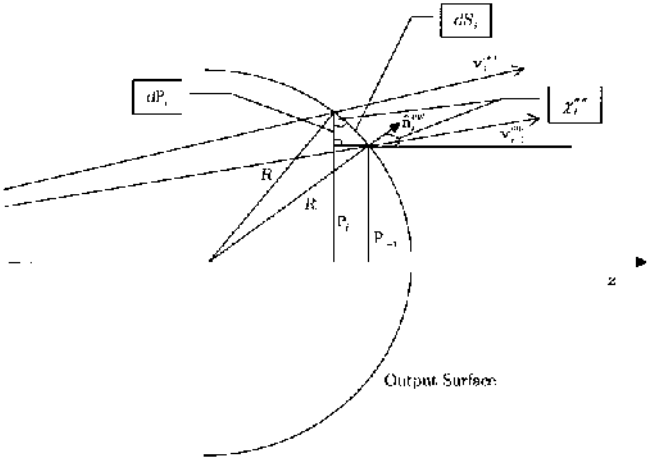


Figure 2 Determination of dA_i. From the figure, one can see that $dP_i = P_i - P_{i-1}$ and that $dS_i = dP_i/\cos(\chi_i^{\text{out}})$. (From Ref. 12).

Also, it is clear from Fig. 1 and Fig. 2 that

$$\hat{\mathbf{n}}^{\text{in}}(\rho) \cdot \mathbf{v}^{\text{in}}(\rho) = \cos(i^{\text{in}}(\rho)) \tag{6}$$

and

$$\hat{\mathbf{n}}^{\text{out}}(P) \cdot \mathbf{v}^{\text{out}}(P) = \cos(i^{\text{out}}(P)). \tag{7}$$

Combining these observations with Eqs. (2)–(5), one has

$$u(P_i) = \sigma(\rho_i)\left(\frac{\cos(i_i^{\text{in}})\rho_i(\rho_i - \rho_{i-1})\cos(\chi_i^{\text{out}})}{\cos(i_i^{\text{out}})P_i(P_i - P_{i-1})}\right). \tag{8}$$

In the examples presented in this chapter, the input irradiance is assumed to be Gaussian, measured in units of rays per unit area:

$$\sigma(\rho_i) = \exp(-\alpha\rho_i^2), \tag{9}$$

where α is a unitless quantity given by $2/\rho_N^2$. Here, the beam waist of the incoming beam is expressed by ρ_N, and is defined as the radius of the circle where the irradiance drops to $1/e^2$ of the central irradiance. The N rays that are traced through the system are distributed uniformly over the input plane according to Eq. (3). Though it need not be, a Gaussian input profile is chosen because it describes typical laser profiles when the laser is in the fundamental mode (TEM_{00}). Eq. (8) expresses the output beam irradiance in terms of the input beam irradiance times a ratio of areas expressing the beam expansion as a result of ray propagation through the optical system. Eqs. (8) and (9), along with the ray trace array, provide an accurate means of calculating the beam profile over any reasonable surface. The accuracy of this method has been verified by calculating the profiles for several benchmark systems (14,15). Calculations of output beam profiles using Eq. (8) are in close agreement with the profiles given in the benchmark papers. Now, a merit function can be developed based on Eq. (8) which allows the GA to distinguish between systems with uniform intensity profiles (which are desired) and non-uniform profiles. This is done in the sections that follow.

II. THEORY AND OPTIMIZATION

Generally, the idea behind optimization is that one has some function f which may be evaluated easily—usually computationally. This function is expressed in terms of several variables which may be discrete or continuous in nature. One wishes to find the values of these variables which make f assume either its maximum or minimum value. The difficulty of

the problem is related to whether one is searching for local extrema, of which there may be many, or the global extrema, which represent the absolute best solutions. The complexity of the problem is related to the number of variables which make up f, in addition to the ease with which f can be calculated (22). The greater the complexity of the problem, the longer it takes to arrive at a solution. Thus, search algorithms which arrive at solutions quickly are to be coveted, which is evident by the voluminous amount of research regarding the subject present in the literature (23,24). In this chapter, the GA search method is presented as one of these treasured methods, but it should be noted that other algorithms exist which produce similar, if not superior performance. Two popular alternative methods, simulated annealing and the Tabu search, are discussed below.

A. Overview of Iterative Computational Optimization Methods

Though there are myriad optimization techniques to choose from, methods such as GAs and simulated annealing are of particular interest because of their ability to solve combinatorial minimization problems. The key feature of such problems is that one or more of the parameters that make up the merit function (which is to be maximized) are discrete, in the sense that they can assume only particular values from a pre-defined set of allowable values. Thus, instead of an N-dimensional space made up N continuous parameters, one is presented with a parameter space whose complexity is factorially large—so large in fact that it cannot be completely explored (25). Without a continuous merit function, concepts such as "downhill" and "uphill" lose their meaning and other optimization techniques, such as the simplex method (26) can no longer be applied. For example, in the problem presented in Sec. III.B, the glass types of the lens elements in the system are chosen from a fixed set of gradient-index materials found in a manufacturer's catalog. It is here that GAs and simulating annealing techniques excel, though they also can be applied to problems that are purely continuous as well. In the literature, there are numerous examples of problems solved using these techniques (27,28), as well as research that compares the performance of one or more of these methods on the same class of problems (29–31). It seems that of the three methods discussed here, no one method is necessarily more efficient than the others, though it does appear that GAs and the Tabu search tend to arrive at solutions more quickly than simulated annealing methods, at least in the papers cited here.

It should be noted that several commercial optical design and analysis packages implement these techniques in their optimization routines to varying degrees. OSLO*, for example, uses an 'adaptive simulating annealing' method. ZEMAX† and CODE V‡ also contain proprietary global optimization methods. The problem with these implementations, among other things, is that the merit functions in these packages are oriented towards imaging systems (ZEMAX, however, allows for user-defined merit functions computed by macros or an external programming interface), limiting one's ability to manipulate the merit function for one's own purpose. Also, since the makers of these packages keep their optimization codes proprietary, one's ability to tweak those routines is all but eliminated. More ambitious goals like parallelization of the optimization code (see Sec. II.C) becomes extremely difficult, at best.

B. Genetic Algorithms

Since GAs are based on a biological paradigm, a lot of the GA nomenclature is borrowed directly from evolutionary biology. The reader may find it useful to have some of this jargon expressed in terms more familiar to the optics community. As discussed in Sec. I, GAs produce a finite number of test solutions to a problem. Individually, these solutions are referred to as "organisms" (or just "individuals"), and collectively as a "generation." A generation is essentially an iteration. With each iteration, the merit function, M, is evaluated for each member of the generation. There may be as few as five or as many as hundreds of individuals per generation, depending on the code used and how it is configured. In the applications presented in this chapter, there are typically five or ten individuals in a generation. A new generation of child systems is produced from the genetic material of the parent generation (the specifics of this process are described below). An individual's genetic code represents a particular system prescription. For example, in the beam shaper/projector example presented in Sec. III.A, the six parameters that collectively define one surface of the beam-shaping element are concatenated into a string (i.e., genetic code). Thus, with each iteration, five or ten new system prescriptions are produced and their respective merit functions evaluated.

---

* OSLO is a registered trademark of Sinclair Optics, Inc., 6780 Palmyra Road, Fairport, NY, 14450.

† ZEMAX is a registered trademark of Focus Software, Inc., P.O. Box 18228, Tucson, Arizona, 85731.

‡ CODE V is a registered trademark of Optical Research Associates, 3280 E. Foothill Blvd., Pasadena, CA, 91107.

The GA method developed here is based on a micro-GA code*† (32). Micro-GAs have several features which distinguish them from other GA codes. The most prominent of these features is the fact that micro-GAs can operate efficiently with small generation sizes (on the order of 10 individuals per generation). This is important for the applications presented here, since evaluation of the merit function for each individual is a very time-consuming process. The micro-GA is able to work with small generation sizes by checking for "stagnancy" in each generation it produces. Stagnancy is determined by taking the average value of the merit function for all the individuals in a generation, $\bar{M}(t)$, and comparing it to average merit functions values for N parent generations, $\bar{M}(t-1)$, $\bar{M}(t-2), \ldots, \bar{M}(t-N)$. If these values do not differ significantly (a parameter which can be set in the micro-GA, and is usually "tweaked" at the beginning of a problem to produce the most efficient code), then the population is defined as stagnant. Essentially, when stagnancy is detected, the code assumes that the GA is stuck in a local minimum and attempts to add some randomness to the process. When such a situation arises, the GA picks the best of the individuals in a generation, kills the remaining and replaces them with new, randomly selected individuals in the child generation.

"Reproduction" is defined as producing a child generation of new individuals from the genetic material of a parent generation. The individual with the highest value of M in a particular generation is most likely to have its genetic material passed on to the next generation. The "genetic material" for a particular individual is defined by concatenating the binary value for each parameter to be optimized into a binary string (consists of only ones and zeros). See Fig. 3 for an example. New generations are created by a "crossover operator", which swaps chunks of genetic material (strings) between two or more individuals in a generation. With the micro-GA code, crossovers are done in a manner that maintains individual "alleles". An allele is the particular value that a parameter assumes, expressed in string format (33). When a crossover occurs, the strings that represent alleles are not broken into pieces, but are transferred from one individual to

* The GA code is based on ga164.f, D.L. Carroll's FORTRAN Genetic Algorithm Driver. See http://www.staff.uiuc.edu/~carroll/ga.html to download ga164.f, along with documentation and useful information on its implementation. One may find several sites related to ga164.f by using it as a keyword on popular search engines. Ga164.f is based on a Micro-GA method. See Ref. 32 for more information on Micro-GAs.

† See the Naval Research Lab's GA archive, http://www.aic.nrl.navy.mil/galist/, for a comprehensive source of GA codes in several different languages and implementations. This site also is the home of a very useful mailing list in which one may find the state of the art in GA techniques and theory.

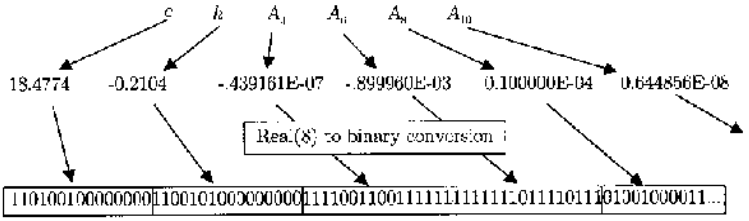


Figure 3 Example of the genetic material for a single individual. Values (Real*8) for each parameter are converted into binary strings, which are in turn concatenated into one long string, the genetic material for that individual (from Ref. 12). Individual alleles (the values that c, k, $A_4, \ldots, A_{10}$ assume, expressed in binary form) are kept intact when crossovers occur.

another intact. Another operation the GA performs is "mutation." Here, the GA randomly selects one or several bits in an allele, and changes the state of these bits. Since the string is binary, this amounts to operating on the bit with "not" (not 0 = 1, not 1 = 0). Mutation adds a built-in randomness to the GA method, which helps the GA avoid local minima. Because of the stagnancy-checking feature it employs, the micro-GA allows one to avoid the constant tweaking of GA parameters (e.g., crossover and mutation rates) which is often necessary with other GA routines. The flowchart for this GA, referred to as the sequential GA, is shown in Fig. 4.

C. Parallelization of Genetic Algorithm

Given that the total execution time for the problems discussed in this chapter is nearly seven hours, it is important to increase the efficiency of the GA method. This can be accomplished by having the code execute in parallel, something facilitated by the nature of the GA. Parallel implementations of GA codes are common in the literature, and several different strategies for parallelization exist (34,35). The two most popular strategies are described below.

For the first strategy, it is interesting to note that, generally, the most computationally intensive step of the optimization process involves the evaluation of the merit function. In these applications, this involves calling optical simulation or optical design packages, such as CODE V, which are external to the GA code itself. For example, in Sec. III.A, the evaluation of the merit function involves calling CODE V and tracing $N = 200$ rays for each system. The calculation time for a single generation with 10 individuals takes about 9 s (on a Sun Ultra 1 170 with 64M of RAM). Of this, about 8 s on average is spent in CODE V. The evaluation of the merit function step is

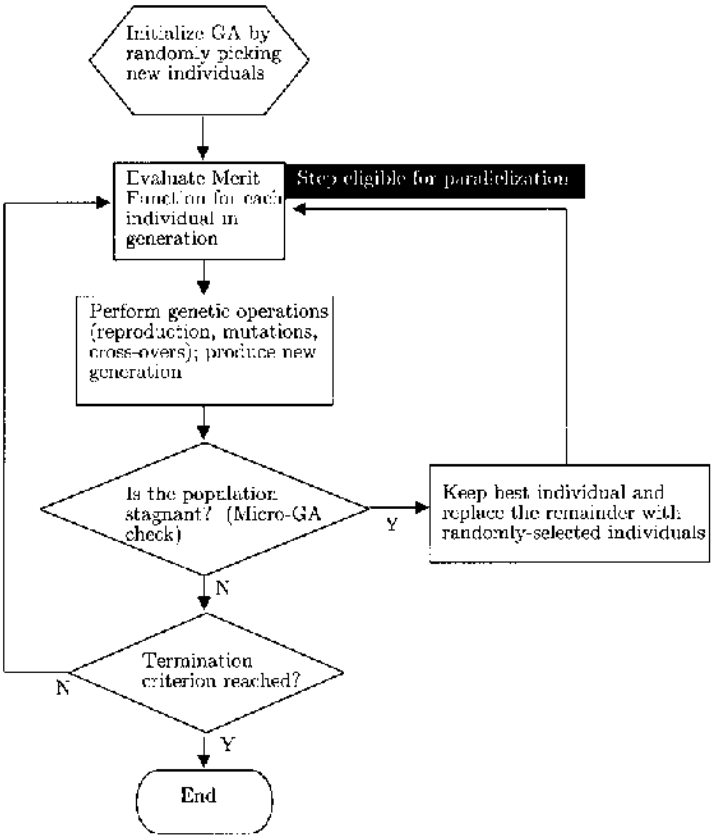


Figure 4 Flowchart of the sequential micro-GA. For definitions of reproduction, mutations, cross-overs and stagnancy see Sec. II.B.

an obvious candidate for parallelization. A potential parallel scheme is one where the GA executes sequentially on one machine, the master node, until it reaches the point where the merit function is to be evaluated. Here, the master initiates processes on each of the slave nodes. The slave nodes, in turn, evaluate the merit functions for individuals in the generation given to them in parallel and return the results to the master. Once all the merit functions are evaluated, the GA code runs normally on the master and executes all GA-related operations, producing the next generation. This parallel scheme, known as the Standard Parallel GA Paradigm, is shown in Fig. 5, along with an alternative parallel scheme, which is explained below.

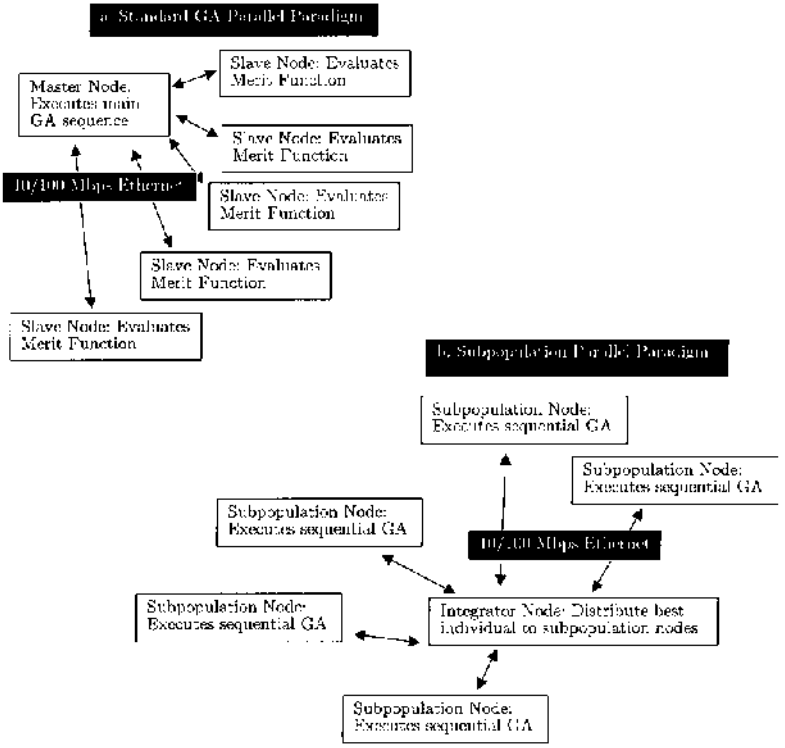


Figure 5 Two parallel GA paradigms. In the first setup, the standard GA paradigm (a), the GA is executed sequentially on the master until the step where the merit function is evaluated. At this point, the merit function is evaluated in parallel on the slave nodes. In the second setup, the subpopulation parallel paradigm (b), the GA executes normally on several.

An alternative scheme, the subpopulation parallel paradigm, essentially runs independent instantiations of the GA code on each node. With time, different nodes produce different best individuals with varying degrees of fitness. The best individuals are periodically sent to a central bookkeeping node, the "integrator." The integrator finds the "best of the best" and distributes this champion to the other slave nodes, where the genetic material of this champion is assimilated by the local subpopulation. If there are enough nodes, each subpopulation can execute on a pod of machines using the Standard Parallel GA Paradigm outlined above. Thus, one must test the efficiency not only of the two paradigms individually, but also a hybrid scheme that incorporates both paradigms.

One could run the parallel GA (PGA) on an inhomogeneous cluster of workstations on the 10/100 Mbps local-area network. Message passing among nodes could be accomplished using the MPI-2 (36,37) libraries, making the PGA code easily scalable to high-performance massively parallel systems. For most research, however, scaling to parallel supercomputers is not feasible since evaluation of the merit function for common problems requires calling proprietary software packages such as CODE V, for which the source code is not (freely) available* (38). If problems are chosen where all source code is available, including that for evaluating the merit function, the GA can be ported to a supercomputer environment with (relatively) little modification.

III. APPLICATIONS

One key feature of the GA method is its broad applicability. Using the theory and tools developed above, one can adapt the GA to solve a multiplicity of problems. The problems presented below are chosen not only to demonstrate this advantage, but to do so while building a logical, concise method that scales from simple to more complex applications. Furthermore, these problems are chosen from current literature and provide a means to compare the solutions generated by the GA with solutions generated by other methods. Hopefully, this will provide insight into where the application of GA methods is appropriate and where they are of little advantage. The three problems chosen here are design and analysis of a beam shaper projector, design and analysis of a gradient-index shaper, and design and analysis of a refractive two-lens shaper.

A. Design and Analysis of a Beam Shaper/Projector

The general goal here is to modify the shape of a lens element (the "shaping element") to uniformly illuminate a spherical surface some distance away. See Fig. 6 for a ray trace of the system. For this particular application, the use of a GA is perhaps a tad overzealous, considering that other more established design methods could be employed to produce solutions both

*It should be noted that there is at least one optical design package for which the source code is available. KDP can be downloaded in compiled form for no charge from http://www.kdpoptics.com. The full source for KDP is available for a reasonable price. Conceivably, one could use KDP for the ray-tracing aspects of the merit function evaluation, and modify the KDP source code so it can be compiled in a supercomputer environment.

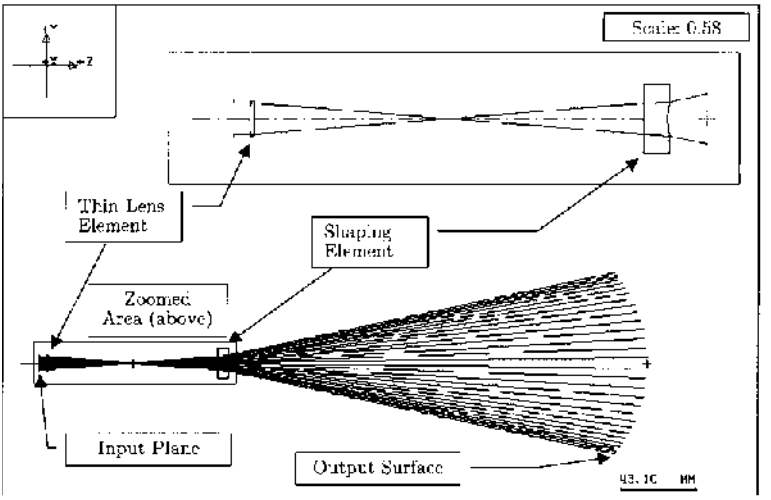


Figure 6 Beam-shaping system with ray trace showing the density of rays increases at the periphery of the output surface, as one would expect to compensate for the Gaussian nature of the input beam. Both the thin lens element and the shaping element are shown. The shaping element is determined by the GA. (From Ref. 12).

easily and efficiently. The goal, however, is a long-term one: the GA technique will be used to attack systems that are difficult to solve with more conventional methods. For example, certain holographic projection systems have fitness landscapes with 20 or more dimensions and extremely complex merit functions, making their solution with conventional methods very tedious (38). For the short term, it must be established that the GA technique produces good solutions in a reasonably efficient manner. This is accomplished by the application of the GA technique to simple, well-understood systems. The insights gained by this application will provide a picture of the fundamental mechanisms that govern this and the nuances involved in its proper implementation.

The application of a GA generally must satisfy two prerequisites. First, one must identify those parameters that fundamentally characterize the system. The parameters must be numerically quantifiable and the modification of these parameters should have direct consequence on the system itself. Second, one must identify those features of a system which best describe the fitness (or ''merit'') of the system. This could be one particular attribute, such as focal length, or, on the other extreme, could involve the blending of many different attributes, each with different weights of influence.

In the application of the GA to the beam profile-shaper, one surface of a refractive lens is modified such that the irradiance profile over a spherical image surface is uniform (that is, $u(P) = \text{constant}$ in Eq. (2)). The rotationally symmetric lens surface is characterized by the conventional surface equation used in optics:

$$z(h) = \frac{ch^2}{1 + \sqrt{1 - (1+k)c^2h^2}} + \sum_{j=2}^{5} A_{2j}h^{2j}, \tag{10}$$

where z is the sag of the surface, c is the curvature of the surface, k is the conic constant and $A_4, \ldots, A_{10}$ are aspherical deformation coefficients. z is an even function of h, the surface radial distance from the optical axis. The choice of this function allows the GA great flexibility in determining the shape of the lens surface, depending on the number of deformation coefficients included in the optimization process. One might speculate that the lens surface must be highly aspherical, based on results from similar beam shaping systems, such as that discussed in Chapter 4 and in Ref. 14. Furthermore, it is desirable to provide the GA with a large (multi-dimensional) parameter space to explore, since this is where GAs are especially powerful. Thus, the GA is given six parameters to optimize: c, k, A_4, A_6, A_8, and A_{10}.

Finally, the GA must be given a means of distinguishing between good systems and bad systems. In this application, a uniform irradiance profile over the output surface is desired. The output surface is a sphere with a radius of $R = 84.12\,\text{cm}$*. Also, the exit pupil have a radius of 50 cm. To accomplish this, the following merit function is defined as follows (12):

$$M = \frac{1}{\mu}\exp[-s(50 - P_N)^2], \tag{11}$$

where

$$\mu = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(u(P_i) - \bar{u})^2} \tag{12}$$

* Ken Baker of Optimetrix Co. provided the initial specifications and requirements for this system. Optimetrix is located at 13659 Victory Boulevard, Van Nuys, CA, 91401. The final system has been fabricated by Optimetrix as part of a holographic projection system. See KM Baker, Highly corrected submicrometer grid patterning on curved surfaces. Appl Opt 38:339–351, 1999.

and

$$\bar{u} = \frac{1}{N}\sum_{i=1}^{N} u(P_i). \tag{13}$$

In Eq. (11), P_N is the radial height (as measured from the optical axis) of the marginal ray on the output surface, which defines the exit pupil in this case. The exponential function is chosen since, for this problem, one wants the merit function to peak sharply at $P_N = 50$ cm. The exponential accomplishes this nicely, but functions other than the exponential may have been chosen for the same purpose. s determines the sensitivity of the merit function to the exit pupil radius constraint: the smaller the value that s assumes, the more prominent the exponent becomes. In this example, s is set to 0.01. This value is adjusted on occasion while the GA is executing to insure that the pupil radius constraint is satisfied. In Eqs. (12) and (13), $\bar{u}$ is the "mean" of the values of the output intensity function, $u(P_i)$ over N points on the output surface. As the beam profile on the output surface becomes more uniform, μ approaches zero, and M increases substantially. Also, as the exit pupil, which is measured by P_N, approaches 50 cm, the value of M peaks as a result of the exponential in Eq. (11). This is illustrated in Fig. 7. Systems with the desired characteristics—a uniform beam profile on the output surface and an exit pupil of 50 cm—will have higher values of M, which is precisely what is required. The GA will find those systems with the highest value of M.

The beam shaping element consists of two surfaces. The first surface is flat and the second surface, of course, has its shape determined by the GA. The GA starts by randomly choosing, within pre-determined constraints, values for the six surface-shape parameters to be optimized. In each generation, ten individuals are produced (see Sec. II.B for explanation of GA nomenclature). The parameters for each individual are passed to a ray-

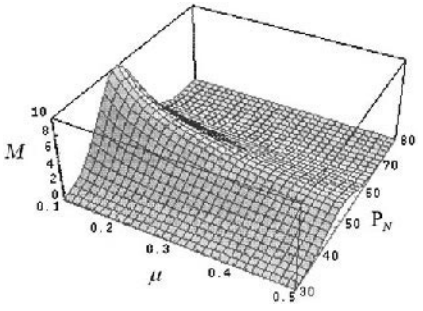


Figure 7 Merit function versus μ and P_N. $N = 200$ in this system. (From Ref. 12).

tracing routine (CODE V is used for ray tracing in this application) where N rays are traced and the value of the merit function, M, is calculated for each individual as above. See Sec. VI for an example of a CODE V macro used to evaluate the merit function and call the GA library. The constraints on each of the surface-shape parameters are given in Table 1.

The GA code was executed simultaneously on four Sun Sparcs, all running Solaris 2.6. The constraints were modified in real time so that each instantiation of the GA code could search different regimes of the parameter space. Once it became apparent that a particular regime contained better solutions, the constraints were narrowed on all machines to search that regime more thoroughly. The constraints given in Table 1 represent the final values of these constraints. Also, when one machine found an individual that was substantially superior to the best individuals on the other three machines, the code on the three other machines was re-initialized using a restart file from the machine with the superior individual. This amounts to a primitive form of parallel processing, an issue which will be better addressed in future applications (see Sec. II.C). Total processing time was not rigorously recorded but was on the order of 12 hours. The fastest machine of the four, a Sun Ultra 1 170 with 64M of RAM, found the best individual. The search was stopped when no significantly better individuals were found over a period of 5 hours, so the best individual actually was discovered after about 7 hours of processing time.

The final profile-shaping system is shown in Fig. 6, with the GA-determined shaping element shown in Fig. 8. In Fig. 6, one notices that the GA-determined shaping element is actually the second element in the system; the first element (the "thin lens element") is an artifact of the initial design requirements and is not part of the optimization process. The purpose of the thin lens element is to focus the incoming collimated beam such

Table 1 Constraints on surface parameters. Each parameter must be between or equal to the end points of the respective constraint.

Surface parameter	Constraint
c	-10 to 20
k	-1.0 to 0
A_4	-1.0×10^{-3} to 1.0×10^{3}
A_6	-1.0×10^{-5} to 1.0×10^{5}
A_8	-1.0×10^{-6} to 1.0×10^{6}
A_{10}	-1.0×10^{-7} to 1.0×10^{7}

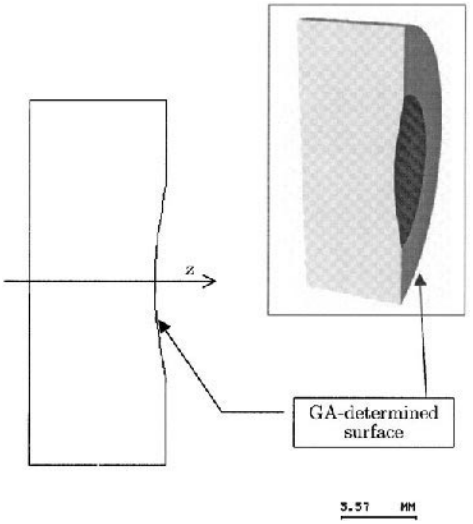


Figure 8 Shaping element showing aspherical surface, which is determined by the GA. The axial thickness of this element is 6 mm (from Ref. 12).

that the numerical aperture of the shaping element is 0.7. The two surfaces of the thin lens element are spherical and the shape factor is set such that spherical aberration is minimized (39). The general parameters for the system are given in Table 2, along with specific parameters for the shaping element and thin lens element in Table 3. The shape of the input beam profile is shown in Fig. 9. It is assumed that the laser beam is circular and is operating in the fundamental mode, TEM_{00}. In Fig. 10, one can see that irradiance on the output surface is nearly uniform, with a average value, $\bar{u}$,

Table 2 Beam shaper/projector system parameters

Parameter	Value
Wavelength	441.57 nm
Radius of the input beam (entrance pupil diameter)	3.9441 mm
Radius of the output aperture	52.5 mm
Glass type for two lens elements	Ohara slah53
Index of ambient medium (air)	1.0
Gaussian constant α in Eq. (9)	0.129 mm^{-2}
Object distance	Infinity
Image distance from shaping surface	242.9 mm

Table 3 Beam shaper/projector lens element parameters

	Thin lens element		Shaping element	
Parameter	Left surface	Right surface	Left surface	Right surface
Aperture radius			8.0 mm	8.0 mm
Thickness	1.0 mm	49.1 mm	6.0 mm	242.9 mm
Vertex radius ($1/c$)	36.606 mm	359.72 mm	infinity	14.235 mm
Surface type	spherical	spherical	spherical	aspherical
Conic constant (k)				0.0
A_4				-0.66411×10^{-3} mm
A_6				0.12400×10^{-5} mm
A_8				-0.31156×10^{-6} mm
A_{10}				-0.61495×10^{-8} mm

of 2.13×10^{-3} rays/mm^2. To check for self-consistency, the irradiance functions, $\sigma(\rho)$ and $u(P)$, are integrated over the input plane and output surface, respectively. The results of these integrations are in close agreement, as expected. The uniformity of the output profile can be characterized by the standard deviation from the mean value of the N points that determine the output profile. The standard deviation for this data set is 3.78×10^{-5} rays mm^2, which is 1.9% of $\bar{u}$.

The GA method produced the desired system—that is, a system with a uniform irradiance profile on the spherical output surface and with an exit pupil very close to 50 mm—in a reasonably efficient manner and while requiring virtually no user input. The GA started with a randomly defined system and found a good solution in about 7 hours. The self-consistency check, which involves integrating the input irradiance function, $\sigma(\rho)$, over the input plane and integrating the output irradiance function, $u(P)$, over the output surface, indicates energy is conserved as required by Eq. (1). The small difference in the two values (1.9% error) can be attributed to small deviations from the mean ($\bar{u}$) in the output irradiance profile data set, as shown in the previous paragraph. Nevertheless, this error would certainly fall within an acceptable range of fabrication for an aspherical surface such as those in this problem (40).

This application illustrates the ability of GAs to solve difficult problems, suggesting the GA may be useful in solving more complex and vexing problems in optics. The efficiency of the GA method can be enhanced by introducing parallelism into the GA code. The most computationally expensive step in the GA routine is the calculation of the merit function, which requires N rays to be traced through the system for each individual in a

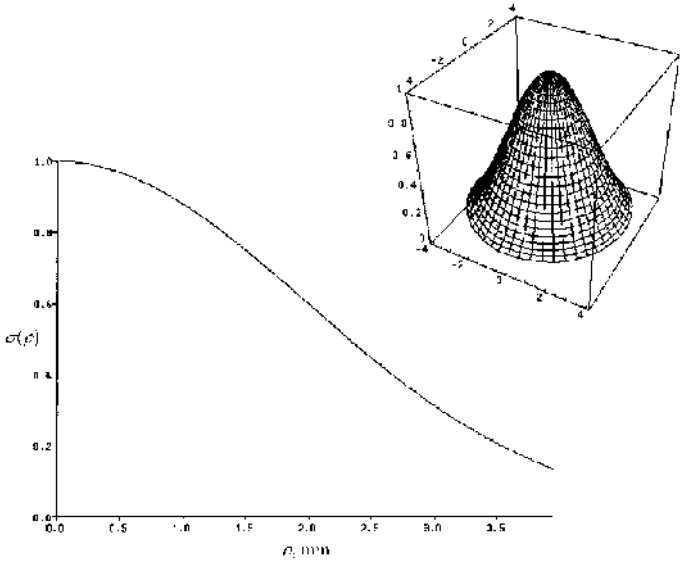


Figure 9 Input beam irradiance profile. The $1/e^2$ diameter of the input beam is 7.882 mm. Integrating $\sigma(\rho)$ over the input plane yields 21.1 units, a quantity which must be conserved according to Eq. (1). (From Ref. 12).

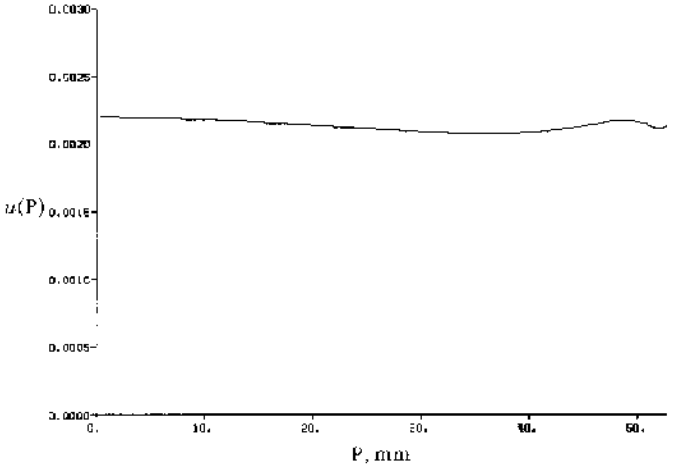


Figure 10 Beam profile on output surface. The radius (P_N) of the output surface is 52.5 mm. The mean value of the profile, $\bar{u}$, is 2.13×10^{-3} rays/mm^2, with a standard deviation of 3.78×10^{-5}. Integrating this mean value ($u(P) = \text{constant} = \bar{u}$) over the output surface yields a value of 20.7 units (from Ref. 12). The beam profile is radially symmetric.

generation. This is done in serial and no other part of the code can execute until the merit function has been calculated. If merit functions for various individuals are calculated in parallel on several slave machines simultaneously, allowing the GA to run unfettered on a master machine, the efficiency of the overall search process could be enhanced significantly.

B. Design and Analysis of a Gradient-Index Shaper

The beam shaper/projector problem presented in the previous section provides an example of a purely continuous merit function. Continuous merit functions can be solved by a number of different methods, and solving it with a GA is not particularly glamorous. As an example of a more complex problem—one difficult to solve using more conventional methods—the GA technique is used to design a gradient-index shaper, which has a merit function that contain both continuous and discrete parameters. This problem was solved using a differential-equation design method in a paper by Wang and Shealy (15). Though Wang was able to produce several perfectly good solutions, the gradients of the lenses reported in Ref. 15 were determined solely by solving the differential equations, and no constraints were imposed that required the indices of refraction of the lenses to correspond to those that can be found in common glass catalogs. It follows that a challenging problem for the GA would be to create a laser shaping system with two gradient-index elements as Wang did, but to do so with the added constraint that the elements can only be chosen from existing glass catalogs. Obviously, this makes the potential for fabricating the system the more easily realized. Furthermore, this is an interesting problem from the perspective of the GA since it is now required to optimize discrete parameters in addition to continuous parameters, adding several nuances to the coding.

The general layout of the gradient-index shaping system is inspired by Wang's system (see Fig. 11). Essentially, there are two shaping elements, each of which are made from a gradient-index glass. The right surface of each shaping element is spherical. The precise shape of the spherical surface and the thickness for each shaping element are optimized by the GA. Furthermore, there is a "connector" between the two shaping elements that is set a priori to be Schott BK7 glass type. The "connector" in Wang's system matches the base index, n_0, of the two gradient elements (see Sec. IV.B in Chapter 4). This is necessary to insure that the marginal ray passes undeviated through the system (15). No such requirement is made for the GA problem, so the glass type of the connector can be set with caprice. The GA is allowed to choose randomly from the Lightpath gradient-index glass catalog available in CODE V. See CODE V documentation for more details on this glass catalog. Thus, the GA optimizes the following

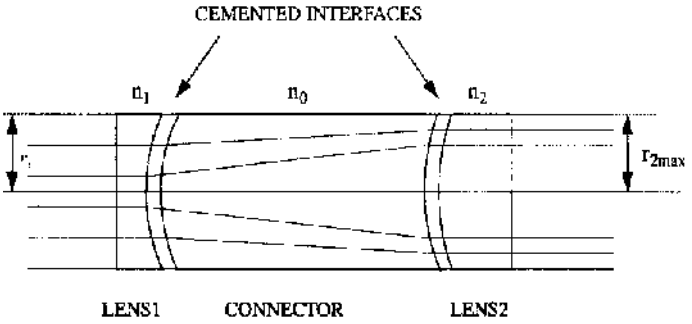


Figure 11 Layout of the gradient-index expander designed by Wang and Shealy (from Ref. 15). This system provides the inspiration for the gradient-index shaper problem.

continuous parameters: thickness of element 1, curvature of right surface of surface 1, thickness of connector, thickness of element 2 and curvature of right surface of surface 2. Also, the GA optimizes the following discrete parameters, Lightpath gradient glass type for surfaces 1 and 2 (four possible types for each surface) and gradient index direction for surface 1 and 2 (may be positive or negative for each respective surface). Of the three problems presented in this chapter, only this problem has a discontinuous parameter space. Because of this, derivative-based methods (e.g. simplex and damped least squares (41) would not be applicable for this problem.

The merit function for this system is designed with three key features. First, the merit function includes a term to characterizes the flatness of the output profile, in the manner shown in Sec. III.A. Second, the merit function includes a term that insures that each ray is perpendicular to the output plane, i.e., a collimated output beam. The form of this term is expressed by the following function:

$$\exp[-(1-\Theta)^2], \tag{14}$$

where

$$\Theta = \prod_{i=1}^{N} \gamma_i^s. \tag{15}$$

γ_i^s in the above equation is the cosine of the angle that ray i makes with the optical axis. The exponents adjusts the sensitivity of the merit function non-parallel ($\gamma \neq 1$) rays. For this application, s is set at six. If each ray traced through the system is parallel to the optical axis, then $\Theta = 1$ and Eq. (14) is

unity. If not, then Eq. (14) is less than one and reduces the merit function, penalizing the system.

Third, the merit function rewards those systems where the radius (P_N) of the output plane is close to 8.0 mm. This is expressed in the following function:

$$\exp[-0.01(8 - P_N)^2], \tag{16}$$

where P_N is the radial height of the marginal ray at the output surface. If P_N is 8.0 mm as desired, then Eq. (16) is unity. If not, then the system is penalized as above. See the Appendix, line 101, to see these terms as they appear in the code, expressed as a complete function. Descriptions of the terms in line 101 are given at the end of the Appendix. Building the merit function in this manner is intended to produce a system that resembles the systems in Ref. 16, which makes similarities and differences more apparent.

To solve this problem, the serial version of the GA is employed and CODE V calculates the merit function. After a total execution time of about 7 hours, the highest value of the merit function (the best individual) failed to significantly improve indicating no better solutions were forthcoming. The final system is shown in Fig. 12 and the shapes of the input and output irradiance profiles, along with consistency checks, are shown in Fig. 13. See Tables 4 and 5 for the parameters of the optimizing system. Examining this best individual, one finds that integrating the irradiance functions, $\sigma(\rho)$ and

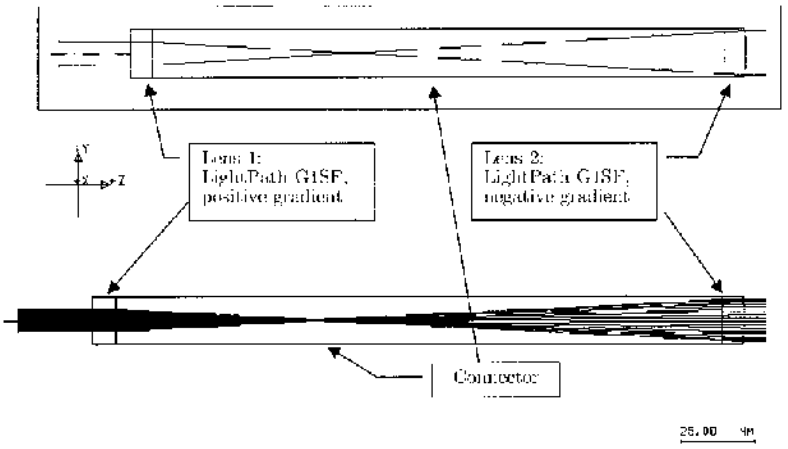


Figure 12 Gradient-index shaper system with ray trace. The materials for Lens 1 and Lens 2 were chosen from a catalog of materials by the GA.

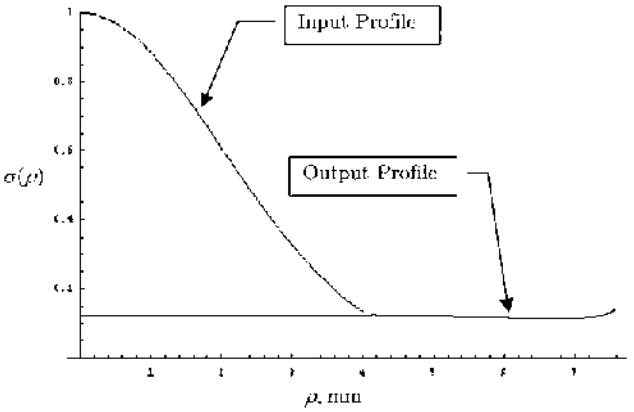


Figure 13 Input and output irradiance profiles for the gradient-index shaper. The $1/e^2$ diameter of the input beam is 8.0 mm. Integrating $\sigma(\rho)$ over the input plane yields 21.7 units. The radius (P_N) of the output plane is 7.56 mm. The mean value of the profile, $\bar{u}$, is 0.121 rays/mm^2, with a standard deviation of 4.45×10^{-3}. Integrating this mean value ($u(P) = \text{constant} = \bar{u}$) over the output surface yields a value of 21.7 units. Energy is conserved, as required in Eq. (1).

Table 4 Gradient-index shaper system parameters

Parameter	Value
Wavelength	589.00 nm
Radius of the input beam (entrance pupil diameter)	4.00 mm
Radius of the output aperture	7.56 mm
Glass type for two lens elements	Lightpath G1SF
Gradient for first lens element	positive
Gradient for second lens element	negative
Glass type for connector	Schott BK1
Index of ambient medium (air)	1.0
Gaussian constant α in Eq. (9)	0.035 mm^{-2}
Object distance	Infinity

$u(P)$ over the appropriate surfaces yields the same value. Energy is conserved as expected. The system satisfies all given constraints and performs with the necessary features; thus, the GA has indeed solved the problem. It is interesting to note that the marginal rays converge inside the connector, resulting in a very long connector length. Such a long connector would be undesirable in any system intended for manufacture. By referring

Table 5 Gradient-index shaper lens element parameters

	First lens element		Second element	
Parameter	Left surface	Right surface	Left surface	Right surface
Thickness	7.75 mm	199.9 mm	7.29 mm	25.0 mm
Vertex radius ($1/c$)	infinity	−99.58 mm	infinity	−58.15 mm
Surface type	spherical	spherical	spherical	aspherical

to Wang's system, one notes that the length of the connector is significantly smaller than the one produced by the GA. This suggests that adding an additional constraint which limits the length of the connector to some small value could force the GA to produce a system with collimated marginal rays. The index of refraction of the connector would also need to be a variable to satisfy the physical constraint that the marginal ray, defined as the ray that enters at the $1/e^2$ point ($\rho_N = 4.0$ mm in this case), passes undeviated through the system.

C. Design and Analysis of a Two-lens Beam Shaper

This system represents the most complex presented in this chapter, where complexity is related to the dimensionality of the merit-function space. This problem is inspired by the system designed, built, and tested by Jiang, Shealy and Martin (14). The system has two lens elements, which are designed to shape an incoming Gaussian beam to an outgoing beam with a uniform irradiance profile. The system expands an 8 mm incoming beam to 12 mm. Both incoming and outgoing beams are parallel to the optical axis. The right surface of the first lens and the left surface of the second lens accomplish the irradiance redistribution and beam expansion. In Ref. 14, these shaping surfaces are highly aspherical.

As above, the merit function must contain a term that quantifies the uniformity of the irradiance profile on an output surface. Furthermore, the merit function must insure that the outgoing rays are parallel to the optical axis and that the radius of the output beam is some predefined value. The precise nature of this merit function is similar to those described in the previous problems. The GA is given 12 parameters to optimize in this problem: $c, k, A_4, A_6, A_8,$ and A_{10} for the aspherical lens surface of each of the two lenses in the system. The merit function for this system has the same form as the one for the beam shaper/projector system outlined in Sec. III.A [Eqs.

Table 6 Two-lens shaper system parameters

Parameter	Value
Wavelength	589.00 nm
Radius of the input beam (entrance pupil diameter)	8.00 mm
Radius of the output aperture	10.7 mm
Glass type for two lens elements	CaF_2
Index of ambient medium (air)	1.0
Gaussian constant α in Eq. (9)	$0.031\,mm^{-2}$
Object distance	Infinity

Table 7 Two-lens shaper lens element parameters

	First lens element		Second element	
Parameter	Left surface	Right surface	Left surface	Right surface
Thickness	10 mm	150 mm	10 mm	25 mm
Vertex radius ($1/c$)	infinity	100.59 mm	−100.01 mm	infinity
Surface type	spherical	aspherical	aspherical	spherical
Conic constant (k)		−0.922971	−0.375469	
A_4		0.843226×10^{-5} mm	0.617298×10^{-4} mm	
A_6		-0.664541×10^{-6} mm	-0.960417×10^{-7} mm	
A_8		0.504624×10^{-8} mm	-0.164098×10^{-8} mm	
A_{10}		0.274667×10^{-14} mm	0.687429×10^{-11} mm	
A_{12}		$-0.999878 \times 10^{-13}$ mm	0.466018×10^{-14} mm	

(11)–(13)], save that "50" in Eq. (11) is "12" in this example (since the radial height of the marginal ray should be 12 mm in this example and not 50 mm). This 12-dimension parameter space is the most complex presented in this chapter. Consequently, the optimization time for this problem is significantly longer, taking some 50 hours on the platforms described above, using the serial processing paradigm. The GA, nevertheless, found a solution, which is presented in Table 6, Table 7, Fig. 14 and Fig. 15.

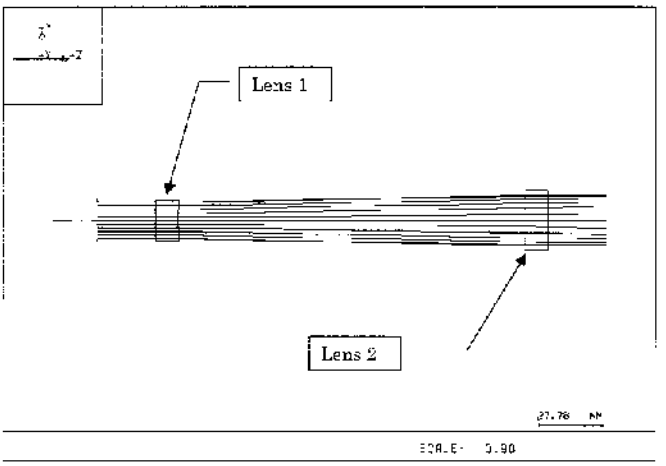


Figure 14 Two-lens beam shaper system with ray trace. The right surface of Lens 1 and the left surface of Len 2 are shaped by the GA.

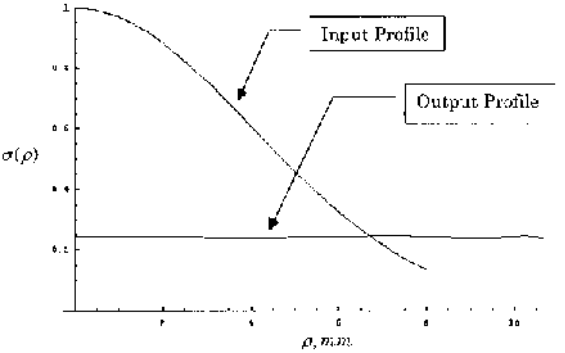


Figure 15 Input and output irradiance profiles for the two-lens beam shaper. The $1/e^2$ diameter of the input beam is 16.0 mm. Integrating $\sigma(\rho)$ over the input plane yields 86.9 units. The radius (P_N) of the output plane is 10.7 mm. The mean value of the profile, $\bar{u}$, is 0.242 rays/mm^2, with a standard deviation of 1.86×10^{-3} rays/mm^2, or 0.8% of $\bar{u}$. Integrating this mean value ($u(P) = \text{constant} = \bar{u}$) over the output surface yields a value of 87.0 units. Energy is conserved, as required in Eq. (1).

IV. CONCLUSIONS

This chapter demonstrates not only that the GA method can solve problems in optical design and theoretical optics, but also that the flexibility of GAs

is great and their range of applicability is broad. After the initial work optimizing the GA software itself is completed, adapting the basic GA kernel to solve the problem at hand follows with ease. Future work on this method lies in decreasing its solution time, which now is on the scale of several hours for the problems considered. This can be accomplished through the creation of parallel versions of the code, as described above, or the implementation of new heuristic models in the GA driver, such as the Tabu search. In any case, these optimization methods provide a means to produce solutions unfettered by convention or human influence.

REFERENCES

1. M Gell-Mann. The Quark and the Jaguar. (New York: Freeman, 1994), Chapter 20, p. 312.
2. DE Goldberg. Genetic Algorithms in Search, Optimization, and Machine Learning. (Reading, MA: Addison-Wesley, 1989, Chapter 1, pp. 1–6.
3. BP Buckles, FE Petry. Genetic Algorithms. Los Alamitos, CA: IEE Computer Society Press, 1992, p. 1.
4. DE Goldberg. Genetic Algorithms in Search, Optimization, and Machine Learning. Reading, MA: Addison-Wesley, 1989, Chapter 1, pp. 7.
5. J Bengtsson. Kinoform-only Gaussian-to-rectangle beam shaper for a semiconductor laser. Appl Opt 35:3807–3814, 1996.
6. C Han, Y Ishii, K Murata. Reshaping collimated laser beams with Gaussian profile to uniform profiles. Appl Opt 22:3644–3647, 1983.
7. X Tan, B Gu, G Yang, B Dong. Diffractive phase elements for beam shaping: a new design method. Appl Opt 34:1314–1320, 1995.
8. FM Dickey, SC Holswade. Gaussian laser beam profile shaping. Opt Eng 35:3285-3295, 1996.
9. PW Rhodes, DL Shealy. Refractive optical systems for irradiance redistribution of collimated radiation: their design and analysis. Appl Opt 19:3545–3553, 1980.
10. D Shafer. Gaussian to flat-top intensity distributing lens. Opt Laser Technol. 14:159–160, 1982.
11. PH Malyak. Two-mirror unobscured optical system for reshaping the irradiance redistribution of a laser beam. Appl Opt 31:4377–4383, 1992.
12. NC Evans, DL Shealy. Design and optimization of an irradiance profile-shaping system with a genetic algorithm method. Appl Opt 37:5216–5221, 1998.
13. M Kuittinen, P Vahimaa, M Honkanen, J Turunen. Beam shaping in the nonparaxial domain of diffractive optics. Appl Opt 36:2034–2041, 1997.
14. W Jiang, DL Shealy, JC Martin. Design and testing of a refractive reshaping system. In: Current Developments in Optical Design and Optical Engineering III. E Fischer; WJ Smith, eds. Proc. SPIE 2000, 64–75, 1993.

15. C Wang, DL Shealy. Design of gradient-index lens systems for laser beam reshaping. Appl Opt 32:4763–4769, 1993.
16. E Betensky. Postmodern lens design. Opt Eng 32:1750–1756, 1993.
17. K Nemoto, T Nayuki, T Fujii, N Goto, Y Kanai. Optimum control of the laser beam intensity profile with a deformable mirror. Appl Opt 36:7689–7695, 1997.
18. JH McDermit, TE Horton. Reflective optics for obtaining prescribed irradiative distributions from collimated sources. Appl Opt 13:1444–1450, 1974.
19. V Oliker, L Prussner, DL Shealy, S Mirov. Optical design of a two-mirror symmetrical reshaping system and its application in superbroadband color center laser. In: Current Developments in Optical Design and Optical Engineering IV. RE Fischer; WJ Smith, eds., Proc. SPIE 2263:10–18, 1994.
20. T Dresel, M Beyerlein, J Schwider. Design of computer generated beam-shaping holograms by iterative finite-element mesh adaptation. Appl Opt 35:6865–6874, 1996.
21. M Born, E Wolf. Principles of Optics. 4th ed. New York: Pergamon Press, 1970, Chapter 3, pp. 115–116.
22. WH Press, SA Teukolsky, WT Vetterling, BP Flannery. Numerical Recipes in Fortran 77: The Art of Scientific Computing. Cambridge: Cambridge University Press, 1992, Chapter 10, pp. 387–388.
23. JE Dennis, RB Schnabel. Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Englewood Cliffs, NJ: Prentice-Hall, 1983.
24. RP Brent. Algorithms for Minimization without Derivatives. Englewood Cliffs, NJ: Prentice-Hall, 1983.
25. WH Press. Numerical Recipes in Fortran 77: The Art of Scientific Computing. Cambridge: Cambridge University Press, 1992, Chapter 10, pp 436–437.
26. GB Dantzig. Linear Programming and Extensions. (Princeton, NJ: University Press, 1963.
27. T Etzion, PRJ Ostergand. Greedy and heuristic algorithms for codes and colorings. IEEE Transactions on Information Theory 44:382–388.
28. RHJM Otten, LPPP van Ginneken. The Annealing Algorithm. Boston, MA: Kluwer, 1989.
29. RR Brooks, SS Iyengar, J Chen. Self-calibration of a noisy multiple-sensor system with genetic algorithms. In: A Wang, ed., Self-calibrated Intelligent Optical Sensors and Systems. Proc SPIE 2594:20–38, 1996.
30. H Kim, Y Hayashi, K Nara. An algorithm for thermal unit maintenance scheduling through combined use of GA, SA and TS. IEEE Trans Power Systems 12, 329–335, 1197.
31. FS Wen, CS Chang. Tabu search approach to alarm processing in power systems. IEEE Proceedings: Generations, Transmission and Distribution 144:31–38, 1997.
32. K Krishnakumar. Micro-genetic algorithms for stationary and non-stationary function optimization. SPIE: Intelligent Control and Adaptive Systems 1196, 1989.
33. DE Goldberg. The Quark and the Jaguar. New York: Freeman, 1994, Chapter 20, 21.

34. P Kraft, NR Harvey, S Marshall. Parallel genetic algorithms in the optimization of morphological filters. J Electronic Imaging 6:504–516, 1997.
35. AB Djuriši, JM Elazar, AD Raki. Simulated-annealing-based genetic algorithm for modeling the optical constants of solids. Appl Opt 36:7097–7103, 1997.
36. CH Still. Portable Parallel Computing via the MPI1 Message-Passing Standard. Computers in Physics 8:533–538, 1994.
37. DW Walker. The design of a standard message-passing interface for distributed memory concurrent computers. Parallel Computing 20:657–673, 1994.
38. KM Baker, DL Shealy, W Jiang. Directional light filters: three-dimensional azo dye formed micro-honeycomb images with optical resins. In: I Cindrich, SH Lee, eds, Diffractive and Holographic Optics Technology II. Proc SPIE 2404:144–158, 1995.
39. P Mouroulis, J Macdonald. Geometrical Optics and Optical Design. Oxford University Press, 1997, Chapter 9, pp. 248–250.
40. W Jiang. Application of a Laser Beam Profile Reshaper to Enhance Performance of Holographic Projection Systems. University of Alabama at Birmingham Dept. of Physics, Birmingham, 1993, Chapter 2, pp. 30–32.
41. WH Press. Numerical Recipes in Fortran 77: The Art of Scientific Computing. Cambridge: Cambridge University Press, 1992, Chapter 10, pp. 425–436.

APPENDIX: CODE SAMPLES

GA macro: ga.seq.1 (CODE V macro language)

```
 1 lcl num ^merit ^meritv(10)
 2 lcl num ^par(200,9) ^fun(200) ^iv
 3 lcl num ^r1(0..200) ^r2(0..200)^n(0..200)^srn(0..200)
 4 lcl num ^sum ^avg ^eprinc ^raymiss ^nsum
 5 lcl num ^i2p ^i2s(0..200)
 6 lcl num ^UDGC11 ^UDGC12 ^UDGC21 ^UDGC22
 7
 8 out n
 9 ver all n
10 exc n
11
12 ^eprinc==(epd)/400.
13
14 for ^iv -1 1000000
15
16 if (^iv=0)
17         ^iv == 2
18 end if
19
```

```
usr ^par ^fun ^iv
!out y
!wri (tim)
!wri ''__________''
!out n


for ^c 1 10

thi s2 ^par(^c,1)
thi s4 ^par(^c,2)
thi s3 ^par(^c,3)
rdy s3 ^par(^c,4)
rdy s5 ^par(^c,5)
^UDGC11 == roundf(^par(^c,6))
if (^par(^c,7)>0)
        ^UDGC12 == 1
els
        ^UDGC12 == -1
end if
^UDGC21 == roundf(^par(^c,8))
if (^par(^c,9) > 0)
        ^UDGC22 == 1
els
        ^UDGC22 == -1
end if

del prv all
prv
pwl 589
'n1' LPT GRADIUM 1.7
UDG
UDG C1 (^UDGC11)
UDG C2 (^UDGC12)
pwl 589
'n2' LPT GRADIUM 1.7
UDG
UDG C1 (^UDGC21)
UDG C2 (^UDGC22)
end
gla s2 'n1'
gla s4 'n2'
```

```
 62
 63 ^merit == 0
 64 ^sum == 0
 65 ^nsum == 1
 66
 67 for ^i 0 200
 68         ^r1(^i) == ^i*^eprinc
 69         ^j = raysin(0,0,0.,^r1(^i),0.,0.)
 70         if (^j <> 0)
 71                ^raymiss==1
 72         els
 73                ^r2(^i) == (y s6)
 74                ^r2(^i) == absf(^r2(^i))
 75                ^n(^i) == (n s6)
 76                ^nsum == ^nsum * ^n(^i)**6
 77                ^srn(^i) == (srn s6)
 78         end if
 79 end for
 80
 81 if (^raymiss = 0)
 82         ^a == -logf((PUI))/((PUX)*(EPD)/2.)**2
 83         for ^i 1 200
 84                ^j == ^i-1
 85                ^k == ^i
 86                ^i2p == (^r1(^i)*(^r1(^k)-^r1(^j))
                          *expf(-^a*^r1(^i)**2) *^n(^i))&
 87                        /(^r2(^i)*(^r2(^k)-^r2(^j)))
 88                ^i2s(^i) ==
 89 (^i2p*^srn(^i))/((^n(^i)*^srn(^i))
                   +(sinf(acosf(^n(^i) ))
                   *sinf(acosf (^srn(^i)))))
 90         !      ^i2s(^i)=^i2p*absf(^srn(^i))
 91                ^sum = ^sum +^i2s(^i)
 92         end for
 93         ^avg = ^sum/199
 94
 95         for ^i 1 200
 96                ^merit = ^merit+(^i2s(^i)-^avg)**2
 97         end for
 98
 99         ^merit = sqrtf(^merit/199)
100
```

```
101         ^meritv(^c) = 1/^merit
                          * expf(-1*((1 -^nsum)**2))
                          * expf(-0.01*((8 -
                          ^r2(200))**2))
102
103  els
104         ^raymiss == 0
105         ^meritv(^c) = 0
106  end if
107
108  ^fun(^c) == ^meritv(^c)
109
110
111  end for
112
113  end for
114
115  ver y
116
```

This appendix illustrates how CODE V can be used to evaluate the merit function for the Gradient-Index system. The above code is written in CODE V macro language. Not shown in this appendix is the GA driver, which is written in Fortran 77. See the footnote on p. 223 for more information on the GA driver. To access the GA driver in CODE V, the GA code was compiled into a library format specified in the CODE V documentation. See 'User-defined functions' in the CODE V manuals for more information. At line 2, the parameters are defined that provide a communications nexus between CODE V and the GA driver. The GA is invoked at line 20, where the parameters defined in line 2 are passed to the GA. The GA interprets the values stored in the "fun(200)" array, and uses those in its calculations. In turn, the GA populates the "^par(200,9)" array, which defines a new generation to be evaluated in CODE V. The "for..end" for loop starting at line 27 and ending at line 111 contains the code that traces the 200 rays for each of the 10 individuals per generation and, most importantly, evaluates the merit function for each. The raytrace is done in lines 67–79. The irradiance profile is calculated in lines 83–92. The precise form of the merit function can be found in lines 95–105. The first term, "1/^merit," is a value proportional to the flatness (uniformity) of the output profile. "^nsum" is a value related to the angle each ray traced through the system make with the optical (Z-) axis. As each of these angles approaches 0, "^nsum" approaches 1. If "^nsum" is one, then it makes no contribution to the merit function.

The last term, ''expf(-0.01*((8-^r2(200))**2))'' is less than one if the radial height of the marginal ray (''^r2(200)'') is some value other than eight, penalizing the system. Also note how the gradient glass type is chosen in lines 47-61. This is a unique feature of this particular gradient-index application. The command 'prv' at line 48 instructs CODE V to begin a private (user-defined) glass type. ''pwl 589'' at line 49 indicates that the private glass is valid for wavelength $\lambda = 589$ nm. ''UDG C1'' and ''UDG C2'' are the two discrete parameters that select the GRIN element from the several available in the CODE V gradient-index catalog. The values are populated from the variables ^UDGC11 and ^UDGC12, which are set by the GA, in lines 52 and 53. This same process is repeated to define a private glass type for the second shaping element in lines 54–58. Finally, the individual glass type for the shaping elements are set to the two newly defined private glasses in lines 60 and 61.

6

Beam Shaping with Diffractive Diffusers

David Renick Brown
MEMS Optical, LLC, Huntsville, Alabama

I. INTRODUCTION

In this chapter, we discuss an approach to beam shaping that often has a different realm of applications than the more conventional techniques. Specifically, we discuss what is called a band limited diffuser. Many diffusers, such as ground glass, diffuse light over an angular volume that is often larger and not as well defined as desired. We will see in this chapter that diffractive diffusers offer a technique to diffuse light over a very well-controlled angular spectral band.

In Sec. II, we describe the properties of diffusers by contrasting their characteristics with other perhaps more familiar optics. We first compare them to conventional single diffractive order beam shapers. We then compare them to gratings and develop the theory used to describe the diffuser.

In Sec. III, we give a very simple yet illustrative design example that covers a basic technique for the design of a binary diffuser.

Section IV is a discussion of fabrication techniques. In this section, we describe fabrication limitations and considerations. It is meant to give the reader an appreciation for what is possible and which fabrication method is appropriate for a given design.

In Sec. V, we describe the major negative aspect to this beam shaping technique, which is speckle. We derive the size of speckle and discuss a few methods for reducing speckle.

We conclude this chapter with a few of the possible applications for diffractive diffusers. Section VI also discusses when and when not to use diffusers for beam shaping.

II. PROPERTIES OF DIFFRACTIVE DIFFUSERS

To help define the properties of beam shaping with diffusers, it is useful to describe the differences between the other beam shaping techniques. There are two general categories of beam shapers, the first of which is a near field beam shaper.

A. Near Field Beam Shapers (Remapping Optics)

Near field diffractive optics generally use a single diffractive order to produce the desired optical effect. In general, a diffractive optic such as a grating can use many diffractive orders. A simple example of a diffractive optic that uses only one diffractive order is a lens. Near field beam shapers are much like a complex aberrated lens that performs a re-mapping of the beam's energy distribution to provide the desired shape. The shaped beam will exist only at some pre-defined plane unless a second optic is used to correct the phase in the beam as shown in Fig. 1. The resulting phase of a near field beam shaper can be canceled to produce a collimated beam which is allowed to diverge to give a shaped beam over an extended but finite range. The corrected shaped beam will experience diffraction and will degrade as the beam propagates. The diffraction of the corrected shaped beam will be as if the beam originated from an aperture function that is the same as the shaped beam. To minimize the diffraction of the edges, it is often advantageous to design the desired shape of the beam to have soft or smooth edges (1). The function that is used to describe the soft edge can have many forms. One such soft aperture function is a high-order Gaussian or super Gaussian of the following form:

$$I \propto e^{2(r/\omega)^{2N}} \tag{1}$$

Where I is the intensity, r is the radius, ω is the waist radius and N is an integer. As the value of N increases, the closer the function approximates a true top hat function.

It is possible to extend the range of a near field beam shaper by optically taking the Fourier transform (2) of the output as shown in Fig. 2. The lens transforms the shaped beam into its Fourier transform at the back focal plain of the lens. As the field propagates beyond the back focal

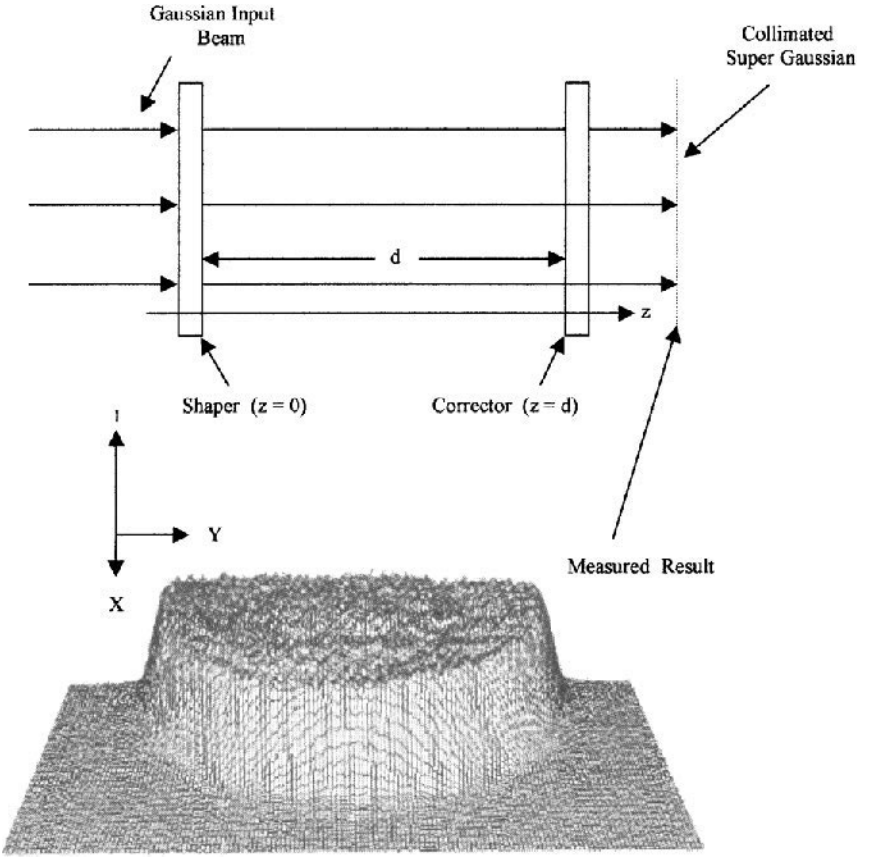


Figure 1 Typical system layout of a near field beam shaper. Only the first optic is required to shape the beam at plane $z = d$. To extend the range at which the top hat will exist requires a second optic to correct for aberrations in the phase of the beam. The empirical result shown has a non-uniformity standard deviation of 5.6%.

plain, the diffraction caused by the propagation transforms the field back into the shaped beam with a spherically diverging phase. This creates a diverging cone of light whose intensity envelope has the desired shape. Experimental results of this set-up for a round super Gaussian are shown in Fig. 2. The structure that is observed in the measured result is caused by multiple reflections within the system due to optics that do not have anti-reflection coatings. The beam can then be collimated at any point by selecting the appropriate lens. This also allows one to size the output beam.

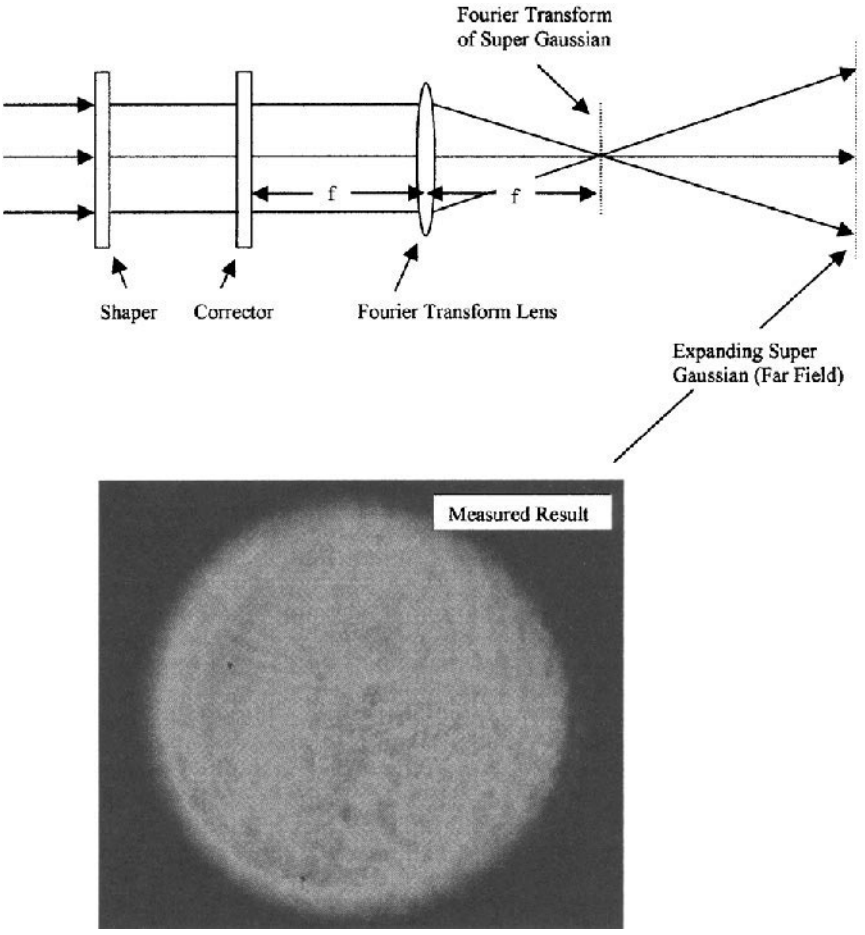
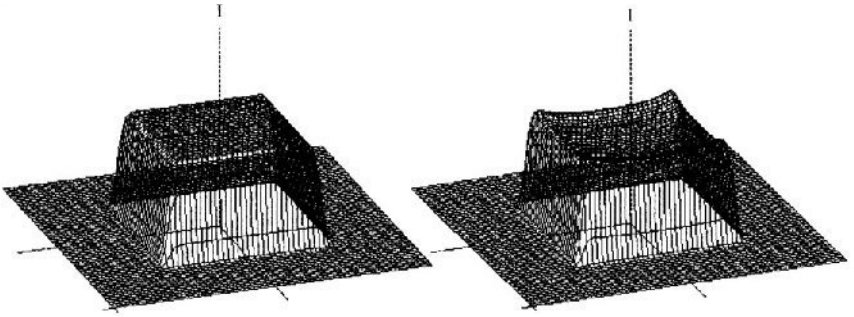


Figure 2 System to transfer the shaped beam into the far field. A simple lens with the appropriate focal length can be added after this system to re-collimate the beam. The measured results shown was taken approximately 300 mm beyond the Fourier transform lens and is approximately 1 cm in diameter. The structure in the beam is from multiple reflections within the system.

Due to the re-mapping nature of a near field beam shaper, the output is highly sensitive to the intensity and phase of the input beam. Any deviation in the input beam size, shape or location relative to the near field beam shaper will cause degradation to the resulting output. Figure 3 shows the output intensity of a simulation of a Gaussian to square top hat beam shaper. There are several methods for designing the beam shaping diffractive



(a) (b)

Figure 3 Simulation results of the output intensity of a Gaussian to square super Gaussian top hat beam shaper. (a) The result with the perfect input beam. The output is >99% efficient and the peak-to-valley non-uniformity is <2%. (b) The result with an input beam that is 5% too large. The peak-to-peak non-uniformity is ~18%.

optic (2, 3). Commercially available ray tracing computer codes, such as ZEMAX and Code V, can be used to design the beam shaping optic if one is careful to include diffraction effects. The results in Fig. 3 were simulated using a scalar wave propagation computer code which accurately models diffraction effects. Figure 3a is a plot of the output intensity with a perfect input beam. The peak-to-valley non-uniformity of the intensity of this top hat is less than 2% and the simulated diffraction efficiency is better than 99%. Figure 3b shows the output of the same optic with the input beam 5% larger than the designed beam. The peak-to-valley non-uniformity of the intensity is now about 18%. From this, we see the sensitivity of the near field beam shaper to input beam variations. In general, the desired intensity footprint is maintained over a fairly large range of variations in input beam. The uniformity of the output intensity is however very sensitive to the input beam. However, with care extremely good results can be obtained. In Fig. 4, experimental results for a UV beam shaper for a lithography application are shown. The non-uniformity (σ/μ) was measured at less than 3% where σ and μ are the standard deviation and mean of the intensity respectively.

B. Far Field Beam Shapers (Gratings, Diffusers)

A diffuser falls into the second category of beam shapers, which is far field beam shapers. Far field diffractive optics such as gratings and diffusers can

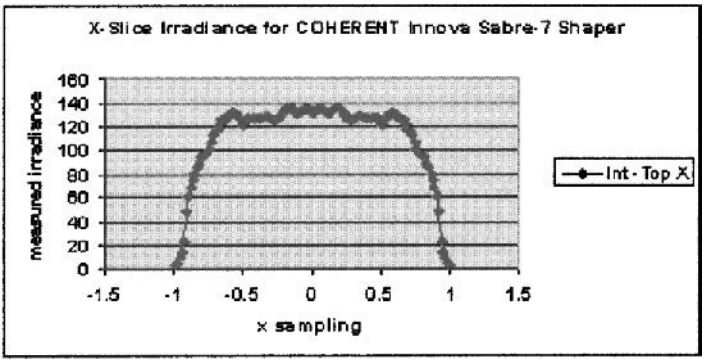


Figure 4 Experimental results of a beam shaper for a Coherent Innova Sabre-7 UV laser (363.8 nm). Non-uniformity (σ/μ) was measured to be less than 3%.

in general use many diffractive orders in the case of gratings or spatial frequencies in the case of diffusers. Far field diffractive optics imparts a defined spatial frequency distribution to the phase of the laser beam. As the beam propagates, the spatial frequencies in the phase cause the beam to interfere with itself. Since the structure of such a device is made up of many very small phase apertures (typical < 10 wavelengths), the beam is in the far field almost immediately beyond the optic. This means that the resulting shape of the beam will continue to propagate with the pre-defined angular divergence as defined by the spatial frequencies in the phase. Figure 5

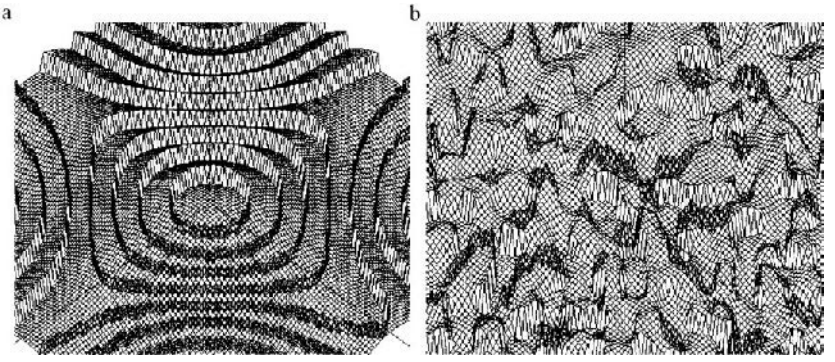


Figure 5 (a) The phase of the Gaussian to square super Gaussian top hat beam shaper shown in Fig. 3. (b) A portion of the phase of a diffuser that projects a square energy envelope.

illustrates the extreme differences in the phase of a near field (single order) optic and a far field (multiple order) optic. Figure 5a is the phase applied to the input field to generate the shaped beam shown in Fig. 3. Figure 5b is the phase of a diffuser that projects a square pattern.

Far field optics have the advantage of being relatively insensitive to the shape, size, and alignment of the input beam. An input beam that is a TEM00 mode will produce a very similar output to an input beam that is a TEM01 mode (4, 5). This is due to the multiplicative property of a Fourier transform. The resulting beam of a far field optic is simply the convolution of the Fourier transform of input beam and the spatial frequencies of the optic. As we will see later, the energy envelope of the output pattern is dominated by the phase function of the diffuser and not the shape of the input beam.

Since gratings and diffusers are both far field diffractive optics they share many characteristics. It is useful to describe a diffuser in terms of a grating due to its familiarity with most readers. In general, a grating is a periodic amplitude and/or phase structure. For the purposes of this discussion, we will limit a grating to a phase only structure.

By starting with the differential form of Maxwell's equations, and making simplifying assumptions for a homogeneous medium, one can arrive at the homogeneous wave equation for the electric field (6):

$$\nabla^2 \mathbf{E} = \mu\varepsilon \frac{\partial^2 \mathbf{E}}{\partial t^2} \tag{2}$$

where $\mathbf{E}$ is the electric field vector, μ and ε are the material property parameters called the permeability and permittivity respectively. A similar equation exist for the magnetic field. This vector equation can be separated into three scalar equations, one scalar equation for each component of the coordinate system. Using Cartesian coordinates and choosing the scalar equation dependent on the z spatial coordinate we have

$$\frac{\partial^2 E}{\partial t^2} - \frac{1}{\mu\varepsilon} \frac{\partial^2 E}{\partial z^2} = 0 \tag{3}$$

The general form of the solution of equation (3) is (6)

$$E(z,t) = Af_+(\omega_0 t - k_{0z} z) + Bf_-(\omega_0 t + k_{0z} z) \tag{4}$$

where A and B are constants, ω_0 is the angular frequency with the units of radians/time and k_{0z} is called the propagation constant with the units of radians/length. Equation (4) is the solution of the scalar equation (3) if

$$\frac{\omega_0}{k_0} \equiv \frac{1}{\sqrt{\mu\varepsilon}} = \upsilon \tag{5}$$

where v is the velocity of the light in the medium. The two terms on the right side of Eq. (4) describe two waves, one traveling in the positive z direction and one traveling in the negative z direction. In general, the wave can travel in any direction. The argument of the first term in Eq. (4) can be written more generally as $\omega_0 t - \mathbf{k}_0 \cdot \mathbf{r}$ where in Cartesian coordinates

$$\mathbf{k}_0 = k_{0x}\hat{x} + k_{0y}\hat{y} + k_{0z}\hat{z} \tag{6}$$

$$\mathbf{r} = x\hat{x} + y\hat{y} + z\hat{z} \tag{7}$$

Equation (4) can be rewritten as

$$E(x, y, z, t) = Af(\omega_0 t - \mathbf{k}_0 \cdot \mathbf{r}) \tag{8}$$

Equation (8) is a wave with amplitude A and velocity v traveling in the $\mathbf{k}_0$ direction. $\mathbf{k}_0$ is called the propagation vector or the wave vector. The direction of the power flow density of a field is equal to the direction of the propagation vector (4). The magnitude of the propagation vector is given by (4,7)

$$|\mathbf{k}_0| = k_0 = \frac{\omega_0}{v} = \frac{2\pi}{\lambda} \tag{9}$$

where λ is the wavelength of the light in a given material. k_0 is a constant while the light is propagating in the material. A wave described by Eq. (8) is often referred to as a plane wave. An arbitrary complex electromagnetic field can be analyzed in terms of its Fourier components. The Fourier components of a complex field are simply a series of plane waves traveling in different directions (5). When analyzing periodic structures such as gratings, it is often advantageous to perform the analysis in the Fourier domain.

Figure 6 shows a circle whose radius is k_0. Along the k_x axis is a periodic structure with a grating vector of $\mathbf{K}_g$ whose magnitude is given by

$$k_g = \frac{2\pi}{\Lambda} \tag{10}$$

where Λ is the period of the grating. $\mathbf{K}_g$ has only an x component and adds to the x component of the propagating wave in discrete multiples. A graphical illustration of this is shown in Fig. 6. Due to the constraint that the wave vector has a constant magnitude of k_0 we see from Fig. 6 that we can graphically determine the direction of the series of plane waves that result from the grating

$$\sin(\theta_t^m) = \frac{(k_{0x} + mk_g)}{k_0} \tag{11}$$

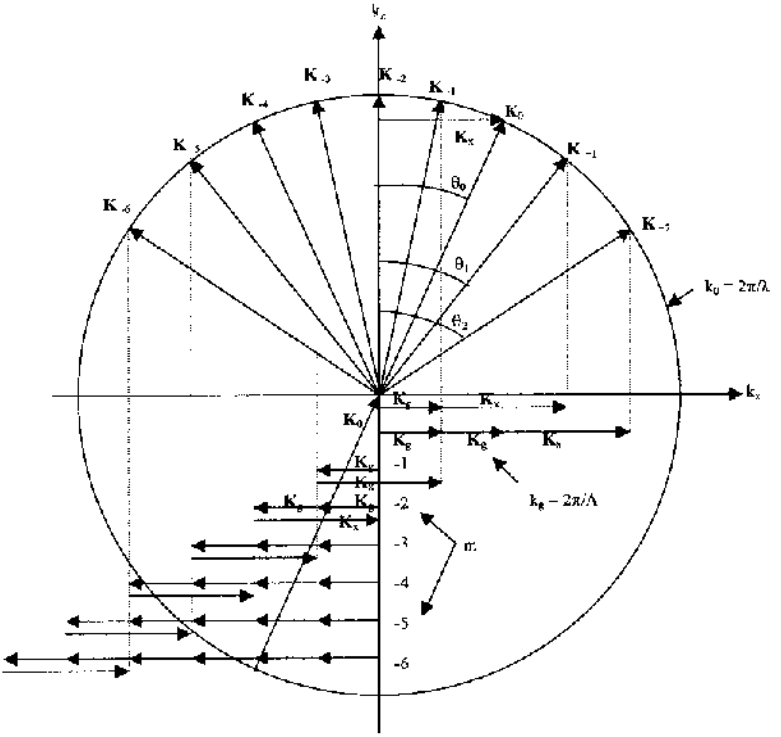


Figure 6 The wave vector map of light as it is transmits through a periodically varying structure such as a grating. The lower portion of the plot shows the x component summations of the undeviated beam ($\mathbf{K}_0$) and the grating vector ($\mathbf{K}_g$). The orders of the grating are the result of an integer number of grating vectors added (or subtracted) to K_0.

which then reduces to the familiar grating equation

$$\sin(\theta_t^m) = \sin(\theta_i) + \frac{m\lambda}{\Lambda} \tag{12}$$

Where θ_i is the incidence angle and θ_t^m is the transmitted angle of a given diffracted order m. Orders (values of m) that require $|\sin(\theta_t^m)| > 1$ are called evanescent orders. The wave vector of an evanescent order has an imaginary z component and thus attenuates exponentially beyond the surface of the grating (7).

For the "real" orders, the resulting electric field at $z = 0$ is of the form

$$\mathbf{E}_g(x, z = 0) = \mathbf{A}(x, 0)e^{j(2\pi/\lambda)(n-1)g(x)} = \mathbf{A}(x, 0)P(x) \tag{13}$$

where n is the index of refraction of the material, $\mathbf{A}(x,0)$ is the amplitude of the input beam, and $g(x)$ is a periodic phase function whose height is $\lambda/(n-1)$. In the far field ($z = z'$) Eq. (13) becomes

$$\begin{aligned}\mathbf{E}_g(k_x, z') &= \mathfrak{I}[\mathbf{E}_g(x,0)] \\ &= \mathfrak{I}[\mathbf{A}(x,0)] * \mathfrak{I}[P(x)] = \mathbf{A}'(k_x,0) * \mathfrak{I}[P(x)]\end{aligned} \tag{14}$$

where $\mathfrak{I}$ stands for Fourier transform, $*$ is the convolution symbol and $k_x = k_0 x'/z'$. From Eq. (14), we see that the resulting field at $z = z'$ is simply the convolution of the spatial frequency content of the phase and the amplitude of the input beam after propagating a distance of z'. If the divergence of the diffuser (high spatial frequency) is significant, then the shape of the energy envelope will be dominated by the phase of the diffuser, rather than the divergence of the input beam.

A grating has very distinct orders due to its periodic structure. The spatial frequency of the phase is simply delta functions with the appropriate weighting factors spaced at angular intervals as defined in Eq. (12). If we now add a second function $r(x)$ with a period much larger than the period of $g(x)(\Lambda_r \gg \Lambda_g)$) to the phase in Eq. (13). We see that

$$\mathbf{E}_d(x, z=0) = \mathbf{A}(x,0)e^{j(2\pi/\lambda)(n-1)[g(x)+r(x)]} = \mathbf{A}(x,0)P(x)R(x) \tag{15}$$

The resulting field after propagating a distance z' then becomes

$$\mathbf{E}_d(k_x, z') = \mathfrak{I}[\mathbf{E}_d(x,0)] = \mathbf{A}'(k_x,0) * \mathfrak{I}[P(x)] * \mathfrak{I}[R(x)] \tag{16}$$

When a second component is added to the phase, the distinct orders become blurred by the spatial frequency components of the function $r(x)$. By choosing Λ_r and Λ_g appropriately, an apparent continuum of spatial frequencies may be obtained, resulting in a "solid" filled region of light in the far field.

Notice that since $\Lambda_r \gg \Lambda_g$ the period of the diffuser Λ_d is approximately equal to Λ_r. Thus, if a grating has a large period such that the orders of that grating, as governed by Eq. (12), are spaced in such a way that the resulting beams significantly overlap, the angular region will be "solidly" filled with light. Any two coherent beams that overlap will interfere. This interference is the source of the speckle indicative in diffuser patterns. The subject of speckle will be covered in some detail later in this chapter.

Also notice the case that as Λ_r becomes so large that the size of the input beam is no longer large enough to sample one period of the phase then the phase is no longer periodic. At this point, the location of the orders is ill defined since optically the phase is not periodic. This is a result of the fact that non-periodic functions have a continuum of frequencies rather than a discrete set of frequencies. However, the phase still has the same spatial frequency spectrum and thus the envelope of the pattern will remain

virtually unchanged. This is of course much like a traditional hologram in that a small piece of a hologram will still produce the same image. We will see later that as the input beam decreases in size, the speckle pattern becomes increasingly coarse.

C. Mathematical Description of a Diffuser

To mathematically describe a diffuser, we first note the shift property of the Fourier transform:

$$A(k_x - k_{x0}) \Leftrightarrow a(x)e^{-\mathrm{j}k_{x0}x} = a(x)e^{-\mathrm{j}(2\pi/\Lambda)x} \tag{17}$$

When designing optics such as diffusers, it is often useful to define things in term of a Discrete Fourier Transform. For a calculation grid of dimension D, the smallest frequency increment is $\delta f = 1/D$ (8). Physically, D is the diameter of the input beam or the period of a grating. Thus, any frequency is an integer multiple of δf. For example, define a frequency f_0

$$f_0 = N\delta f = \frac{N}{D} \tag{18}$$

or in terms of the wave number

$$k_{x0} = 2\pi f_0 = 2\pi\frac{N}{D} \tag{19}$$

where N is an integer. From Eq. (17) and (19) it then follows that

$$\Lambda = \frac{N}{D} \tag{20}$$

Substituting Eq. (20) into Eq. (12), we find that any discrete spatial frequency can be described as

$$\sin(\theta_t) = \sin(\theta_i) + \frac{N\lambda}{D} \tag{21}$$

Recognizing the fact that $D = \delta d(M-1)$ where δd is the smallest distance increment and M is the number of data points across the calculation grid, Eq. (21) becomes

$$\sin(\theta_t) = \sin(\theta_i) + \frac{N\lambda}{\delta d(M-1)} \tag{22}$$

Finally, solving Eq. (22) for N we have

$$N = [\sin(\theta_t) - \sin(\theta_i)]\,\frac{\delta d(M-1)}{\lambda} \tag{23}$$

This equation is useful for computational reasons to calculate a particular grid point number on a discrete Fourier grid to produce a phase function of a given dimension that will bend light of a wavelength λ by an angle θ_t.

III. SIMPLE DESIGN EXAMPLE

As a simple illustrative design example, suppose we wish to design a binary diffuser with a clear aperture of $D = 1.0\,\text{mm}$ that projects a ring of laser light with a wavelength of $\lambda = 0.6328\,\mu\text{m}$ between 1° and 2°. In this simple design approach we perform the following functions:

1. Define system/calculation variables.
2. Calculate spectral limits.
3. Define ring on a Fourier grid.
4. Randomize grid.
5. Take inverse FFT of grid.
6. Truncate phase to form binary (two-phase level) optic.
7. Verify performance with a scalar wave propagation simulation.

First we define our system variables. Let

$$\lambda = 0.6328\,\mu\text{m}$$

$$D = 1.0\,\text{mm}$$

$$M = 512$$

$$1.0^\circ < \theta_t < 2.0^\circ$$

$$\theta_i = 0.0^\circ$$

From Eq. (23) $N_1 = N(\theta = 1^\circ) = 28$ and $N_2 = N(\theta = 2^\circ) = 55$. Notice that we have to round to the nearest integer. From this information, we now prepare a grid of 512 by 512 points with a width and height of 1.0 mm that is zero everywhere except grid points whose radius falls between 28 and 55.

$$28 < R < 55$$

$$R = \sqrt{i^2 + j^2}$$

Where i and j are grid indices which have the range $-257 < i, j < 256$. Figure 7 shows the result.

The next step is to randomize the non-zero values of the amplitude between 0 and 1, and randomize the phase between 0 and 2π. The randomization step reduces the output dependence on the input beam. Effectively,

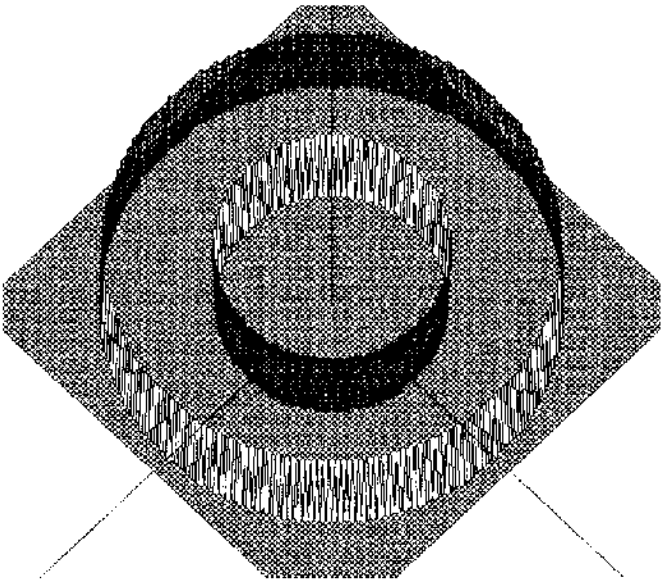


Figure 7 A plot of the spatial frequency band that is desired in the ring diffuser design example. The plot represents a subsection of the 512×512 calculation grid.

the high-frequency random function being multiplied by the desired frequency envelope insures that the spectral content of the envelope function is distributed over the full area of the binary diffuser. This removes any input beam alignment tolerances and any input beam intensity profile requirements. It is convenient at this point to insure that the complex grid is conjugate-symmetric; that is

$$f(i, j) = f^*(-i, -j) \tag{24}$$

Once this is done, the inverse FFT of the complex grid is calculated. Due to the conjugate-symmetry the result is real. All of the desired spatial frequency information is contained in the real part of the complex FFT. The real component of the grid is used as the phase for our optic, which contains the desired spatial frequency information. To reduce the phase to a binary diffractive optic, the phase is truncated to only two-phase levels, 0 and π. Assigning any positive value to π and any negative value to 0, will do this simply. Figure 8 shows the resulting phase profile. This simple design process is summarized in Fig. 9.

It might be noted that the process of using only the phase or even binary versions of the phase to reconstruct a function has a long history in

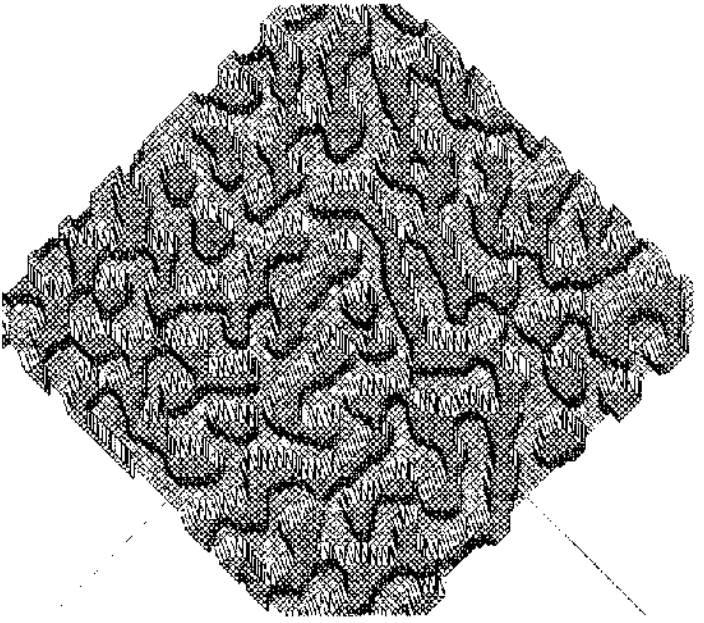


Figure 8 A portion of the binary (two levels) phase structure for the ring diffuser.

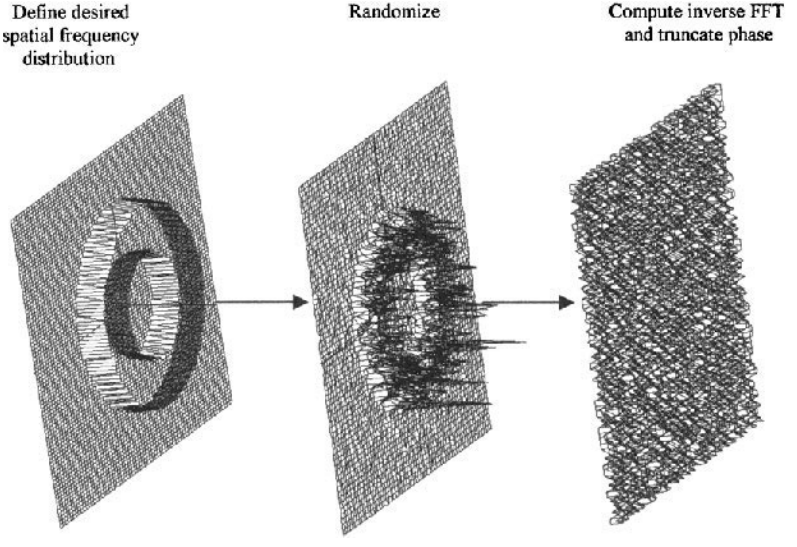


Figure 9 The steps for a simple diffuser design approach.

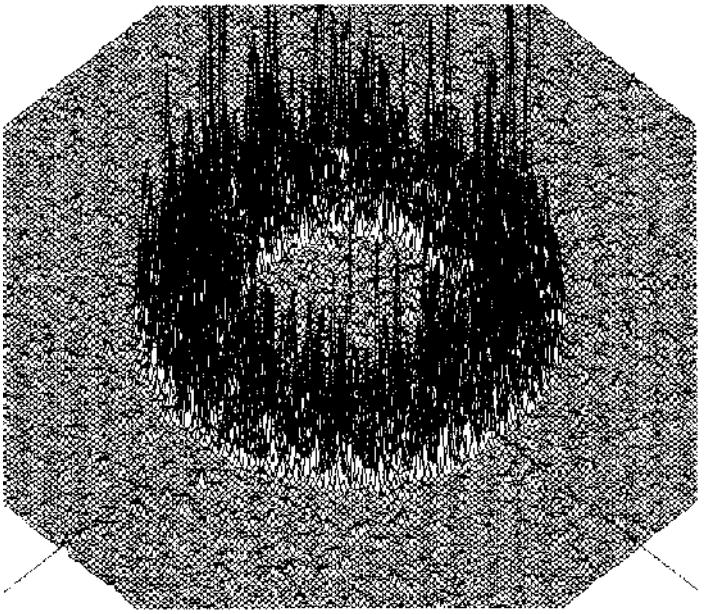


Figure 10 Simulated intensity plot of the ring diffuser. This simulation used a Gaussian input beam of diameter 0.5 mm. A spherical phase curvature was then applied to simulate a lens with a focal length of 10.0 mm. Using scalar wave theory the field was propagated 10.0 mm to the focal plane of the lens.

the image processing and the optical pattern recognition fields. The significance of the phase of the image Fourier transform in reconstructing the image is discussed by Stark (9). It has been demonstrated that the Fourier transform phase can provide a good reconstruction of an image. This is especially true for images with a lot of high-frequency content such as edges. Image reconstruction from the Fourier transform phase can be viewed as a high-frequency enhanced filtering process. This is certainly the case of the randomized function with sharp boundaries in the example just discussed. Flannery and Horner (10) discuss the applications of phase-only and binary phase-only filters to Fourier optical signal processing and pattern recognition.

This example was for a symmetric pattern. For a non-symmetric pattern the conjugate-symmetric condition is not possible, which results in a complex field following the inverse Fourier transform. One way to design a more complicated pattern that can still be generated with a binary diffuser is to define the center of the desired pattern far enough off-axis so that it does not overlap its symmetric counterpart. One way to produce a truly non-

symmetric pattern is to simply use the phase of the resulting inverse Fourier transform as the phase for the diffractive diffuser.

To verify our design we first assign an arbitrary amplitude function: Gaussian with $\omega_0 = 0.25\,\text{mm}$. We then simulate an optical Fourier transform by first applying a spherical phase to the field which simulates a lens of a given focal length and then propagating the field to the focal plane of the lens using a scalar theory code. The result is shown in Fig. 10. Notice the high-frequency content within the ring. This is the speckle characteristic of diffusers. Also notice the excellent suppression outside of the ring area. We know that at the focal plane $r = f\tan(\theta)$ where r is the radius from the optical axis, f is the focal length of the lens and θ is the angle of the incoming ray. From this, we can verify that the divergence angles of the design are indeed correct.

These binary diffusers generally have an efficiency of about 70%. More complicated techniques can be used to obtain higher efficiency (~90%) non-symmetric patterns with multi-level or continuous phase profiles. However, fabrication limitation may determine the level of complexity of the device.

IV. FABRICATION CONSIDERATIONS

There are many ways to transfer a designed diffractive optic into the actual physical diffractive optic. To mention a few: binary mask(s) photolithography followed by an etch(es) (11), gray scale mask photolithography followed by an etch, e-beam lithography, focused ion-beam milling, plastic molding, embossing and so on. The limiting factor in the fabricating of a diffractive optic is the level of accuracy that the process can reproduce the designed structure. This is generally limited to the resolution of the process. This is usually defined by δd in Eq. (23). This is the smallest feature in the grid. In our design example, we had a very easy resolution requirement of $\delta d \sim 2\,\mu\text{m}$.

There is a fabrication factor that is missing in Eqs. (13)–(16), and that is the pixelization effect in the phase. The very small squares in the phase will manifest themselves in the far field as a large $\text{sinc}^2(x, y)$ function envelope. The desired pattern (in this case, a top hat function) will be present in the center of the main lobe and repeated at the null points of the $\text{sinc}^2(x, y)$ function as shown in Fig. 11. The location of the null points are at $\sin(\theta_m) = m\lambda/\delta d$. This problem can become very significant if desired diffuse pattern contains angles that are large enough to approach these null points. From Eq. (23) we see that this occurs when

$$N \sim \frac{M}{2} \tag{25}$$

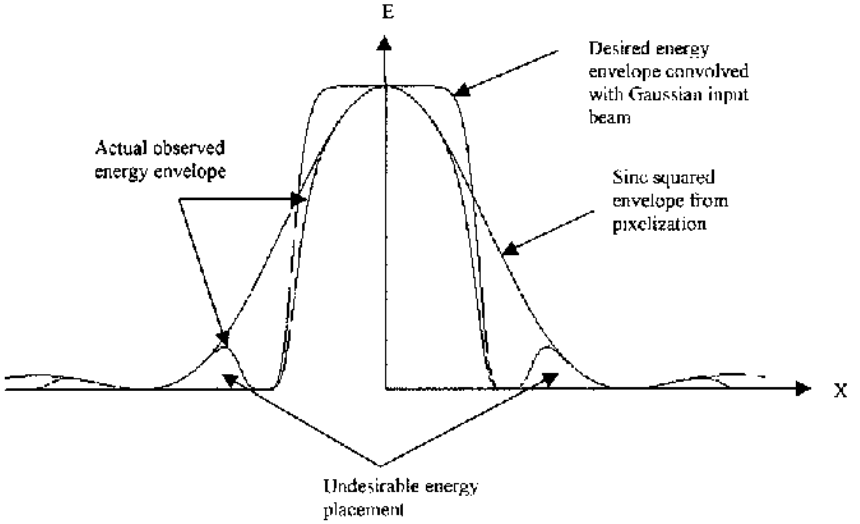
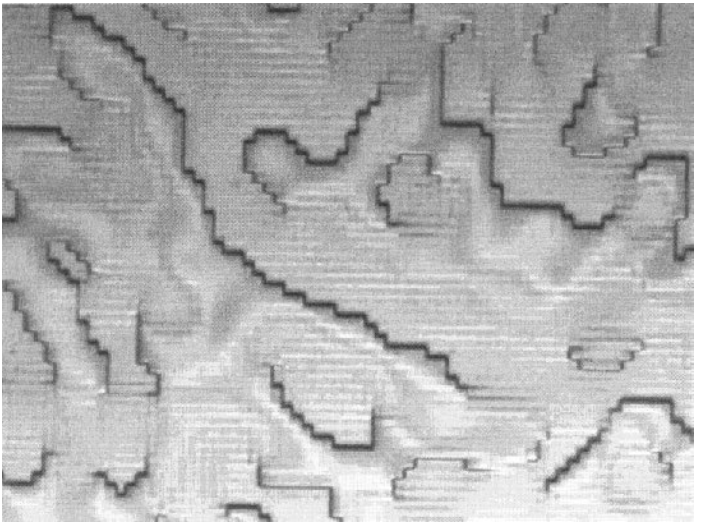


Figure 11 An illustrative energy envelope plot of the output of a diffuser and the effects of a pixelated phase.

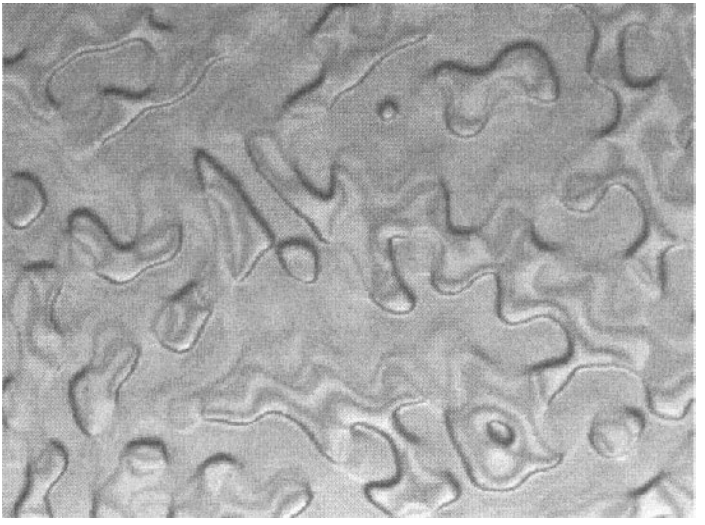
In the example shown in Fig. 11, N is 50 and M is 64. It is therefore best if $N \ll M/2$. For a given divergence angle, decreasing δd is a way to decrease N. However, the fabrication method of choice may not be able to support such a reduction in δd. Also notice that as $\delta d \to 0$ the $\text{sinc}^2(x, y)$ envelope becomes very large and thus less of a factor. In single step lithography, such as gray scale or direct write methods, the resolution limits actually help minimize the $\text{sinc}^2(x, y)$ envelope effect by poorly representing the pixelated nature of the design grid producing more of a smooth continuous phase. This is often the phase that a coarse calculation grid is trying to approximate.

In multi-step lithography processes, such as a binary multi-mask fabrication sequence, the small δd (sub-micron) can become difficult due to resolution limits and layer-to-layer registration errors. These errors usually manifest themselves as very small peaks and/or valleys in the phase creating potentially a very large number of scattering sites.

The phase structures of two diffusers are shown in Fig. 12. In Fig. 12a $\delta d \sim 2.0\,\mu\text{m}$ and in Fig. 12b $\delta d \sim 0.33\,\mu\text{m}$. Both of these diffusers where fabricated with gray scale mask photolithography. In this case, the phase structures have 64 phase levels and are only in a photo-resist layer on a fused silica substrate. In the case of Fig. 12b, one can see that the final result is not pixelated and δd is only a consideration for the design grid. To make them



(a)



(b)

Figure 12 Pictures of a portion of the phase structures of two diffusers with the same desired output. Both phase functions have 64 phase levels and are in a photoresist layer on a fused silica substrate. (a) A pixelated version with a pixel size of $\sim 2.0\,\mu m$. (b) A non-pixelated version. The alternating dark and light contour fringes are the results of thin film interference from the illuminating source.

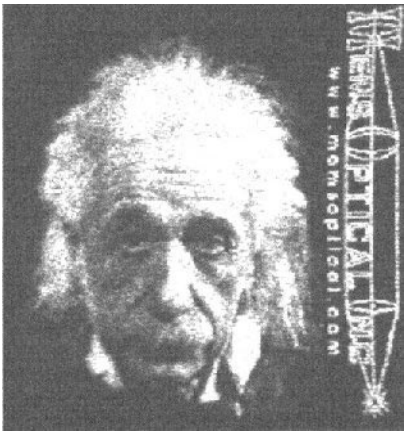


Figure 13 Empirical results of a 632.8 nm laser illuminating the phase of the diffuser pictured in Fig. 11b. This demonstrates the complexity and clarity that is possible with complex diffusers. Note that some of the fine detail present in the actual output was not faithfully transferred to this figure. For instance, individual hairs on the very top of his head are distinguishable in the live presentation.

more robust, a reactive ion etch would be used to transfer the surface relief structure into the fused silica. Shown in Fig. 13 is the experimental measurement of the output of the diffuser whose phase is shown in Fig. 12b. The sinc squared envelope indicative in a pixelated phase is not present in the output from the phase shown in Fig. 12b since the pixel structures are not represented in the phase. This example also demonstrates that the complexity that is available in the desired pattern is almost limitless provided that the fabrication method is appropriate in turning the design into reality.

It is interesting to note that the diffuser shown in Fig. 12b was also tested in white light and worked fairly well. The face was clear but the letters were blurred. From Eq. (22), the divergence angle is dependent on the wavelength of light (λ). This causes the pattern to be chromatically blurred. The separation of colors is most evident in the letters which are the farthest points from the optical axis. A blue-green dot at the zero order (optical axis) was present due to the lower diffraction efficiency of the off wavelengths (11).

V. SPECKLE

The largest drawback of a diffuser is the presence of speckle in the output pattern. Speckle is the high-frequency modulation of the intensity within the

desired energy envelope. As alluded to earlier, the origin of speckle in the output of a diffuser is from the overlapping nature of a coherent beam. Speckle is simply the result of interference.

A. Size of Speckle

The average size of the speckle is of great interest since some applications can tolerate speckle if it is small enough. From Eq. (21), we see that the smallest angular increment is given by (12)

$$\delta\theta \sim \frac{\lambda}{D} \tag{26}$$

Here we are assuming that the input beam is the same size as the calculation grid. If the phase is not periodic within the illumination area of the input beam the D in Eq. (26) is simply the diameter of the input beam. For our design example, $\delta\theta = 0.63\,\text{mrad}$. At a distance of 1 m the average speckle size would then be $\sim 0.63\,\text{mm}$.

At the focal plane of a lens the size of the speckle is (13)

$$\delta r \sim f\,\delta\theta \sim f\,\frac{\lambda}{D} = f/\#\,\lambda \tag{27}$$

which is roughly the size of the focused spot if the diffuser were not present. For an $f/2$ system, the average speckle size in our example is $\sim 1.2\,\mu\text{m}$. If the resolution limit of the system being illuminated, such as a detector array or a material processing application, is much greater than the speckle size, then there will be multiple speckle lobes integrated within the resolution area. This can significantly decrease the speckle effect by integrating the energy over the larger area defined by the resolution limit of the system. From Eq. (27), we see then that reducing the focused spot size of the non-diffused system will also decrease the speckle size when the diffuser is added. This fact also agrees with Eqs. (11)–(16) since the amplitude $A(k_x, z')$ is simply being convolved with the randomized orders of the diffuser. Also notice that the frequency cut-off point is much sharper at the focus of a lens than in a free space propagation. This is due to the fact that the amplitude function A has a smaller diameter at focus. Thus we see that if the application will allow the diffuse pattern to exist only on a particular plane along the optical axis, the focal plane of a lens generally gives the best results.

The number of speckle lobes within the desired pattern can be seen from Eq. (21). For a solid pattern, such as a square or a circle, the number of speckle lobes across the diameter or along a side is simply $\sim 2N$. In our example, there are $\sim 55 - 28 = 27$ speckle lobes across the ring's width.

B. Speckle Reduction

The reduction of the amount of speckle modulation is accomplished by three basic techniques (14). The first is to illuminate the diffuser with temporally or spatially partially coherent light. Since speckle is caused by coherent interference, then it stands to reason that by reducing the coherence of the beam will also reduce the contrast within the speckle pattern. Contrast is defined as (13,14)

$$C = \frac{\sigma_I}{\langle I \rangle} \tag{28}$$

Where σ_I is the variance of the intensity and $\langle I \rangle$ is the mean value of the intensity. For coherent light the contrast is equal to unity. The contrast of M mutually incoherent speckle patterns superimposed with the same wavelength and equal average intensities then becomes (13,14)

$$C = M^{-1/2} \tag{29}$$

From Eq. (29), we see that a laser with many incoherent modes will have lower contrast speckle than a single-mode coherent laser.

The second technique to reduce speckle is to time-average the speckle. This involves physically moving the beam or the diffuser very quickly and integrating the speckle pattern over a short period of time. This can work quite well for applications where the sensor, such as the human eye, a CCD array or photographic film, have a finite exposure time. The final technique is to spatial filter the speckle by observing the pattern through a finite aperture.

The subject of speckle is usually described in terms of the statistics of the speckle within the pattern. The details of the statistics of the speckle pattern is beyond the scope of this book. Excellent sources for further reading on the subject are in Refs. 13 and 14.

VI. APPLICATIONS OF DIFFUSERS

The applications for diffuser beam shapers as opposed to near field beam shapers depend on the system limitation and the application requirements. Diffusers should be used in applications were the input beam quality and/or system alignment capabilities are not sufficient for a near field beam shaper. Factors that affect the input beam quality are the ability to measure the intensity and the phase of the beam, the stability of the beam with time, and the consistency of the beam from laser to laser. Near field beam shapers can only shape one mode of the laser. Other modes that may be present will be

background noise. Diffusers will shape all modes of the laser and an increased amount of modes will lower the contrast of the speckle.

When designing a near field beam shaper for many different lasers, a series of measurements on many lasers must to be taken to evaluate the variance of the input beams to determine if the variance in the resulting output is acceptable. It is prudent to design the optical system with a spatial filter and methods to adjust the beam to better match the designed input beam. Alignment to the beam shaper is critical. Depending on the specifics of the beam shaper, the errors in alignment often have a multiplicative effect on the errors observed in the output. For example, a 2% translational misalignment in the beam shaper may result in a 10% tilt in the top hat.

Diffusers should not be used in applications where the speckle is not acceptable and can not be reduced to a tolerable level. They are quite effective for high-power applications where the laser that is being used has a large number of mutually incoherent modes. They should also not be used when collimation of the shaped pattern is a requirement.

To mention just a few specific applications, diffusers are used to homogenize a light source, even a broad band light source. It should be noted that diffractive diffusers function with white light even though they are dispersive. They are used to illuminate a specific region for scanning applications. They are also used in alignment applications where a specific pattern is desired, such as laser targeting systems for firearms, machine tooling and assembly alignment systems, and even for space station to shuttle docking alignment. A grid diffuser pattern can be used to map the topography of a region. They can be used with illuminators for night vision systems, product marking systems, pen pointers with corporate logos or sports team logos, laser light shows or increasing the viewing angle of a display. Diffusers could even be used in a wireless, free space inter-office communication network to reduce alignment requirements between components.

REFERENCES

1. AE Siegman. Lasers. Mill Valley, CA: University Science Books, Chapter 18, pp. 736–739.
2. G Borek, DR Brown, D Brown, R Clark. High performance diffractive optics for beam shaping. SPIE Proceedings 3633, San Jose, CA, January 1999.
3. D Brown, A Kathman. Multi-element diffractive optical designs using evolutionary programming. SPIE Proceedings 2383, San Jose, CA, February 1995.
4. HA Haus. Waves and Fields in Optoelectronics. Englewood Cliffs, NJ: Prentice Hall, 1984.
5. JW Goodman. Introduction to Fourier Optics. New York: McGraw-Hill, 1968.

6. PP Banerjee, TC Poon. Principles of Applied Optics. Boston: Aksen Associates, 1991.
7. EG Loewen, E Popov. Diffraction Gratings and Applications. New York: Marcel Dekker, 1997.
8. AV Oppenheim, RW Schafer. Discrete-Time Signal Processing. Englewood Cliffs, NJ: Prentice, 1989.
9. MH Hayes, H Stark, ed. In: Image Recovery: Theory and Application. San Diego: Academic Press, 1987, Chapter 6.
10. DL Flannery, JL Horner. Fourier optical signal processors. Proceedings of the IEEE, 77(10):1511–1527, 1989.
11. GJ Swanson. Binary optics technology: the theory and design of multi-level diffractive optical elements. Massachusetts Institute of Technology Lincoln Laboratory, Technical Report 854, 1989.
12. N George. Speckle. SPIE Proceedings 243, San Diego, July 29–30, 1980.
13. P Hariharan. Optical Holography Principles, Techniques, and Applications. Cambridge Studies in Modern Optics, Cambridge: Cambridge University Press, 1996.
14. JC Dainty. Laser Speckle and Related Phenomena. New York: Springer-Verlag, 1984.

7

Multi-aperture Beam Integration Systems

Daniel M. Brown
MEMS Optical, LLC, Huntsville, Alabama

Fred M. Dickey and Louis S. Weichman
Sandia National Laboratories, Albuquerque, New Mexico

I. INTRODUCTION

Various high-power laser applications, such as laser heat processing, cutting, marking, photolithography, and fiber injection require a laser irradiance that is substantially uniform on a target over a specified area at a fixed longitudinal distance from the source. The irradiance pattern may be circular, hexagonal, rectangular, ring-shaped, or practically any other shape that can be defined by the boundary of an aperture. If the laser beam mode is well defined and constant in time then the beam shaping (field mapping) methods of Chapters 2, 3, 4, and 5 and the near-field beam shapers discussed briefly in Chapter 6 can be used. The field mapping approach discussed in these chapters is also applicable if the output is required to be collimated. However, in cases where the laser modes are unknown or change with time, and collimation is not required, a multi-faceted or multi-aperture beam integrator may be more desirable. This approach to beam shaping is especially suitable to excimer lasers (1,2) and other multi-mode laser beams, laser diode arrays (3,4), or other light sources with highly irregular irradiance distributions (5).

A multi-aperture integrator system basically consists of two components: (a) a subaperture array component consisting of one or more lenslet

arrays which segments the entrance pupil or cross section of the beam into an array of beamlets and applies a phase aberration to each beamlet, and (b) a beam integrator or focusing component which overlaps the beamlets from each subaperture at the target plane. These elements can be refractive, reflective, or diffractive. Generally, the subaperture elements all have the same shape and phase function to simplify fabrication, but varying their phase and shape within the array can provide greater irradiance uniformity in the target plane as shown by Pepler et al. (6) and others (7). The target is located at the focal point of the primary focusing element, where the chief rays of each subaperture intersect. Thus, the amplitude of the irradiance distribution on the target is a Fourier transform of the incoming wavefront aberrated by the lenslet array (8). Although this chapter primarily addresses the multi-aperture beam integration problem from the standpoint of refractive optics the concepts and analysis are directly applicable to reflective optics. Dickey and O'Neil treat multi-faceted reflective systems in considerable detail (5).

All beam integrators can be loosely divided into two categories; diffracting and imaging. A simple diffracting beam integrator (also called a non-imaging integrator (9)) is illustrated in Fig. 1, consisting of a single lenslet array and a positive primary lens. The target irradiance is the sum of defocused diffraction spots (point spread functions) of an on-axis object point at infinity (assuming a collimated input wavefront). If the source is spatially coherent over the lenslet aperture, or can be defined by a single field point from a ray optics point of view, then the diffraction spot will closely replicate the shape of the subaperture with diffraction rings (determined by the degree of defocus and other aberrations) superimposed. The defocus is caused by the additional optical power of the subaperture.

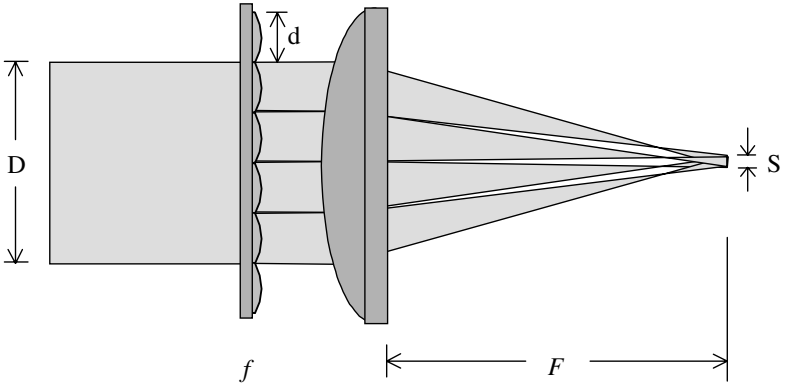


Figure 1 Diffracting multi-aperture beam integrator concept.

The diffracting beam integrator is based on the assumption that the output is the superposition of the diffraction fields of the beamlet apertures. The diffraction field is obtained using the Fresnel integral. If the beam is not spatially coherent over each beamlet aperture a more complicated integral is required and, generally, one would not be able to obtain a reasonable replica of the lenslet aperture. For example, a spatially incoherent field is approximated by a Lambertian source that radiates over a large angle and would not produce a localized irradiance distribution at the output plane.

Figure 2 illustrates an imaging multi-aperture beam integrator. This type of integrator is especially appropriate for spatially incoherent sources. From a ray optics perspective, these sources produce a wavefront incident over a range of field angles on the lenslet apertures. The first lenslet array segments the beam as before and focuses the beamlets onto a second lenslet array. That is, each lenslet in the first array is designed to confine the incident optical radiation within the corresponding aperture in the second array. A second lenslet array, separated from the first by a distance equal to the focal length of the secondary lenslets, together with the primary focusing lens forms a real image of the subapertures of the first lenslet array on the target plane. The primary lens overlaps these subaperture images at the target to form one integrated image of the subapertures of the first array element. Re-imaging the lenslet apertures mitigates the diffraction effects of the integrator in Fig. 1.

In this chapter we describe the theory and design of diffracting and imaging multi-aperture beam integration systems. We show how the subaperture shape and phase function determines the irradiance pattern in the image plane. We discuss the diffraction effects and the interference between the subapertures.

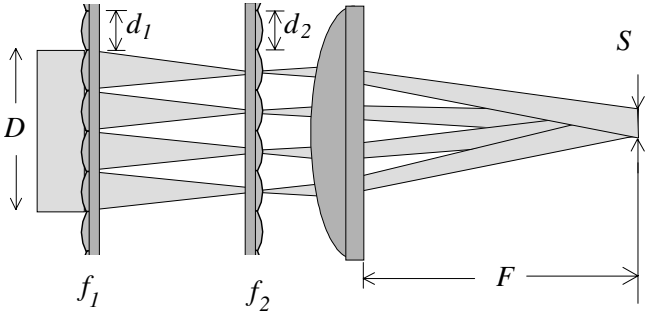


Figure 2 Imaging multi-aperture beam integrator concept.

In Sec. II we outline the basic theory and design considerations for multi-faceted beam integrators. A number of different optical configurations exist for multi-faceted beam integrators. Minimizing the interference effects produced by the finite number of subapertures and making the irradiance uniform in the target pattern is the design goal of these systems. We discuss these problems in this section. Multi-aperture beam shaping design methods are discussed in Sec. III. In this section we show how to use ZEMAX (10) and other ray tracing codes to design multi-faceted beam integrators. The effects of geometric aberrations are discussed. Fabrication considerations are discussed in Sec. IV, and applications and experimental data are presented in Section V.

Throughout this chapter we use the following fundamental design parameters to describe multi-aperture beam integration systems:

D = diameter of input beam at multi-aperture integrator
d = diameter of subaperture or lenslet
F = focal length of primary lens
f = focal length of array lenslet
S = diameter of target spot
λ = wavelength
$k = 2\pi/\lambda$
R_w = radius of curvature of wavefront
R_0 = radius of curvature of reference sphere centered on target

All other parameters will be defined as they are introduced.

II. THEORY

A major assumption in the multi-faceted approach to beam integration is that the laser beam does not have a time-varying divergence that is significant over the distance required to accomplish the integration; that is, only the irradiance fluctuates with time. This divergence requirement corresponds to a slowly varying, nearly uniform phase across each subaperture. This condition is required to guarantee a good overlap of the beamlets in the output plane. Further, it is required that the input beam has a high degree of spatial coherence over each facet. If this is not the case, the diffraction pattern of the beamlets will be dominated by the coherence function, not the aperture function defining the beamlets.

The analysis and design of laser beam integrators should include the effects of averaging, diffraction, interference, and imaging. With multi-faceted integrators, it is primarily the averaging process that is used to produce the desired irradiance distribution. Diffraction and interference

tend to produce undesired fluctuations in the irradiance distribution. Imaging can be used to control diffraction effects as well as the size (scale) of the irradiance distribution. Aberrations, which are inherent in the imaging process, tend to degrade integrator performance.

The beamlet geometry basic to multifaceted mirror integrators is illustrated schematically in Fig. 3. The figure does not, of course, describe the various lens or mirror geometries required to accomplish the integration. In the figure, the array at the left represents beamlets redirected from the laser beam to overlap at the square on the right. The x_0 and y_0 coordinate direction cosines for the beamlets are given by

$$\alpha_m = \frac{x_{1m}}{\sqrt{x_{1m}^2 + y_{1n}^2 + z^2}}, \qquad \beta_n = \frac{y_{1n}}{\sqrt{x_{1m}^2 + y_{1n}^2 + z^2}}. \tag{1}$$

The integrated optical field is the sum of the diffracted fields associated with the individual beamlets. The diffraction integral should be developed with respect to the limiting apertures that produce the beamlets. Also, imaging can be included as generalized diffraction. An assumption appropriate to multi-faceted integrators is that the optical field amplitude (or equivalently, irradiance) is approximately uniform over each limiting aperture forming the beamlets. Also, the angle between beamlets should be kept small to reduce the size of the following optics and associated aberrations.

There are four major assumptions in the development of diffracting beam integrators. We will list them here and discuss their impact. They are as follows:

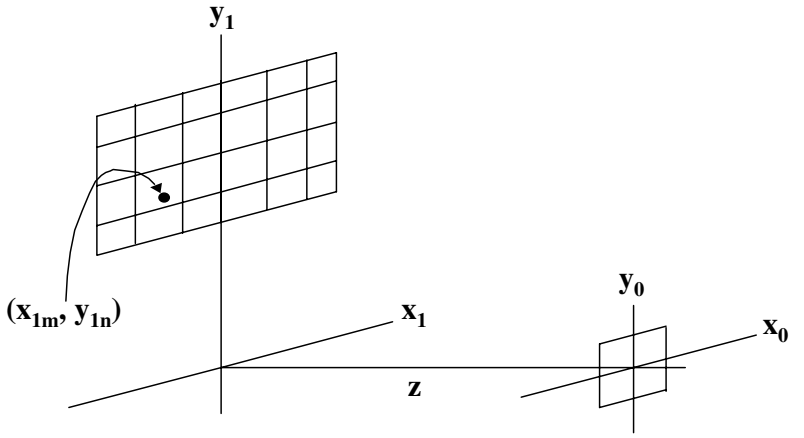


Figure 3 Beam integrator geometry.

1. The input beam amplitude (or equivalently irradiance) is approximately uniform over each subaperture. This allows for the output to be the superposition of the diffraction patterns of the beamlet defining apertures. It is expected that small deviations will average out in the output plane. That is, the errors associated with a particular aperture will not dominate.
2. The phase across each subaperture is uniform. The discussion in 1. applies in this case also. In addition, a linear phase across a subaperture results in a redirection of the beamlets, causing a misalignment in the output.
3. The input beam divergence does not vary significantly with time. Generally, an input beam divergence will result in a non-overlapping of the beams in the target plane. This can be corrected in many cases with correction optics in the input beam. However, a time-varying divergence would negate the possibility of correction.
4. The input beam field should be spatially coherent over each subaperture. This is inherent in assumption 1. since the diffraction patterns are assumed to be described by a Fresnel integral.

The imaging integrator does not require assumption 4. since it does not necessarily require that the output pattern be described by a diffraction integral.

A. Diffraction Considerations

The basic problem of multi-aperture beam integration is to map the input fields in the input apertures in Fig. 3 (x_1–y_1 plane) into the desired irradiance in the output plane (x_0–y_0 plane). It is assumed that the irradiance in the input plane is relatively uniform (constant) over each aperture. This assumption leads to the requirement of small apertures; however, there is a limit to how small one would make the apertures. This is the averaging problem discussed in Sec. II.C. Also, since the beamlets are superimposed in the output plane, small deviations from uniformity of the beam irradiance over the apertures should average out.

The theory of mapping a uniform input irradiance into a uniform output irradiance is developed in Appendix A of Chapter 3. The basic optical layout of a system for accomplishing this mapping, for each subaperture, using a diffracting multi-aperture beam integrator is illustrated in Fig. 1. The system consists of a lenslet array and a primary lens. The target plane is located at the focal point of the primary lens. A collimated beam of diameter D is incident on an array of lenslets, each of diameter d and focal length f, which segments the beam into multiple beamlets. The primary lens

of focal length F overlaps the beamlets, bringing the chief rays of each beamlet to a common focus at the back focal point of the primary objective where the integrated irradiance pattern is formed. The primary lens produces a field distribution at the focal plane that is proportional to the Fourier transform of the product of the functions representing the input beam and the lenslet array (8).

It is shown in Chapter 3, Appendix A using the method of stationary phase that the optical element that effects the mapping is a quadratic phase element, that is, a simple lens. The analysis is done in one dimension, but it can be extended to two dimensions. The phase of the optical element is given by

$$\Phi\left(\frac{x}{d}\right) = \beta\left(\frac{x}{d}\right)^2. \tag{2}$$

The stationary phase solution includes a parameter β that is a measure of the quality of the solution. This parameter, given by

$$\beta = \frac{\pi\, dS}{\lambda F}, \tag{3}$$

has the same form as the Fresnel number, differing only by a constant factor. Note that β is a dimensionless constant. The significance of β is discussed in considerable detail in Chapters 2 and 3. In Chapters 2 and 3, β is shown to be related to the mathematical uncertainty principle. Although the numerical coefficient in front of the factor, $dS/\lambda F$, may vary with the problem, the main result is that control of the shape of the beam can not be maintained if β is sufficiently small. Further, it can be seen from the form of Eq. (2) that if β is fixed the solution for the phase of the lenslet is fixed.

It can be shown, using diffraction theory (see Chapter 3, Sec. II.E), that if the phase function representing each lenslet is an even function then the negative of the phase also gives the same output irradiance. If the lenslets in the array are positive, as shown in Fig. 1, each beamlet will have a focus ahead of the focal point of the primary lens. If they are negative, the individual beamlets will either have a real focus after the focal point of the primary lens or a virtual focus ahead of the lenslets, depending on the relative optical powers of the lenslets and the primary lens. The numerical aperture of the beamlets and the distance between beamlet foci and target plane determines the spot size, S. It can be shown from paraxial geometrical optics that the spot size S on the target is equal to the focal length F of the primary lens divided by the f-number of the subaperture lens,

$$S = \frac{F}{f/d}. \tag{4}$$

This result is also obtained using diffraction theory and Fourier optics. It is apparent from Eq. (4) that the lenslet diameter can be varied over the array as long as its focal length is also varied proportionately such that the f-number (f/d) remains constant. Also, one can change the focal length F of the primary lens to scale the diameter S of the target spot and not change β.

The diffraction pattern of a single subaperture determines the shape of the spot on the target. The Fresnel number is useful for estimating this irradiance distribution on the target. The Fresnel number equals the number of half waves of optical path difference (OPD) and is approximately given by

$$Fresnel = 2 \times OPD = \frac{r^2}{\lambda}\left(\frac{1}{R_w} - \frac{1}{R_o}\right), \tag{5}$$

where R_w is the radius of curvature of the wavefront, R_o is the radius of curvature of a reference sphere centered on the observation point, and r is the radial coordinate in the subaperture. For a uniformly illuminated aperture the Fresnel number also equals the number of peaks in the aberrated point spread function (PSF) cross-section. Even integer Fresnel numbers have an on-axis minimum in the diffraction PSF. Odd integer Fresnel numbers have an on-axis peak intensity in the PSF. This is shown in Fig. 4(a) and (b). Generally, the Fresnel number is a measure of the complexity of the diffraction pattern; the number of maxima increases with increasing Fresnel number while the depth of the modulation decreases with increasing Fresnel number.

Assuming a collimated beam incident onto the lenslet array, the optical powers of the lenslets and the primary lens combine to produce a spherical wavefront converging with a radius of curvature of R_w. Substituting $1/R_w = 1/f + 1/F$, $1/R_o = 1/F$, and $d = 2r$ into Eq. (5) gives the Fresnel number in terms of lenslet parameters,

$$Fresnel = \frac{d}{4\lambda(f/d)}. \tag{6}$$

Alternatively, substituting Eq. (4) into Eq. (6), the Fresnel number can be written in terms of the target spot size and primary lens focal length,

$$Fresnel = \frac{dS}{4\lambda F}. \tag{7}$$

Since $\lambda F/d$ is proportional to the width of the point spread function of a single subaperture, Eq. (7) shows the Fresnel number is proportional to the number of PSFs across the target pattern. The lower the Fresnel number, the more rounded the target pattern becomes.

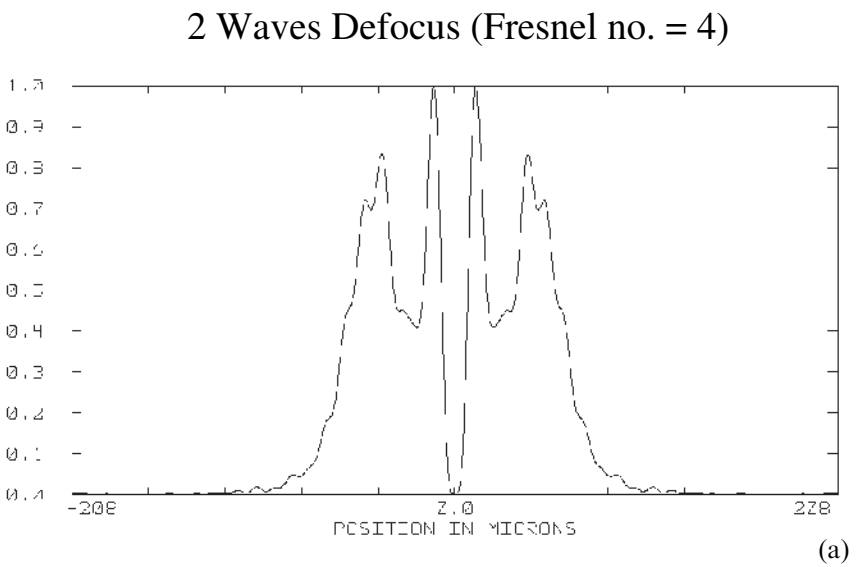


Figure 4 (a) Diffraction patterns of on-axis subaperture pupil functions as a function of Fresnel number 2 waves defocus and (b) 2.5 waves defocus.

B. Interference Effects

The output irradiance in Fig. 3 is the superposition of the diffraction patterns of the input aperture fields. Depending on the degree of coherence of the source, the output irradiance will contain a component of an interference or speckle pattern. For these conditions, the integrated irradiance of the coherent component is adequately described by

$$I(x,y) = \left| \sum_{0,0}^{M,N} A_{mn} \exp\{i[k(\alpha_m x + \beta_n y) + \theta_{mn}]\} \right|^2 |F(x,y)|^2 \tag{8}$$

where α_m and β_n are the direction cosines defined in Eq. (1), θ_{mn} is the phase of the beamlet, A_{mn} is the amplitude of the beamlet field, and the function $F(x,y)$ is the diffraction integral of the beamlet-limiting aperture. $F(x,y)$ is the Fourier transform of the aperture function for the optical configuration in Fig. 1.

The first factor in Eq. (8) describes the averaging and interference effects of the integrator. The interference effect is a result of the sum of linear (in x and y) phase terms, which can be viewed as a Fourier series. The spatial period of the resulting interference pattern is given by

$$P = \frac{\lambda}{\alpha} \tag{9}$$

where α is the angle between adjacent beamlets.

For a spatially coherent source the interference pattern will generally result in large fluctuations in the integrated irradiance. The only practical way to negate the effects of interference is to choose a sufficiently large value for α so that the interference pattern is too fine to be resolved in the application. If this is done, the effective integrated irradiance will be the local average of the irradiance in Eq. (8). It is easy to show that the averaged irradiance is

$$I(x,y) = \sum_{0,0}^{M,N} |A_{mn}|^2 |F(x,y)|^2. \tag{10}$$

This result represents the ideal performance of a multi-faceted beam integrator. Note that this result does not depend on θ_{mn}, the relative phase of the beamlets. The effects of the diffraction term are discussed with respect to specific configurations in the following sections.

The above results can be obtained for the system in Fig. 1 using Fourier transform theory. Since the target is located at the focus of the primary lens, the irradiance pattern on the target is simply the magnitude squared of the Fourier transform of the pupil function modified by the

lenslet array (assuming a spatially coherent source). If all the lenslets are identical, the aberrated pupil function is approximately the convolution of a two-dimensional delta function array (array of lenslet centers) and the lenslet pupil function, all multiplied by the laser beam amplitude profile. The Fourier transform is then the product of the Fourier transform of the delta function array (properly scaled) and the aberrated PSF of the subaperture, convolved with the PSF of the laser beam focused by the primary lens. Mathematically, the field just past the lenslet array is given by

$$E = E_b(x,y)[\Psi(x,y) * A_{\delta\delta}(x,y)], \tag{11}$$

where $E_b(x,y)$ is the field of the laser beam, $\Psi(x,y)$ is the lenslet pupil function, $A_{\delta\delta}(x,y)$ is the delta function array, and $*$ denotes convolution. The irradiance at the target plane is proportional to the magnitude squared of the Fourier transform of E,

$$I \propto |\Im[E]|^2 = \left|\tilde{E}_b * \left\{\tilde{\Psi}(x,y)\tilde{A}_{\delta\delta}(x,y)\right\}\right|^2, \tag{12}$$

where the tilde denotes the Fourier transform operation. Since the Fourier transform of a periodic two-dimensional delta function array is a two-dimensional delta function array, the function, $\tilde{A}_{\delta\delta}(x,y)$, is the source of the interference effects (11). The subaperture PSF, $\tilde{\Psi}(x,y)$ represents the diffraction effects. These diffraction and interference effects produce undesirable fluctuations in the irradiance distribution.

For a rectangular aperture array, the spatial period of the interference pattern is given by the ratio of the wavelength to the sine of the angle between adjacent beamlets:

$$Period = \frac{\lambda}{\sin\theta}. \tag{13}$$

Since $\sin\theta$ approximately equals the lenslet spacing divided by the focal length of the primary lens, the interference periodicity is also given by

$$Period = \frac{\lambda F}{d}. \tag{14}$$

Equation (13) and (14) are obtainable from $\tilde{A}_{\delta\delta}(x,y)$ scaled to the focal plane of the primary lens.

C. Averaging

The averaging aspect of multi-aperture beam integrating systems consists of making tradeoffs between lenslet aperture size and β or, equivalently, the Fresnel number. Increasing the lenslet aperture size, d, increases β which

reduces diffraction effects. However, averaging is reduced. What is desired is to have d as small as possible while maintaining adequate β.

Assuming the intensity distribution within each subaperture is relatively uniform, the superposition of all the subapertures on the target plane will give a relatively uniform irradiance distribution (except for diffraction and interference effects). The assumption of uniform intensity within a subaperture is of course more valid the smaller the diameter of the subaperture. However, holding the f/d constant in order to maintain a spot size in accordance with Eq. (4), Eq. (6) shows the Fresnel number decreases linearly with subaperture diameter, which results in fewer peaks in the diffraction pattern with a correspondingly greater depth of modulation and a more gradual roll-off on the edges of the pattern at the target. This may require expanding the input beam in order to make the subaperture diameters larger or reducing the focal length of the primary lens. Varying the lenslet diameter according to Eq. (6) across the lenslet array can also be used to superimpose different diffraction patterns such that the peaks of one pattern fall into the valleys of another. Varying the subaperture diameters by integer multiples allows 100% fill factor to be maintained. This is the approach taken by Pepler, et al. (6).

D. Coherence Effects

As discussed in Sec. II.B, multi-aperture beam integration systems will generally exhibit a degree of interference or speckle. The amount of speckle is determined by the spatial coherence of the source. A spatially incoherent source will not produce an interference/speckle pattern and a spatially coherent source will produce the maximum interference/speckle. The results for intermediate cases will depend on the spatial coherence function representing the source. A general formulation of the problem in terms of coherence theory is quite involved and beyond the scope of this chapter. However, we will give a simplified formulation that will illustrate the major aspects of the problem.

The multi-aperture beam integrator can be viewed as a multi-beam interferometer since all of the beamlets are superimposed at the target plane. The superposition of any two of the beamlets mimics Young's experiment. The visibility of fringes in Young's experiment is the definition of spatial coherence. In the following, we will give only the rudiments of coherence theory needed to develop the simplified model of beam integration systems. For the basics of coherence theory the authors recommend the book, *Statistical Optics*, by J. W. Goodman (12).

Assuming a quasimonochromatic source, a generally good assumption for lasers, the spatial coherence of a laser is adequately described by the mutual intensity

$$J_{12} = \langle u(P_1, t)u^*(P_2, t)\rangle, \tag{15}$$

where $u(P_1, t)$ is the analytic signal describing the optical field, P_1 and P_2 are points in the plane in which the coherence of the beam are being represented, and $\langle \cdot\cdot \rangle$ denotes a time average. The mutual intensity is a correlation function when the functional dependence is an explicit function of the difference in coordinates, that is, $J_{12}(P_1, P_2) = J_{12}(P_2 - P_1)$. Note that by definition the irradiance (frequently called intensity) of the optical field is obtained for $P_1 = P_2$ as

$$I(P_1) = J_{12}(P_1, P_1). \tag{16}$$

The complex coherence factor is defined as the normalized mutual intensity,

$$\mu_{12}(P_1, P_2) = \frac{J_{12}(P_1, P_2)}{[I(P_1)I(P_2)]^{1/2}}. \tag{17}$$

When one produces an interference pattern by combining radiation from points P_1 and P_2 in a Young's interferometer configuration, J_{12} may be regarded as the phasor amplitude of the spatial sinusoidal fringe pattern (on axis), whereas μ_{12} is the normalized fringe pattern. The complex coherence factor has the property,

$$0 \leq |\mu_{12}| \leq 1. \tag{18}$$

When $\mu_{12} = 0$ there are no interference fringes, and the two optical fields are mutually incoherent. When $\mu_{12} = 1$ the two optical fields are perfectly correlated, and the two fields are mutually coherent. For intermediate values of μ_{12} the fields are partially coherent.

The mutual intensity J_{12} can be computed at an output plane given J_{12} in the input plane using a generalized Van Cittert–Zernike theorem (12) and the relation

$$J_t(P_1, P_2) = t(P_1)t^*(P_2)J_i(P_1, P_2). \tag{19}$$

Equation (19) relates the mutual intensity transmitted by the object with transmission t to incident mutual intensity. In Eq. (19) the numerical subscripts have been dropped and replaced by i and t, which represent the incident and transmitted mutual intensity respectively. The generalized Van Cittert–Zernike theorem is a fourth-order integral over four variables. To develop such an integral for the system of Fig. 1 would be very difficult

and may not be very enlightening. For this reason we will make a simplifying assumption that will illustrate the basic concepts. The basic assumption that we will use is that the field is mutually coherent over each subaperture in Fig 1, and will generally be partially coherent over greater distances. The assumption that the field is mutually coherent over each subaperture is fundamental to the performance of diffracting multi-aperture beam integration systems (see Sec. II). This assumption can be reduced for the case of imaging integrators. The following analysis will apply specifically to diffracting beam integrators.

With this assumption, we can approximate the field at the output plane as the sum of the coherent diffraction from each aperture,

$$E(x,y) = \sum_{0,0}^{M,N} [A_{mn} \exp\{i[k(\alpha_m x + \beta_n y) + \theta_{mn}]\}] F(x,y). \tag{20}$$

The functions and variables in Eq. (20) are defined following Eq. (8). Assuming a degree of partial coherence between the beamlets, the coherence aspects of the problem are contained in the correlation between the amplitudes, A_{mn}. Given this, we can write the intensity at the target plane as

$$\begin{aligned} I(x,y) = \Bigg\langle & \sum_{0,0}^{M,N} [A_{mn} \exp\{i[k(\alpha_m x + \beta_n y) + \theta_{mn}]\}] F(x,y) \\ & \times \sum_{0,0}^{M,N} [A_{pq}^* \exp\{-i[k(\alpha_p x + \beta_q y) + \theta_{pq}]\}] F^*(x,y) \Bigg\rangle. \end{aligned} \tag{21}$$

Noting that the time average only involves the amplitudes, the time average can be written as

$$\begin{aligned} I(x,y) = & \sum_{0,0}^{M,N} |A_{mn}|^2 |F(x,y)|^2 \\ & + \sum_{m,n \neq p,q}^{M,N} [\langle A_{mn} A_{pq}^* \rangle \exp\{i[k([\alpha_m - \alpha_p]x + [\beta_n - \beta_q]y) \\ & + \theta_{mn} - \theta_{pq}]\}] |F(x,y)|^2. \end{aligned} \tag{22}$$

Noting the time average bracket can be interpreted as a mutual intensity, we can write the last equation as

$$I(x,y) = \sum_{0,0}^{M,N} |A_{mn}|^2 |F(x,y)|^2 + \sum_{m,n\neq p,q}^{M,N} [J_{mn,pq} \exp\{i[k([\alpha_m - \alpha_p]x + [\beta_n - \beta_q]y) + \theta_{mn} - \theta_{pq}]\}]|F(x,y)|^2. \quad (23)$$

Using Eq. (17), the last equation can be written in terms of the complex coherence factor as

$$I(x,y) = \sum_{0,0}^{M,N} |A_{mn}|^2 |F(x,y)|^2 + \sum_{m,n\neq p,q}^{M,N} \sqrt{I_{mn} I_{pq}}[\mu_{mn,pq} \exp\{i[k([\alpha_m - \alpha_p]x + [\beta_n - \beta_q]y) + \theta_{mn} - \theta_{pq}]\}]|F(x,y)|^2. \quad (24)$$

The reader should note that the subscripts refer to aperture elements in the lenslet array and the intensities, I_{mn}, are assumed constant over each aperture. With this in mind, the reader can see that the second term in Eq. (24) is responsible for the interference (speckle) effects, and the first term is the sum of the irradiances at the output associated with each aperture. Clearly, if the fields in each aperture are mutually incoherent, $\mu_{mn,pq} = 0$, we have

$$I(x,y) = \sum_{0,0}^{M,N} |A_{mn}|^2 |F(x,y)|^2, \quad (25)$$

which is just Eq. (10). When all of the aperture fields are mutually coherent, $\mu_{mn,pq} = 1$, we have maximum interference, and the exact form is dominated by the I_{mn} and θ_{mn}. In all other cases the interference pattern will be influenced by the correlation between the various aperture fields, $\mu_{mn,pq}$, as well as I_{mn} and θ_{mn}. In all cases, the envelope of the irradiance pattern is given by the function, $|F(x,y)|^2$.

To estimate the effect of the coherence of the input laser beam on the interference pattern in the output using Eq. (24) one needs to estimate the complex coherence factor. The complex coherence factor can be measured by repeating Young's experiment for pinhole pairs with different spacings between the holes. This is generally a tedious process. It is not uncommon, in practice, to design a prototype beam integration system for a given laser and experimentally measure its performance as part of the of the engineering design process. Experimental evaluation of the effects of spatial coherence is

presented in Sec. V.A. This section presents experimental data for spatially coherent (between subapertures) beams, partially coherent beams, and spatially incoherent beams.

E. Imaging Integrators

Diffracting integrators (illustrated in Fig. 1) are restricted to sources with a relatively high degree of spatial coherence over each subaperture. In addition, a change in the angle of incidence of the incoming light on the array causes a lateral shift in the irradiance spot on the target. As spatial coherence decreases or the angular spread, or field angle of the incident light increases, the irradiance blurs to a spot larger than that given by Eq. (4). This problem is eliminated with an imaging integrator, illustrated in Fig. 2. The angular spread or field angle of the incident light can be quite large and still maintain overlap of the beamlets on the target plane. The imaging integrator can also offer improved integrator performance when the beam is collimated with a high degree of spatial coherence by reducing the diffraction effects. This effect is discussed in detail for reflective systems in Reference 5.

An imaging integrator requires two lenslet arrays. The first lenslet array segments the input beam into multiple beamlets and directs these onto the second lenslet array. The first lenslet array serves to produce an intermediate image plane on or near the second lenslet array so that the second lenslet array can reimage the subapertures of the first array onto the target plane. The magnification of the subaperture images is given by the ratio of the focal lengths of the primary integrator lens and of the second lenslet array. Thus, the spot size is again given by Eq. (4) with the slight modification that the second array lenslet focal length and first array lenslet diameter are used.

$$S = \frac{F}{f_2/d_1}. \tag{26}$$

The second lenslet array serves as a field lens to redirect the off-axis chief rays back toward the optical axis. If the distance between the two lenslet arrays equals the focal length of the first array, then the integrator can receive incident light over a maximum field angle equal to the second array subaperture semidiameter divided by the focal length of the first array.

$$\theta = \frac{d_2}{2f_1}. \tag{27}$$

The simplest form of the second lenslet array element incorporates identical lenslets with mutually parallel optical axes. This is not a necessary

requirement, for example, if a continuously variable tilt or decenter is applied to each of the lenslets as a function of position in the array, the system can be made into a zoom integrator with a continuously adjustable spot size. Zoom integrators are discussed further in Sec. III.B and several configurations are reported in Reference 5.

III. DESIGN CONSIDERATIONS

The first step in the design process is to decide between building a diffracting or imaging beam integrator, and whether to use refractive, reflective, or diffractive components. Generally, an imaging beam integrator will produce lower diffraction effects and better homogenization, particularly for sources with low spatial coherence. A diffracting integrator allows greater flexibility in shaping the irradiance spot through aberrations and aperture flipping (see Sec. III.C). Imaging integrators generally introduce more complexity since there are more optical elements and associated alignment issues. Equation (4) is used to determine the first-order parameters of the lens elements of a diffracting integrator. The first-order parameters for imaging integrators are obtained using Eqs. (26) and (27).

Equation (3) or (7) is used to insure that the parameters in a first-order design result in β or Fresnel number values required for a good beam shaping result. Depending on requirements the Fresnel number should be at least 3.0 and the beta parameter should be at least 40. The period of the interference pattern is obtained using Eq. (14). Generally, it is desirable for the period of the interference to be small to reduce the interference effects. Source spatial incoherence also reduces the contrast of the interference pattern. Various software programs, either geometrical ray tracing or wave propagation, can be used to evaluate and optimize the lens aberrations to achieve the desired integration and smoothing.

A. Diffracting Integrator Layout

The optical layout of a basic multi-aperture beam integrator is shown in Fig. 5. For clarity, only the rays for two outer lenslets are drawn. A collimated beam from a laser source is incident from the left. A lenslet array breaks up the incoming beam into an array of beamlets which are then overlapped at the target by a primary integrator lens. Either positive lenslets or negative lenslets (or a combination of the two) may be used in the array. Positive lenslets will produce a real beamlet focus ahead of the target plane (as shown in the figure). Negative lenslets will have a virtual beamlet focus either ahead of the lenslet array or behind the target plane, depending on

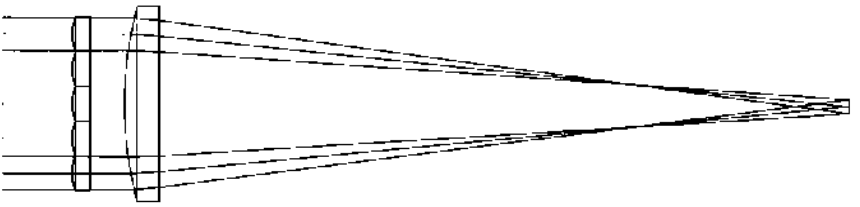


Figure 5 Optical layout of a refractive diffracting beam integrator.

the relative optical powers of the lenslets and primary lens. In the simplest configuration, all the lenslets are identical. Note, the spacing between the lenslet array and the primary integrator lens is not critical to first order.

The target plane is located at the focal point of the primary lens. The positive primary lens focuses the chief rays of each beamlet to a common point on the target, thus overlapping the defocused beamlets at the target. If the lenses are sufficiently free of aberrations then the spot formed at the focal point of the primary will replicate the lenslet aperture. Square lenslet apertures will form a square irradiance pattern, and circular lenslet apertures will form a circular irradiance pattern. If the lenslet has positive optical power, the spot will be an inverted replica of the aperture. If it has negative optical power, the spot will be an upright replica of the aperture. This is the basis for the aperture flipping technique discussed in Sec. III.C. As shown in section Sec. III.D, aberrations can be added to the lenslets to significantly distort the irradiance pattern into almost any arbitrary shape.

B. Imaging Integrator Layout

Imaging integrators are well suited to sources with a low degree of spatial coherence (13). Their drawback is the loss of flexibility in irradiance pattern shaping through lenslet aberrations. Figure 6 shows an optical layout for a simple imaging integrator with an extended source. For clarity, only the rays for two outer lenslets are drawn. The first lenslet array segments the input into multiple beamlets and focuses the beamlets onto the corresponding

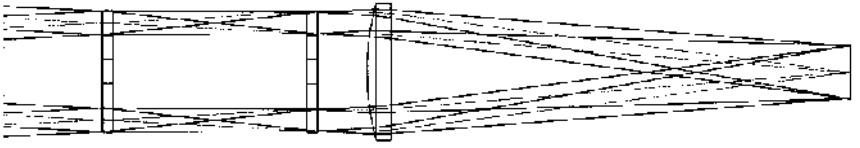


Figure 6 Optical layout of a refractive imaging integrator.

lenslets of the second array. To minimize stray light outside the target area, the beamlet diameters must not be greater than the lenslet clear apertures at the second array. Minimum spot size on the second array elements occurs when the array separation distance equals the first array lenslet focal length, in which case Eq. (27) applies. Each element of the second lenslet array combined with the primary lens forms a two-lens telescope that produces a real image of the pupils (of the first array lenslets) at the target plane. The geometry of each telescope is such that the pupil images are superimposed at the output plane.

To make a zoom imaging integrator that allows continuous adjustment of the target spot diameter, one simply needs to add a second integrator lens to the system, either ahead of or behind the primary integrator lens. The first or primary integrating lens forms a virtual image of the integrated irradiance pattern at a finite conjugate distance. The secondary integrator lens reimages the virtual image onto the target plane with a magnification equal to the ratio of the conjugate distances.

Alternatively, one can add to each lenslet in the second array a tilt that is proportional to the lenslet's radial distance from the system optical axis. This is equivalent to superimposing the primary integrator lens onto the lenslet array and thus eliminating one element. Figure 7 shows such a zoom imaging integrator.

C. Subaperture Shape

The fact that the spot has the same shape as the subaperture allows almost any spot shape to be produced. Usually, the designer wishes to stack the lenslets with 100% fill factor in order to maximize the energy on the image

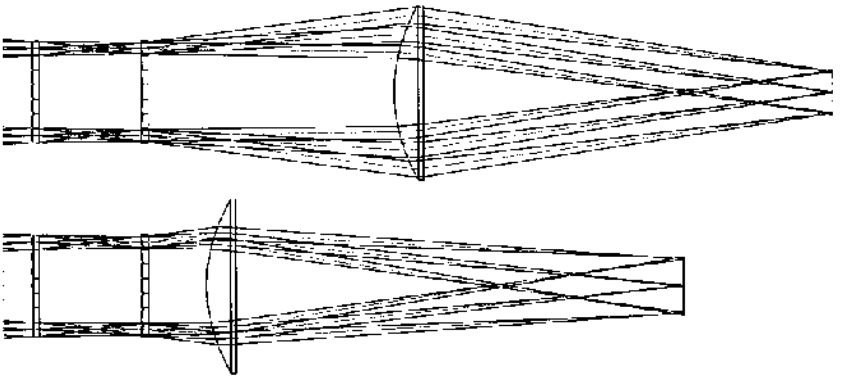


Figure 7 Zoom imaging integrator for variable spot size.

plane. Lenslet aperture shapes which are easily stacked with 100% fill factor include square, rectangular, and hexagonal. If a diffracting integrator is used, one can form a triangular irradiance pattern with 100% fill factor at the lenslet array. If a triangular subdivision of a square, rectangle, or hexagon results in only two different triangles which are inverted images of each other, then the sign of the phase function can be flipped for inverted apertures resulting in a single integrated image of one of the triangles.

An optical layout of an aperture-flipped diffracting integrator is illustrated in Fig. 8. As shown in Fig. 9, the lenslet apertures are equilateral triangles in a hexagonal packing. Triangles oriented with apex up have positive optical powers. Triangles with apex down have negative optical powers but of the same magnitude. The positive apex-up apertures become inverted at the target but the negative apex-down are not. Their superposition results in a single apex-down triangular irradiance pattern. Figure 10 illustrates triangular apertures in a rectangular packing.

Aperture flipping is practical only with diffracting integrators. To use aperture flipping with imaging integrators would require at least a third lenslet array, adding greatly to system complexity. In order to incorporate aperture flipping and maintain 100% fill factor at the lenslet array, the subdivision of the basic square, rectangle, or hexagon shape must result in two subapertures which are inverted images of each other. The phase functions in these two apertures have the same magnitudes but opposite signs.

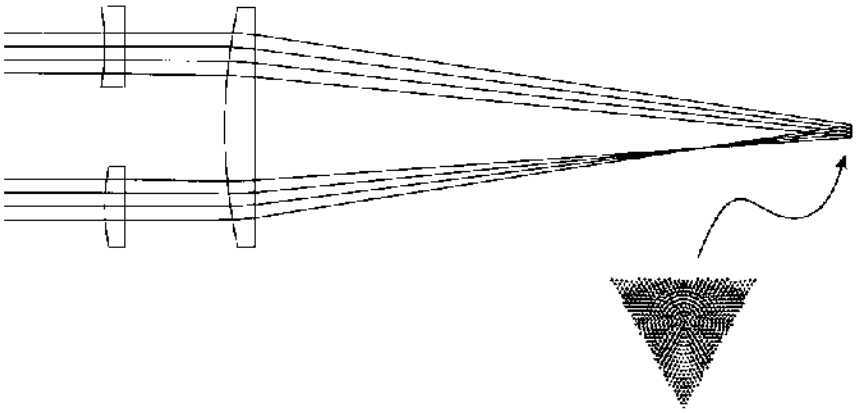


Figure 8 Diffracting integrator incorporating aperture flipping.

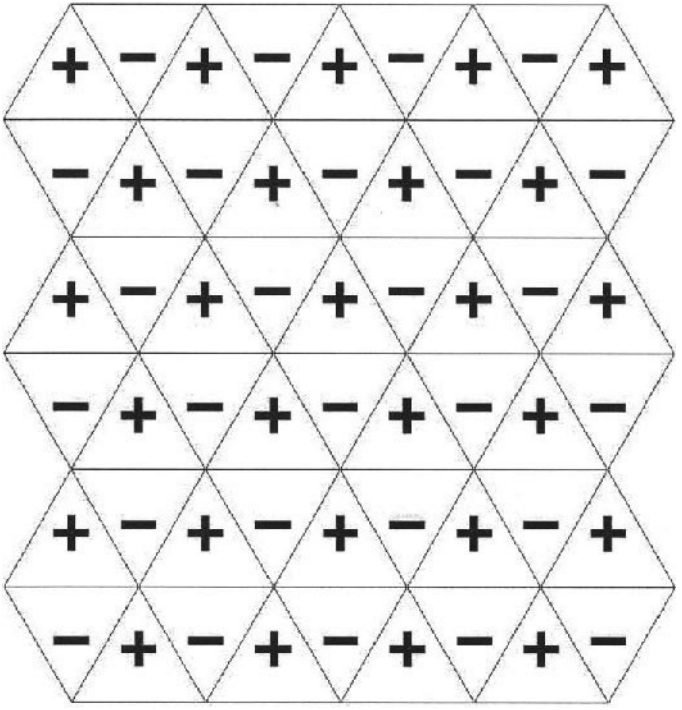


Figure 9 Triangular apertures in a hexagonal packing with sign-flipped optical powers.

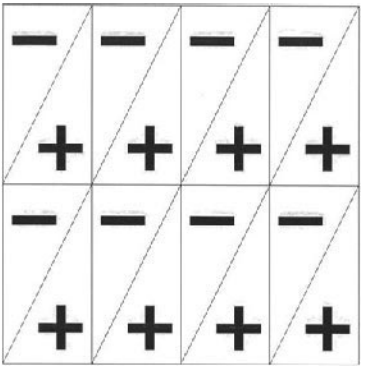


Figure 10 Triangular apertures in a rectangular packing with sign-flipped optical powers.

D. Lens Phase Functions and Aberrations

Once the first-order layout and aperture shape are determined, aberrations in the lens elements can be adjusted to fine-tune the irradiance distribution on the target. Aberrations in both the lenslet array and the primary integrator lens affect the irradiance uniformity on the target. Aberrations in the primary lens (e.g. spherical aberration) will result in a lateral displacement between the overlapped beamlet diffraction patterns at the target which will tend to compensate for the diffraction-induced irradiance modulation. Equivalently, a slight defocus of the target from the ideal focal point produces a similar effect, as shown by Deng et al. (14).

The phase function of the lenslet array elements affects the diffraction pattern or image of the lenslet apertures. For diffracting integrators, nonquadratic lenslet phase functions can be used to significantly modify the shape of the irradiance pattern. This is particularly useful in laser machining where one desires to shape the laser irradiance to the clear aperture of a fabrication mask in order to increase the laser power through the mask (15). For example, a phase function which is a linear function of radius, $\phi = \alpha r$, where r is the radial coordinate in each subaperture, will produce a ring pattern. This is illustrated in Fig. 11. Horizontal and vertical rectangular subapertures with decentered lens functions can be used to form hollow square patterns. Rooftop prisms can be superimposed onto the lenslets, giving a linear phase tilt in one axis, to produce two parallel bars in the irradiance pattern (16). Alternatively, lenslets can be decentered to produce multiple spots on the target (17).

Figure 12 illustrates three different lenslet surface profiles and their corresponding irradiance patterns on the target (determined by ray tracing).

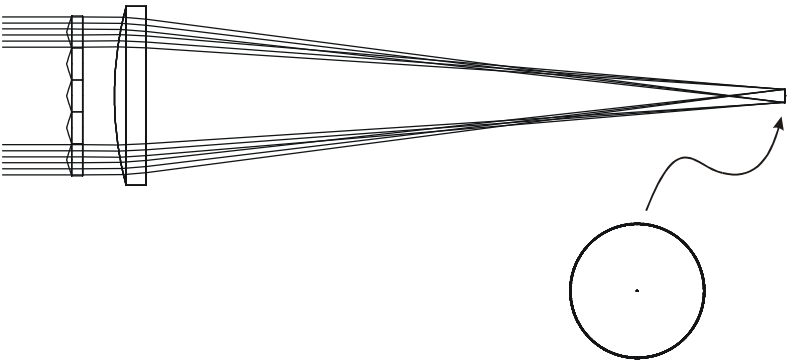


Figure 11 Diffracting integrator forming a ring pattern on target.

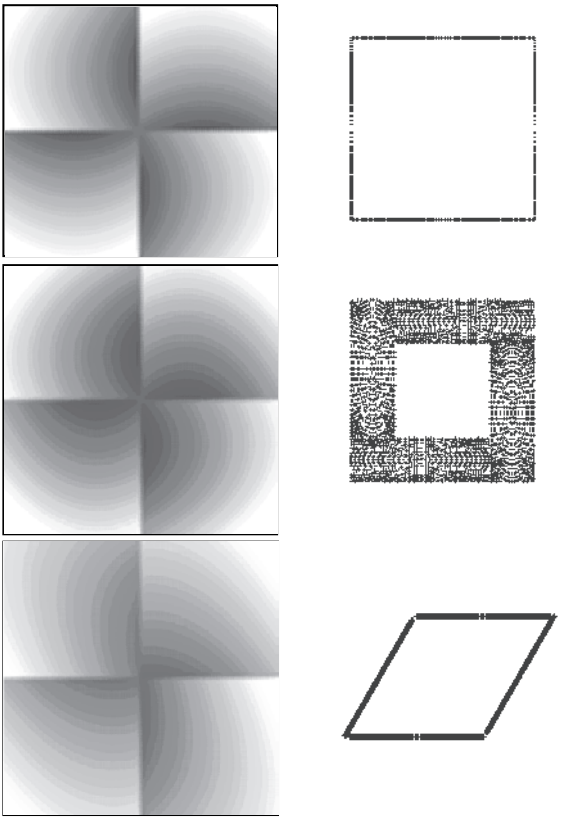


Figure 12 Lenslet surface profiles (represented as gray scale levels) and resulting irradiance patterns; hollow square, thick hollow square, and hollow trapezoid.

The surface profiles, represented as gray scale patterns, are similar to first order but differ in the higher-order (aberration) terms. Variation of the aberrations within the same basic lenslet profile allows producing a thin hollow square, thick hollow square, or hollow trapezoid irradiance patterns. The square aperture of each lenslet has been subdivided into four square subapertures to form the four sides of the patterns. A lenslet with only tilt but no optical power will produce a decentered point on the target. Optical power in one direction in the lenslet aperture will spread the point in the same direction on the target. The hollow patterns are easily made with cylindrical lenlets with a tilt or decenter term superimposed. Thick hollow patterns are made with a small anamorphic term or optical power in the orthogonal direction.

IV. FABRICATION CONSIDERATIONS

The key element in multi-aperture beam integrators and most difficult element to fabricate is the lenslet array. This element can be refractive, reflective, diffractive–transmissive, or diffractive–reflective. Metallic reflective anamorphic or non-rotationally symmetric elements can be diamond turned (18) Various methods of fabricating microlenses are discussed in detail in the literature (19,20), so we will only briefly outline two of the more common technologies for fabricating these elements. Photolithographic technology can be used to fabricate any of the above types of lenslet arrays. Photoresist of appropriate thickness is spun on the glass wafer and hardened by baking. The lens surface profile is formed in the photoresist by exposure to UV light through chrome or gray scale masks and then development of the photoresist. The lens surface profile is then transferred into the glass by reactive ion etching or ion milling.

Inherent limitations of this fabrication technology vary from vendor to vendor, but generally include wafer size limitations, wafer material limitations, maximum etch depth limitations, minimum feature size limitations, etch uniformity across the wafer, and lens surface profile accuracy. Acceptable wafer thickness typically ranges from about 300 microns to several millimeters. The maximum wafer diameter that can be easily processed is about 150 mm. Almost any material can be ion milled but only a few materials can be reactively ion etched. The common material choices for reactive ion etching include fused silica, silicon, and ZnSe. A few other more exotic materials can also be etched with reactive ions, but costly process development is often required with these materials.

Etch selectivity (ratio of wafer etch rate to photoresist etch rate) can be varied over a limited range with reactive ion etching by changing the etchant gas mixture. Selectivity for the ion milling process is usually limited to a ratio of about one. Limits on selectivity and maximum workable photoresist thickness places a limit on the maximum etch depth or lens surface sag. The increase in surface roughness with increasing etch depth also limits maximum etch depth. The maximum etch depth for fused silica is about 20 μm. Silicon, because of its higher etch rate can be etched much deeper, to 60 μm or more. The above numbers are loose approximations and vary from vendor to vendor, but they provide the reader a general idea of the types of fabrication limitations involved.

Photoresist heating and reflow is a common inexpensive method of fabricating microlenses. A single binary chrome mask is used to produce pillars of photoresist which are subsequently reflowed into a lenslet surface by heating. The profile is then etched into the glass wafer by reactive ion etching or ion milling. A few drawbacks of this method are noted here. First,

it is difficult to get 100% fill factor (ratio of lens surface area to wafer surface area) with this method. A sufficiently large gap must exist between adjacent lenses to prevent the merging of photoresist from separate lenses. Second, the reflow method cannot produce negative lens elements. Third, the lens elements tend to have focal lengths which are too short for beam integrators.

The alternative gray scale mask fabrication method solves the above problems as can be seen in Fig. 13. The drawback of gray scale mask technology is greater process development and cost. The inherent nonlinearities of this process often require iterative corrections to the gray scale mask to accurately produce the desired lenslet surface profile.

Reliable surface profilometry equipment, whether contact or optical, is essential for fabricating microlenses. Contact surface profilometers which do not raster scan the lenslet profile, but rather make only a single scan (cross section) of the lens, can be difficult to use as the stylus must scan through the vertex of the lens in order to obtain an accurate measurement of the surface profile. Optical profilometers, such as interferometric microscopes, can capture the entire three-dimensional surface profile of the lens. However, optical profilometers designed to measure flat surfaces can introduce erroneous spherical aberration into the measurements when measuring the steep surface curvatures of microlenses. A Fizeau interferometer can give better

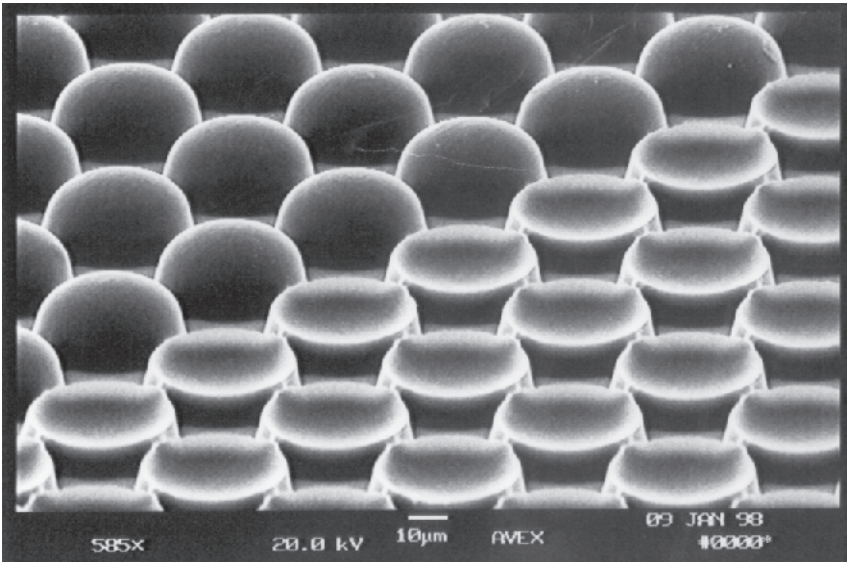


Figure 13 Array of positive and negative elements fabricated by gray scale mask technology.

results for surface profiles which are close to spherical. A ball lens or metal microsphere can be used to check a particular instrument to determine if it introduces erroneous aberrations into the measurements.

V. APPLICATIONS AND EXPERIMENTAL EVALUATION

Numerous applications exist for multi-aperture beam integrators. A few of these applications include; laser heat processing (including medical and dental applications), laser machining, product marking, laser diode array integration for laser pumping and fiber injection, photolithographic mask aligners and steppers, and fiber injection systems. We show experimental results of a fiber injection application below.

The light sources of photolithographic steppers and contact mask aligners are typically high-pressure mercury or xenon arc lamps which supply the required high-intensity UV irradiance. The energy produced by these highly compact and relatively noisy arc sources must be uniformly distributed over the area of the photolithographic mask. A highly uniform irradiance at the mask plane is particularly critical for gray scale mask processes. Multi-aperture beam integrators have been successfully used on photolithographic equipment to homogenize arc sources for many years. Examples of such beam integrators are described in patents by T. Mori (21) and H. Komatsuda (22).

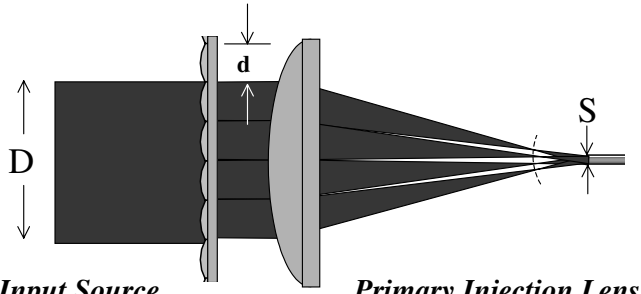
Arc sources can be modeled as a series of small concentric ellipsoids, located near the cathode tip, whose radiance decreases with increasing size of the ellipsoids. Most of the radiant power originates from an ellipsoidal region less than 1 mm in diameter near the cathode tip. Although these sources are often referred to as "point sources," their finite size still results in a finite angular distribution of intensity in the collimated beam. Thus, arc sources are only partially spatially coherent. Nonsequential ray tracing can be used to determine the intensity distribution of collimated arc sources for a given collimator system. Due to the limited spatial coherence of arc sources, imaging beam integrators are ideally suited for forming uniform irradiance patterns with these sources. The Van Cittert–Zernike theorem can be used to model the spatial coherence of collimated arc sources.

A related but slightly different application of beam integrators is the combining of the multiple emitters of laser diode arrays to form a single irradiance pattern. For example, a laser printer application might require the magnified line images of each emitter to be superimposed on the target. Instead of segmenting a collimated source input, the lenslet array reimages multiple sources at finite conjugates. An integrator lens overlaps the images at the target.

A. Experimental Evaluation of Diffracting Beam Integrator

The theory and effectiveness of the design approach discussed in this chapter for the diffracting or non-imaging beam integrator can be illustrated using experimental data recently collected on a compact fiber injection system (23). For this application a single lenslet array and a plano-convex lens were employed as the fiber injection elements to couple a multimode, Q-switched, laser to the fiber optic transmission system. Characterization of the intensity profiles produced by the diffracting beam integrator using various laser sources and primary lens focal lengths illustrate those diffraction, interference, and averaging effects discussed in Sec. II.

The optical components and their specifications that form the basis of the fiber injection system are shown in Fig. 14. This system comprises a multi-aperture, refractive, lenslet element fabricated by MEMS Optical Inc. and a primary injection lens that overlaps or integrates the beamlets from each subaperture at the lens focal plane. Given the set of specifications shown in Fig. 14 and the design equations presented earlier in this chapter the functional injection parameters are calculated and summarized for the reader in Table 1.



Input Source

Wavelength: 1061nm
Diameter (D): 5.0mm
Divergence: 2mR (Full Angle)

Primary Injection Lens

Type: Plano/Convex
Material: Fused Silica
Focal Length (F): 17.1mm

Lenslet Array

Fill Factor: >98% (Hexagonal)
Subaperture (d): 1.25mm
Focal Length (f): 68.95mm

Integrator Output

Spot Size (S): 0.310mm
Fiber Size (Core): 0.365mm
Fiber NA: 0.22

Figure 14 Optical architecture and design specifications for the diffracting beam integrator and fiber injection system.

Table 1 Calculated injection parameters for the diffracting beam integrator

Equation	Calculated value	Measured value
$\beta = \frac{\pi dS}{\lambda F}$, Eq. (3)	67	NA
$S = \frac{F}{f/d}$, Eq. (4)	310 μm	353 μm
$Fresnel = \frac{dS}{4\lambda F}$, Eq. (7)	5.34	Modeled
$Period = \frac{\lambda F}{d}$, Eq. (14)	14.3	16.3 μm

Verification of the design equations and illustration of the lenslet's performance is based on capturing and characterizing intensity profiles distributed along the optical axis of the primary injection lens. The collection of these intensity images was accomplished using a CCD camera* configured with a suitable objective lens to achieve an approximate 10× magnification. Post processing and manipulation of the image data was performed using analysis software.† With this analysis package a qualitative comparison of the peak intensity value can be made for the various profiles presented in the sections that follow. This figure of merit, referred to as the peak-to-average (P/A) value, is defined as the ratio of the peak pixel intensity count to the average pixel intensity count within a user defined diameter. Subsequently, qualitative comparisons of intensity profiles are made by maintaining a consistent user defined diameter for all cases in which the peak and average pixel counts are calculated. (All P/A values presented herein are based on a 365 μm diameter positioned at the centroid of the imaged intensity profile.) With this definition a perfect flat top intensity profile extending over the entire 365 μm diameter would have a P/A value of unity – the theoretical ideal for most fiber injection applications.

1. Diffraction and Interference Effects

The theory of the effects of laser coherence on the performance of the diffracting beam integrators discussed in Sec. II can be illustrated by

* COHU 4800 Camera active picture pixel size – Horizontal: 23.0 μm, Vertical: 27.0 μm.
† Beamcode is an optical beam diagnostic system providing real time analysis of captured intensity profiles.

observing the intensity distribution located at the focal plane of the primary injection lens when such a system is illuminated by a variety of laser sources – representing varying degrees of spatial coherence. As described in Sec. II.A, the diffraction pattern of a single aperture of the lenslet array determines the shape of the spot on the target and the Fresnel number (or β) yields a general measure of the diffraction structure or deviation from the geometric ideal (i.e. flat top profile). In addition, it was shown that both diffraction and interference effects play an integral role in determining the ultimate intensity distribution. For example, a source that is characterized as "highly spatially coherent" produces an intensity pattern dominated by interference effects characterized by large intensity fluctuations (or spikes) with a well-defined periodicity at the lens focal plane – while sources having a "lesser degree" of spatial coherence exhibit a mixture of both interference and diffraction effects.

Three different laser sources were used to "separate" and better illustrate the effects of diffraction and interference on the ultimate performance of the diffracting beam integrator described in Fig. 14 and Table 1. Although each laser source that we evaluated presents a different spatial intensity pattern at the input to the lenslet element, it is the spatial coherence of the source that appears to dominate the resulting profiles at the injection lens focal plane (23).

Using a multimode, Q-switched, Nd:YAG laser (Laser Photonics Model YQL-102), diffraction effects are clearly distinguishable in the intensity profiles shown in Fig. 15a. Visible in the intensity profiles are the shape of the lenslet elements (hexagonal) and the expected diffraction pattern. With this laser source lenslet diffraction dominates the intensity structure and the effects of the Fresnel number can be experimentally observed and analytically verified. Overlaying the calculated one-dimensional diffraction profile with the experimental results displayed in Fig. 15b, reveals excellent agreement with the measured data. This suggests that the structure observed is indeed dominated by lenslet diffraction effects.

At the other performance extreme, a CW Cr:Nd:GSGG, TEM_{00} laser (AMOCO Model 1061-40P) was used to illustrate the interference effects resulting from a source that can be characterized as highly spatially coherent. As expected, the intensity profile shown in Fig. 16a displays both diffraction and interference effects. The hexagonal shape of the lenslet elements and a slowly varying intensity modulation earlier attributed to diffraction effects (see Fig. 15b) are apparent, however, the intensity profile is clearly dominated by narrow spikes – indicative of interference effects. Confirmation that these features are a consequence of interference effects can be made by comparing the calculated and measured periodicity of this

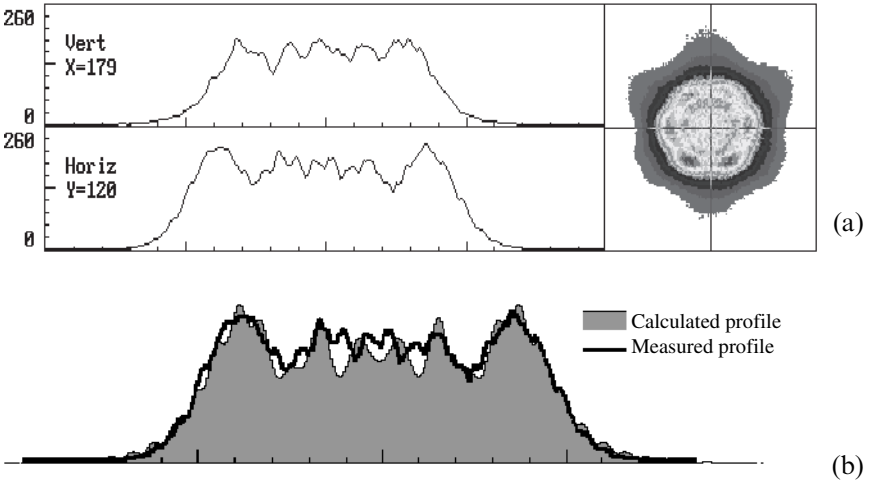
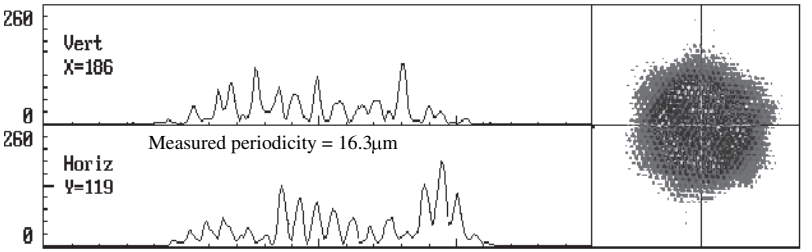


Figure 15 (a) Measured intensity profile at the focal plane of the diffracting beam integrator with a low spatially coherent source ($P/A = 2.50$, source laser: Laser Photonics YQL-102). (b) Comparison of measured and calculated intensity profile at the focal plane of the diffracting integrator.

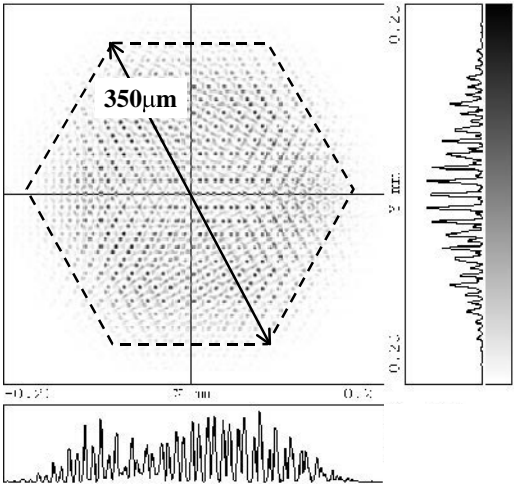
structure (see Table 1). It should be noted that Eqs. (9), (13), and (14) derived for the periodicity of the interference pattern are based on paraxial approximations and assume a one-dimensional lenslet pattern. The calculated intensity profile and its periodicity can be further studied to include the effects of both the hexagonal lenslet array structure and system aberrations using more advanced optical modeling software. Analytical results, obtained using an optical modeling package,* shown in Fig. 16b reveals the effects of the lenslet geometry and system aberrations on the intensity profile recorded at the primary lens focal plane. Assuming a perfectly coherent source, the model displays the expected two-dimensional dependence for both the interference periodicity (14.2 μm) and intensity.

For the condition of partial spatial coherence a third laser source was evaluated against the diffracting beam integrator described in Fig. 14. In this case, the source was a "home built", multimode, Q-switched, Cr:Nd:GSGG laser. (This same laser was used extensively to characterize other performance parameters of the beam integrator further described in Ref. 23). Illumination of the diffracting beam integrator with this source produces

* Advanced System Analysis Program – optical software program for geometrical and physical modeling. ASAP is a trademark of Breault Research Corporation.



(a)



(b)

Figure 16 (a) Measured intensity profile at the focal plane of the diffracting integrator with a high spatially coherent source ($P/A = 8.08$, source laser: AMOCO 1061-40P). (b) Calculated intensity profile at the focal plane of the diffracting beam integrator using the ASAP* modeling software (period = 14.2μm).

the intensity profile shown in Fig. 17 at the primary lens focal plane. Close observation of this profile illustrates effects of both diffraction and interference. The spot geometry and Fresnel structure displayed earlier in Fig. 15 are visible, however, superimposed on the diffraction pattern are narrow, high intensity "spikes." Comparing the periodicity of the high-frequency structure observed in Fig. 17 with that shown in Fig. 16 implies that these features are indeed generated by interference effects. Moreover, comparing the peak-to-average (P/A) pixel response from Figs. 15, 16, and 17 (2.50, 8.08, and 3.19) further suggests that for this beam shaping configuration interference effects will be a major contributor in determining the localized peak intensity at the intended target (23).

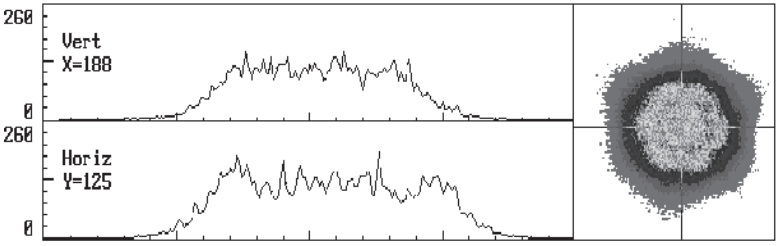


Figure 17 Measured intensity profile at the focal plane of the diffracting beam integrator with a partial spatially coherent source ($P/A = 3.19$, source laser: Q-switched, Cr:Nd:GSGG).

2. Spot Diameter and Averaging

Characterization of the intensity profiles collected along the optical axis of the diffracting beam integrator described in Fig. 14 yields a location and estimate for the minimum beam diameter and further describes how the profile evolves past the lens focal plane. The spot size as a function of distance along the optical axis is plotted in Fig. 18. In this case a functional description of the spot size is defined as the diameter in which 98% of the energy is contained. As expected, the minimum spot size is achieved at the

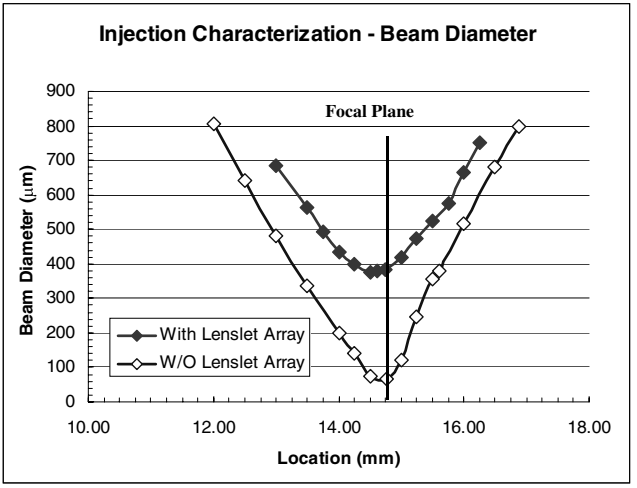


Figure 18 Measured beam diameter along the optical axis with and without the lenslet array (injection NA = 0.17).

focal plane of the injection lens and that the growth in the beam diameter behind this plane is described by the paraxial approximation given by the input beam diameter and the primary lens focal length (i.e. *f*-number). Comparing the beam diameter calculated in Table 1 (310 m) to the measured minimum value shown in Fig. 18 (353 m) suggests reasonable agreement between the intended design and actual performance. Differences between the calculated and measured results are easily accounted for in those assumptions inherent in Eq. (4) for the calculated beam diameter (i.e. aberration free optical system and an ideal source with essential zero beam divergence) and the functional definition for beam diameter presented earlier. Another potentially important feature displayed in Fig. 18 is the greater effective depth of focus around the focal plane or target location for the diffracting integrator when compared to the simple lens configuration. The importance of this characteristic could become significant when the integrator is incorporated into an assembly where alignment or position insensitivity are considered desirable (i.e. fiber injection, photolithography, laser drilling, etc.). This characteristic will be discussed in more detail in the context of the compact fiber injection system presented in Sec. V.B.

The averaging aspect of the diffracting beam integrators and the subsequent insensitivity of the intensity distribution at the target plane to spatial perturbations of the input source is a highly desirable characteristic of many beam shaping systems. Once again, intensity profiles collected at the focal plane using various laser sources and different input beam diameters have been used to illustrate this feature. Applying various apertures to the input source and hence effectively exposing different near-field features to the lenslet array can yield some insight into the sensitivity of the diffracting beam integrator to these changes. The results of such an experimental characterization using the Q-switched, Cr:Nd:GSGG described earlier are shown in Fig. 19. The contour sequence displayed in Fig. 19 reveals little change in the output intensity profile (or P/A value) recorded at the lens focal plane until the input beam diameter is comparable in size to the diameter of a single lenslet element on the array. When this occurs, the primary injection lens no longer ''overlaps'' or integrates inputs from multiple lenslet elements. In the limit of illuminating a single lenslet, this effectively yields the spatial profile that would result using just the primary injection lens slightly defocused from its focal plane. (This slight ''defocus'' is a result of the non-zero optical power provided by the lenslet element.)

B. Compact Fiber Injection System

Various system applications exist in which the advantages offered by the multi-aperture integration systems discussed in the chapter can readily be

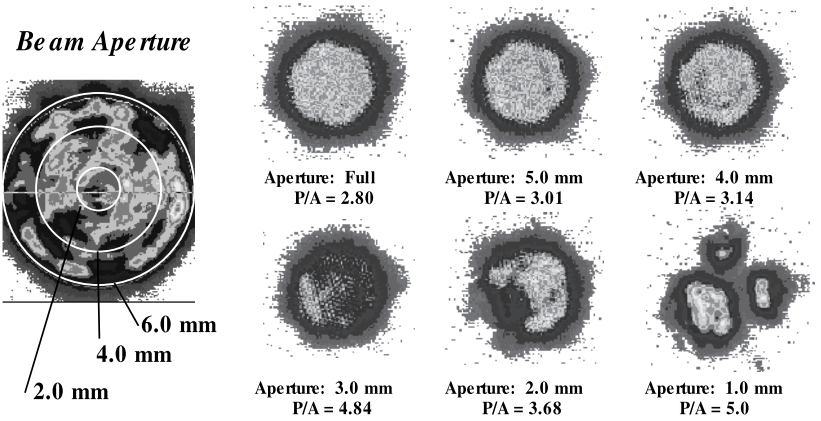


Figure 19 Measured intensity profiles at the focal plane of the diffracting beam splitter with various apertures at the source laser.

applied. Specifically, a number of medical and industrial laser applications exist in which the limiting factor for performance or additional capability resides in the transmission of the optical power (or energy) from the source to the target. Typical transmissions systems for these applications include the use of fiber optics or "open air" designs requiring numerous optical elements to direct and "reshape" the laser radiation along the intended path to the target. Depending on the specific application and configuration (i.e. target accessibility, laser wavelength, peak power, etc.) the use of fiber optics is often the preferred transmission system. Although the use of optical fibers provides the system designer and end user effective control over a number of application parameters (i.e. beam diameter, spatial profile, target accessibility, etc.), it is often limited by the maximum power level that can be reliably injected into and ultimately transmitted to the target.

Maximizing power throughput and minimizing fiber damage requires a thorough understanding of fiber damage mechanisms and the control of a number of fiber injection criteria (24,25). In contrast to the "simple" injection lens, effective control of a number of these parameters is provided by the diffracting beam integrator discussed in Sec. V.A and shown in Fig. 14.

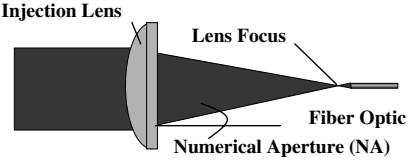
Performance characterization of a compact fiber injection system featuring the diffracting beam integration approach has recently been completed at the author's laboratory.* For this application, geometrical

* Sandia National Laboratories – Sandia is a multi-program laboratory operated by Lockheed Martin for the United States Department of Energy under Contract DE-AC04-94AL85000.

and mechanical constraints were significant factors in determining the optical architecture employed to effectively integrate and optically couple a miniaturized, Q-switched laser source with the desired fiber optic transmission system. Mechanically constrained in the overall length of the fiber injection and alignment system to less than 25 mm, more complex beam shaping techniques were quickly abandoned and the more traditional simple injection lens approach was initially evaluated. However, as Fig. 20 summarizes, the simple lens approach for high-power applications imposes a number of functional limitations on the power or energy that can be reliably injected and transmitted by the optical fiber. With a single injection lens element, the ultimate limitations and performance variability are in practice a combination of the low air breakdown threshold and the strong interdependence of the laser source characteristics and injection alignment to the peak optical fluence incident on the fiber face. It is worth noting that these issues are further exasperated by the use of short focal length lenses and high peak power, multimode lasers – both conditions inherent in this application.

Effective control of a number of those fiber injection issues shown in Fig. 20 are provided by the diffracting beam integrator evaluated and discussed in Sec. II. Beyond those results presented earlier illustrating the effects of diffraction, interference, and spatial averaging on the intensity profile at the target (or fiber) plane, a brief performance comparison of the simple injection lens and the diffracting beam integrator is discussed in the following material.

With the simple injection lens approach the conditions imposed to avoid air breakdown and provide "adequate" filling of the fiber NA are diametrically opposed. Avoiding air breakdown with the simple lens implies controlling the minimum spot size and hence the lens focal length.



- Low air breakdown thresholds - ***all*** energy focused at ***single*** foci
- Spatial profile @ fiber strongly ***dependent*** on input laser and alignment
- Beam diameter @ fiber face is ***strong*** function of lens displacement
- NA of injection system ***limited*** by need to avoid air breakdown
- Exit profile and potential for damage internal to the fiber ***dependent*** on the initial Mode Power Distribution (MPD) incident on the entrance face

Figure 20 Fiber injection issues using the simple plano-convex injection lens.

Unfortunately, as the lens focal length is increased to accommodate a higher threshold breakdown, the injected entrance angle or NA to the fiber is correspondingly decreased. In contrast, the diffracting integrator distributes the input laser energy over a larger cross sectional area and into multiple foci representing the number of active lenslet elements. Consequently, eliminating air breakdown allows the designer the freedom to select the primary lens focal length to better match the acceptance angle of the optical fiber. However, as discussed in Sec. II.A, as the focal length of the primary lens is varied the f-number of the lenslets must be correspondingly adjusted to maintain the desired spot size per Eq. (4). Maintaining the spot size is achieved at the expense of changing β or the Fresnel number defined in Eqs. (3) and (7) respectively. It is the interdependence of the lens focal length and β that must be optimized given the specifics of the desired injection geometry.

Another important characteristic provided by the diffracting beam integrator in the evaluated fiber injection system is the large depth of field discussed in Sec. V.A.2 (Fig. 18) and the alignment insensitivity to the input source that is a result of the lenslet's spatial averaging behavior (Fig. 19). The large depth of field enables the designer to use a very simple mechanical mount that is required to provide only gross adjustment capabilities along the lens optical axis.

A more troublesome design issue with fiber injection systems and a parameter substantially relaxed by the use of the diffracting beam integrator is the alignment or centralization of the input beam on the primary injection lens. As shown in Fig. 21, when the input beam is displaced across the

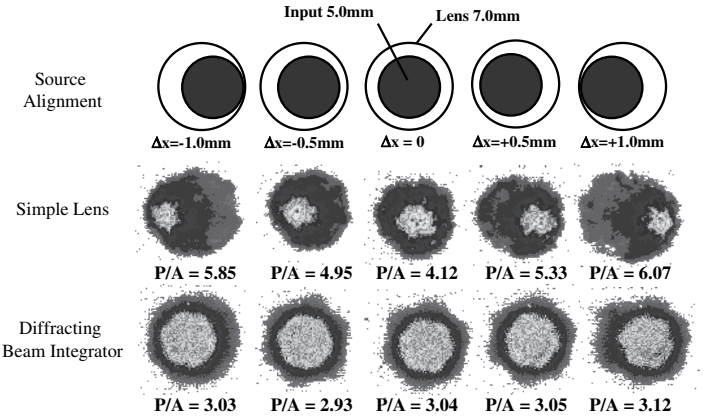


Figure 21 Measured intensity profiles at the fiber face with lateral misalignment of the source.

injection lens without the lenslet array the spatial intensity profile incident at the fiber face is strongly affected by system aberrations – resulting in higher peak fluences physically decentered from the fiber core. In contrast, when the lenslet array is added the spatial profile (or P/A value) remains unchanged over a wide range of lateral misalignment. Once again insensitivity to alignment significantly relaxes the mechanical requirements for orienting the laser source to the fiber injection system.

VI. SUMMARY

In this chapter we have presented the theory and design of multi-aperture beam integration systems. These systems are especially applicable to the shaping of multimode laser beams that are characterized by an irregular irradiance pattern that frequently varies with time. The major assumptions applicable to the design and analysis of a multi-aperture beam integration system are stated explicitly.

As discussed in Chapter 2 and 3, β, or equivalently the Fresnel number, is again an important parameter in determining the performance of the beam shaping system. The basic concepts and equations needed for system design are developed. Diffraction and interference effects, associated with multi-aperture beam integration systems are treated in detail. The impact of the input beam spatial coherence on beam integrator performance is analyzed.

Experimental data illustrating the effects of diffraction, interference and spatial coherence on beam integrator performance is presented. Finally, we discuss the design and present data for a diffractive multi-aperture beam integration system for optical fiber injection.

ACKNOWLEDGMENT

The authors would like to acknowledge Richard N. Shagam for ASAP modeling analysis and Stephen E. Yao and Jeremy D. Brown for their help with manuscript preparation.

REFERENCES

1. HJ Kahlert, U Sarbach, B Burghardt, B Klimt. Excimer laser illumination and imaging optics for controlled microstructure generation. SPIE Proc 1835: 110–112, 1992.

2. L Unnebrink, T Henning, EW Kreutz, R Poprawe. Excimer laser beam homogenization for materials processing. SPIE Proc 3573: 126–129, 1998.
3. JR Leger, WC Goltsos. Method and apparatus for efficient concentration of light from laser diode arrays. US Patent No. 5 787 107, July 28, 1998.
4. J Endriz. Diode laser source with concurrently driven light emitting segments. US Patent No. 5 594 752, January 14, 1997.
5. FM Dickey, BD O'Neil. Multifaceted laser beam integrators: general formulation and design concepts. Optical Engineering 27(11): 999–1007, 1988.
6. DA Pepler, CN Danson, IN Ross, S Rivers, S Edwards. Binary-phase Fresnel zone plate arrays for high-power laser beam smoothing. SPIE Proc 2404: 258–265, 1995.
7. K Kamon. Fly-eye lens device and lighting system including same. US Patent No. 5 251 067, October 5, 1993.
8. JW Goodman. Introduction to Fourier Optics. New York: McGraw-Hill, 1968.
9. L Unnebrink, T Henning, EW Kreutz, R Poprawe. Optical system design for excimer laser materials processing. SPIE Proc 3779, 1999.
10. Commercial lens design software sold by Focus Software, Inc, Tucson, AZ 85731-8228.
11. KF Cheung, RJ Marks II. Imaging sampling below the Nyquist density without aliasing. J Optical Soc America A 7(1): 92–105, 1990.
12. JW Goodman. Statistical Optics. New York: John Wiley & Sons, 1985.
13. T Henning, L Unnebrink, M Scholl. UV laser beam shaping by multifaceted beam integrators: fundamental principles and advanced design concepts. SPIE Proc 2703: 62–73, 1996.
14. X Deng, X Liang, Z Chen, W Yu, R Ma. Uniform illumination of large targets using a lens array. Applied Optics 25: (3), 377–381, 1986.
15. J Bernges, L Unnebrink, T Henning, EW Kreutz, R Poprawe. Mask adapted beam shaping for material processing with eximer laser radiation. SPIE Proc 3573: 108–111, 1998.
16. B Burghardt, HJ Kahlert, U Sarbach., Device for homogenizing a light beam. US Patent No. 5 414 559, May 9, 1995.
17. HJ Kahlert, B Burghardt. Optical apparatus for the homogenization of laser radiation and the generation of several lighting fields. US Patent No. 5 796 521, August 18, 1998.
18. T Henning, M Scholl, L Unnebrink, U Habich, R Lebert, G. Herziger. Beam shaping for laser materials processing with non-rotationally symmetric optical elements. SPIE Proc 3209, 126–129, 1996.
19. S Sinzinger, J Jahns. Microoptics. Weinheim: Wiley-VCH, 1999.
20. M Kufner, S Kufner. Micro-optics and Lithography. Brussels: VUBPress, 1997.
21. T Mori, H Komatsuda. Optical integrator and projection exposure apparatus using the same. US Patent No. 5 594 526, January 14, 1997.
22. H Komatsuda, H Hirose, T Kikuchi. Illumination device with allowable error amount of telecentricity on the surface of the object to be illuminated and exposure apparatus using the same. US Patent No. 5 594 587, January 14, 1997.

23. L Weichman, F Dickey, R Shagam. Beam shaping element for compact fiber injection systems. SPIE Proc 3929, 2000.
24. R Setchell. An optimized fiber delivery system for Q-switched, Nd:YAG lasers. SPIE Proc 2966: 608, 1997.
25. R Setchell. End-Face preparation methods for high intensity fiber applications. SPIE Proc 3578: 743, 1998.

8

Classical (Non-laser) Methods

David L. Shealy
University of Alabama at Birmingham, Birmingham, Alabama

I. INTRODUCTION

In this chapter, the design and analysis of non-laser optical systems used for beam shaping will be discussed. Geometrical optics is used to evaluate the irradiance throughout the optical system. A method based on differential equations is presented for evaluating the contour of an optical surface (mirror or lens) that will transform a given input beam profile into a specified output beam profile or irradiance distribution over a detector surface. Non-laser beam shaping differs from laser beam shaping in two major ways. First, it deals with more general sources, such as Lambertian sources, line sources, or LEDs. Second, it is only concerned with providing a certain irradiance distribution at a particular surface. It is not concerned with propagating a beam beyond that surface or form of the irradiance distribution at intermediate points.

Early thoughts of beam shaping in non-laser systems can be traced to before the days of Archimedes and his burning glass (1,2), where optics was reported to concentrate – to increase the power density of – solar radiation. The literature is rich with reports of various optical systems used as solar collectors (3–9). Welford and Winston (10) have presented a good accounting of non-imaging (non-focusing) optics used as solar collectors, including an ideal light collector (11,12), which concentrates a beam by the maximum amount allowed by phase space considerations. Burkhard and Shealy (13) have used a differential equation method to design a reflecting surface, which distributes the irradiance over a receiver surface in a prescribed manner. McDermit and Horton (14,15) presented a generalized technique for designing a rotationally symmetric reflective solar collector, which can heat

the collector surface in a prescribed manner. Beam shaping has also been used in opto-electronics to achieve maximum power transfer between a micro-optics light source and an optical fiber (16,17), in radiative heat transfer (18–20) in illumination applications (21–24), and for reflector synthesis (25,26).

In the next section, ray tracing and the flux flow equation are discussed within the context of design and analysis of non-laser beam profile shaping. For analysis, the flux flow equation can be used to compute the irradiance over any surface within an optical system. For design, the flux flow equation has been inverted to give a differential equation for the shape of one optical surface of the system, when the input and output beam profiles are known. Section III discusses application of this design method for a point and Lambertian source of radiation.

II. THEORY OF NON-LASER BEAM PROFILE SHAPING

For non-laser systems, a typical beam shaping system design goal is to illuminate a detector or substrate surface with a specified irradiance distribution. For laser-based systems, Chapter 4 (Sec. II.A) demonstrates that shaping an irradiance profile can be achieved by using the energy balance condition to determine the geometrical shape of one optical surface (lens, mirror, or GRIN profile) of the system. In addition, laser beam shaping applications often seek to retain the output irradiance profile as the beam propagates (a collimated beam). This can be achieved by requiring that the system has a constant optical path length between the input and output surfaces, as demonstrated in Chapter 4, Sec. II.B. This propagation constraint is not needed in non-laser applications. The following theory, then, extends the laser analysis of Chapter 4 to the more general sources found in non-laser systems.

Ray tracing (27) is widely used to simulate the performance of both imaging and non-imaging optical systems. By assigning to each incoming ray equal energy density, then by counting the number of rays crossing a unit of area within the optical system, the irradiance can be computed throughout the optical system. Kock (28) reports a method to simplify photo-radiometric calculations of optical systems by using a reference sphere and ray tracing. The flux flow equation (29,30) offers an alternate approach for evaluating the irradiance within an optical system. The flux flow equation along with the ray trace equations are used to monitor the change in size of an element of area of a bundle of rays (31) as the wavefront propagates through the optical system. The flux flow equation depends on the beam parameters and the shape/orientation of the optical surfaces and

allows the irradiance to be computed along a ray path as it propagates through an optical system. The flux flow equation can also be considered as a differential equation of the optical surface contour, which can be solved if the input and output beam profiles are known. In the next section, conservation of energy within a bundle of rays is used to derive two alternate expressions of the flux flow equation. Then, in Sec. II.B, the flux flow equation is used to formulate a method based on differential equations for design of non-laser beam shaping systems.

A. Irradiance (Illuminance) Analysis with the Flux Flow Equation

In this section, a formula is derived for the flux density (irradiance or illuminance, which is the energy per unit area per unit time) of a ray passing through an optical system. This formula, which has been labeled the flux flow equation in the literature, can be expressed as the ratio of the products of the principal radii of curvature* of the wavefront as it approaches and leaves an optical surface. The principal radii of curvature and torsion of the incident and reflected wavefront are related to similar quantities of the reflecting surface using a generalization of the Coddington equations (32) – a procedure also known as generalized ray tracing (33).

Assume that the flux density incident upon an optical interface surface s along the direction of a particular ray is denoted by $\sigma(\mathbf{r})$. If i is the angle of incidence of a ray upon the element of surface area da on s, the total flux incident upon da is given by

$$dF = \sigma(\mathbf{r}) \cos i \, da. \tag{1}$$

The radiation is reflected (or refracted) to the element of area dA on surface S as shown in Fig. 1. Then, flux density over the surface S is given by

$$\frac{dF}{dS} \equiv E_{ds-dS} = \sigma(\mathbf{r})\tau(\mathbf{r}) \cos i(\mathbf{r}) \left(\frac{da}{dA}\right) \tag{2}$$

where $\tau(\mathbf{r})$ is the reflection or transmission coefficient at the point $\mathbf{r}$ on surface s. If the radiation incident upon surface S is reflected or refracted to another surface, then the process of evaluating the flux density along a ray path within an optical system can be generalized to an arbitrary number of surfaces (29). The problem now is to evaluate the ratio of elements of surface areas, (da/dA), along a ray path.

* See Appendix A for a discussion of the principal radii of curvature of a surface and other basic concepts from differential geometry and the theory of surfaces.

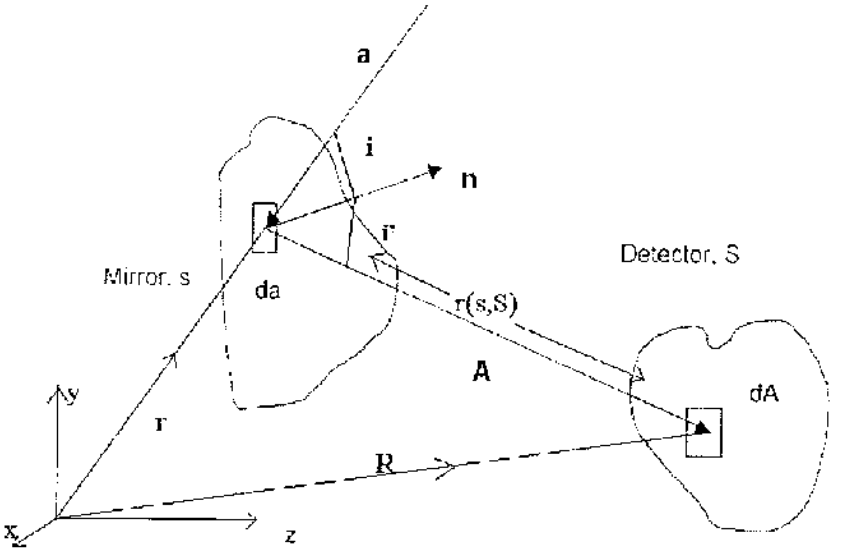


Figure 1 Illustration of beam reflected from surface s to surface S.

The ray trace equations, Chapter 4, Eq. (17) can be regarded as a coordinate transformation between the elements of surface area *da* and *dA*. Then, the Jacobian, *J*, of this transformation enables the connection $dA = J\,da$. This approach has been used by Shealy and Burkhard (30,19). A simpler approach follows by recognizing that the flux is carried by the wavefront (32). From conservation of energy along a bundle of rays, Chapter 4, Eq. (10), an element of area of the wavefront must be related to an element of area on the optical surfaces s and S. Equation (2) may then be replaced by either a quadratic equation in the distance $r(s, S)$ between *da* and *dA* along the ray path or an equation involving the wavefront elements of area before and after reflection (or refraction), which in turn may be replaced by an expression involving wavefront curvatures.

The position vector of *dA* along the ray after it leaves *da* is given by

$$\mathbf{R} = \mathbf{r}(u, v) + r(s, S)\mathbf{A}(u, v) \tag{3}$$

where

$$\mathbf{r}(u, v) = \hat{\mathbf{i}}\, x(u, v) + \hat{\mathbf{j}}\, y(u, v) + \hat{\mathbf{k}}\, z(u, v) \tag{4}$$

is the equation of the mirror surface s expressed in terms of the parameters (u, v). The quantity $r(s, S)$ is the distance along the ray measured from the

point of reflection da on surface s to dA on surface S. $\mathbf{A}(u, v)$ is the unit vector along the reflected ray and is related to the incident ray vector $\mathbf{a}$ by

$$\mathbf{A} = \mathbf{a} - 2\mathbf{n}(\mathbf{a} \cdot \mathbf{n}), \tag{5}$$

where $\mathbf{n}$ is the unit normal vector to the mirror. Equation (3) can also be viewed as an equation of the reflected wavefront in the vicinity of the ray, when the constant phase condition as measured from the source is considered.

To map a bundle of rays continuously across an optical surface s, the following relations hold

$$da \cos i = dW(s), \tag{6}$$

$$da \cos i' = dW'(s) \tag{7}$$

where i' is the angle of reflection (or refraction), $dW(s)$ is an element of area on the wavefront incident upon da, and $dW'(s)$ is an element of area on the wavefront reflected from da. At surface S, similar relations hold:

$$dA \cos I = dW(S), \tag{8}$$

$$dA \cos I' = dW'(S). \tag{9}$$

Now, derivations for two alternate expressions of the flux flow equation will be presented, which involve evaluating the ratio (da/dA) – the first case – in terms of the partial derivatives of the direction cosines of the reflected ray vector $\mathbf{A}$ and the equation of the mirror surface s. In the second case, the flux flow equation is expressed in terms of the principal curvatures of the incident and reflected wavefront at S and s, respectively.

1. Flux Flow Equation – First Case

The first expression for the flux flow equation is obtained by evaluating the ratio (da/dA) from the equation of the surface and the ray trace equation relating da to dA. An element area on the surface s is equal to the magnitude of the vector cross product of the independent surface-tangent vectors

$$da = |\mathbf{r_u} \times \mathbf{r_v}| \, du \, dv = \sqrt{g} \, du \, dv \tag{10}$$

where

$$g = g_{uu} g_{vv} - g_{uv}^2, \tag{11}$$

$$g_{uu} = \mathbf{r}_u \cdot \mathbf{r}_u, \quad g_{vv} = \mathbf{r}_v \cdot \mathbf{r}_v, \quad g_{uv} = \mathbf{r}_u \cdot \mathbf{r}_v \tag{12}$$

$$\mathbf{r}_u \equiv \left(\frac{\partial \mathbf{r}(u, v)}{\partial u} \right) \quad \text{and} \quad \mathbf{r}_v \equiv \left(\frac{\partial \mathbf{r}(u, v)}{\partial v} \right). \tag{13}$$

See Appendix A for derivation of Eq. (10). From Eq. (8), the element of area dA can be expressed in terms of an element of area $dW(S)$ on the wavefront incident upon S as follows:

$$dA = \frac{dW(S)}{\cos I}. \tag{14}$$

Evaluating $dW(S)$ in terms the coordinates (u, v) in similar manner as in Eq. (10) leads to

$$dW(S) = \mathbf{A} \cdot \left(\frac{\partial \mathbf{R}(u,v)}{\partial u} \right) \times \left(\frac{\partial \mathbf{R}(u,v)}{\partial v} \right) du\, dv \tag{15}$$

where the magnitude of an element of area on the wavefront is obtained by projecting the vector cross product along the direction of the ray vector $\mathbf{A}$, which is also normal to the wavefront. Using Eq. (3) to simplify Eq. (15) leads to the following:

$$\begin{aligned} dW(S) = \{ & \mathbf{A} \cdot [\mathbf{r}_u \times \mathbf{r}_v] + r(s,S)\mathbf{A} \cdot [\mathbf{r}_u \times \mathbf{A}_v + \mathbf{A}_u \times \mathbf{r}_v] \\ & + r^2(s,S)\mathbf{A} \cdot [\mathbf{A}_u \times \mathbf{A}_v] \}\, du\, dv \end{aligned} \tag{16}$$

where the subscripts (u, v) of vectors $\mathbf{r}$ or $\mathbf{A}$ represent the partial derivatives with respect to u or v, as defined in Eq. (13). Putting Eq. (16) into Eq. (14) and using this result with Eq. (10) leads to the following expression for the ratio of (da/dA):

$$\frac{da}{dA} = \frac{\cos I}{\left\{ \mathbf{A} \cdot [\mathbf{r}_u \times \mathbf{r}_v] + r(s,S)\mathbf{A} \cdot [\mathbf{r}_u \times \mathbf{A}_v + \mathbf{A}_u \times \mathbf{r}_v] + r^2(s,S)\mathbf{A} \cdot [\mathbf{A}_u \times \mathbf{A}_v] \right\} / \sqrt{g}}. \tag{17}$$

Define the denominator of Eq. (17) to be $\mathscr{L}(s)$

$$\mathscr{L}(s) \equiv L_0(1) + r(s,S)L_1(1) + r^2(s,S)L_2(1) \tag{18}$$

where

$$L_0(1) = \mathbf{A} \cdot [\mathbf{r}_u \times \mathbf{r}_v] / \sqrt{g} = \cos i', \tag{19}$$

$$L_1(1) = \mathbf{A} \cdot [\mathbf{r}_u \times \mathbf{A}_v + \mathbf{A}_u \times \mathbf{r}_v] / \sqrt{g}, \tag{20}$$

$$L_2(1) = \mathbf{A} \cdot [\mathbf{A}_u \times \mathbf{A}_v] / \sqrt{g}. \tag{21}$$

Substituting Eq. (17) into Eq. (2) gives the following expression for the irradiance over surface S

$$E_{ds-dS} = \frac{\sigma(\mathbf{r})\tau(\mathbf{r})\cos i \cos I}{\mathscr{L}(s)} \tag{22}$$

with an immediate generalization to n surfaces (29). Equation (22) has been called the flux flow equation. So far, only the conservation of energy within a bundle of rays has been used to derive the flux flow equation. The law of reflection is introduced at each optical surface when evaluating $\mathscr{L}$. Detailed discussions of using the flux flow equation to compute irradiance distributions for reflective and refractive optical systems have been reported in Refs 34 and 35. For example, Fig. 2 shows the contours of equal irradiance over a

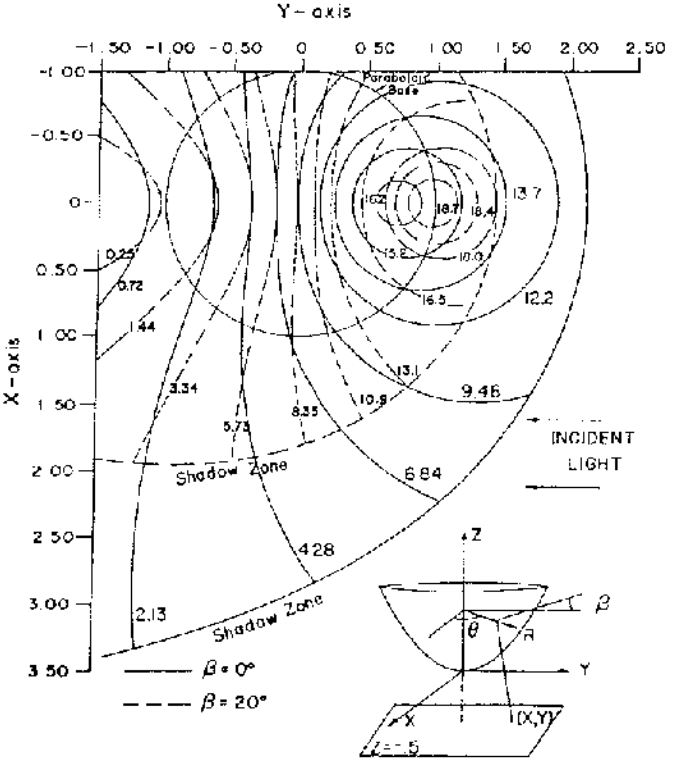


Figure 2 Contours of equal irradiance on a plane for light reflected from a paraboloid. Flux density values associated with each contour represents the percentage of light incident upon a specific region of the receiver plane. The reflection coefficient of the mirror was assumed to be one. (From Ref. 34.)

plane for light reflected from a paraboloid; Figure 3 shows the contours of equal irradiance over a plane for light reflected from a cone; and Fig. 4 shows the contours of equal irradiance over a plane for light refracted by a convex-plano lens.

Explicit expressions for the terms L_0, L_1, and L_2 in Eq. (22) have been reported in the literature for reflection or refraction within multi-interface optical systems (29). In case of reflection of collimated light from a mirror, these coefficients are given by

$$L_0(1) = \cos i' \tag{23}$$

$$L_1(1) = 4H\cos^2 i + 2K_n \sin^2 i \tag{24}$$

$$L_2(1) = 4K\cos i' \tag{25}$$

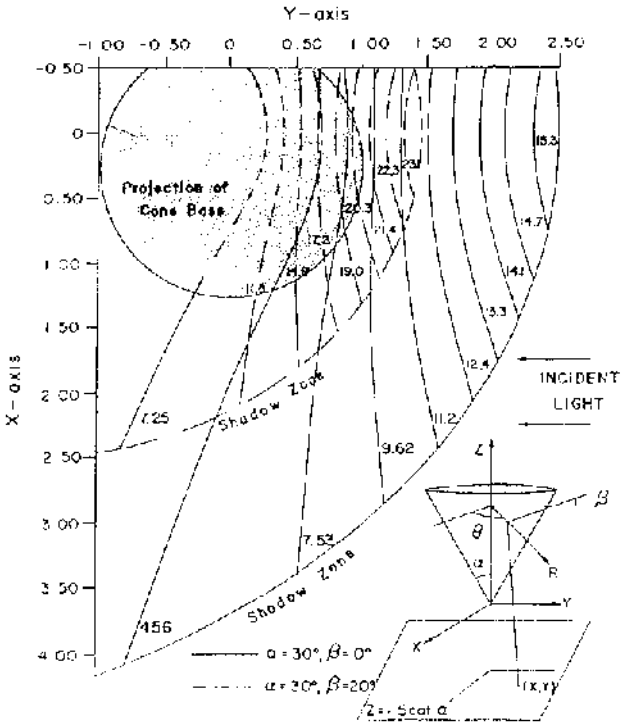


Figure 3 Contours of equal irradiance on a plane for light reflected from a cone. Flux density values associated with each contour represents the percentage of light incident upon a specific region of the receiver plane. The reflection coefficient of the mirror was assumed to be one. (From Ref. 34.)

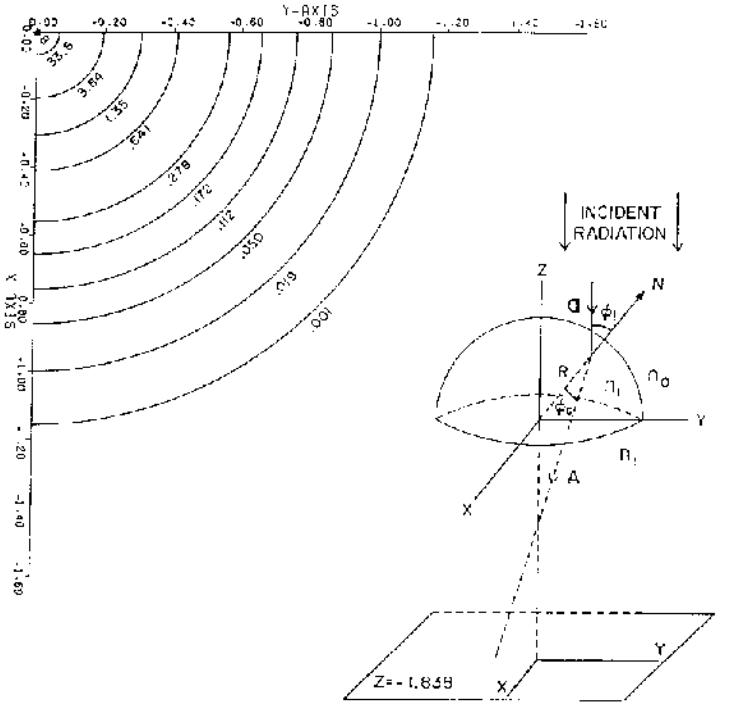


Figure 4 Contours of equal irradiance on a plane for light reflected from a plano-spherical lens of index of refraction 1.544. Flux density values associated with each contour represents the percentage of light incident upon a specific region of the receiver plane. The transmission coefficient of the lens was assumed to be one. (From Ref. 35.)

where H, K, and K_n are the mean, Gaussian, and normal curvatures, respectively, of the mirror and can be expressed in terms of the equation of the mirror surface and the direction of the incident light using the following expressions:

$$K = \frac{[b_{uu}b_{vv} - b_{uv}^2]}{g} \tag{26}$$

$$H = \frac{[g_{uu}b_{vv} - 2g_{uv}b_{uv} + g_{vv}b_{uu}]}{2g} \tag{27}$$

$$K_n \sin^2 i = \sum_{i,j=u}^{v} a^i a^j b_{ij} \tag{28}$$

$$b_{uv} = g^{-1/2}\left|\left(\frac{\partial \mathbf{r}}{\partial u}\right) \times \left(\frac{\partial \mathbf{r}}{\partial v}\right) \cdot \left(\frac{\partial^2 \mathbf{r}}{\partial u\, \partial v}\right)\right| \tag{29}$$

$$a^i = g^{i1}\mathbf{a} \cdot \left(\frac{\partial \mathbf{r}}{\partial u}\right) + g^{i2}\mathbf{a} \cdot \left(\frac{\partial \mathbf{r}}{\partial v}\right) \tag{30}$$

$$g^{uu} = \frac{g_{vv}}{g}; \quad g^{vv} = \frac{g_{uu}}{g}; \quad g^{uv} = -\frac{g_{uv}}{g}. \tag{31}$$

For a brief discussion of some concepts from differential geometry, see Appendix A or Ref. 36. Alternate formulas for these curvatures will be presented in Sec. II.B when the flux flow equation is used to design a rotationally symmetric optical system for non-laser beam shaping.

Combining Eqs. (18), (23)–(25) with Eq. (22) leads to the following expression for the flux flow equation for reflection of collimated light from a mirror:

$$E_{ds-dS} = \frac{\sigma(\mathbf{r})\rho(\mathbf{r}) \cos i \cos I}{|\cos i' + 2r(s,S)\left[2H \cos^2 i + K_n \sin^2 i\right] + 4r^2(s,S)K \cos i'|}. \tag{32}$$

Fock (37) has reported a similar expression for the intensity of a beam cross-section which has been reflected from a surface.

It is interesting to note that the flux flow equation (22) depends on the first and second derivatives of the equation of the optical surface with respect to the coordinates (u, v). Thus, it follows conceptually that if both the input and output beam profiles are known functions of the aperture coordinates, then the flux flow equation represents a differential equation which can be used to determine the shape of the optical surface s. This approach to optical design of non-laser beam shaping systems will be discussed in Sec. II.B.

2. Flux Flow Equation – Second Case

The second expression for the flux flow equation is obtained by expressing the ratio (da/dA) in terms of element of area on the wavefront before and after reflection or refraction. These results are then expressed in terms of the wavefront curvatures which lead to the second alternate expression of the flux flow equation. Both expressions of the flux flow equation are equivalent (32). Using Eqs. (7) and (8), (da/dA) can be written in the form

$$\frac{da}{dA} = \frac{\cos I}{\cos i'} \frac{dW'(s)}{dW(S)}. \tag{33}$$

Then, from Eq. (2) the flux density over surface S can be written as

$$E_{ds-dS} = \sigma(\mathbf{r})\tau(\mathbf{r}) \frac{\cos i \ \cos I}{\cos i'} \frac{dW'(s)}{dW(S)}. \tag{34}$$

We will show (justify) that an element of area on the wavefront can be expressed in terms of the principal radii of curvature of the surface. The principal radii of curvature of an element of area on a surface are the maximum and minimum curvatures of the surface at that point*. They are found by taking a plane through the surface normal at a point and rotating it around the normal. The intersection of the plane and the surface forms curves.

Born and Wolf have shown that there are two focal (imaging) points for each point on a wavefront (38). The caustic surface has also been defined (32) to be the loci of the focal (imaging) points of an optical system. The irradiance, as computed from the flux flow equation (see Eq. (22), for example), is infinite on the caustic surface, which occurs in geometrical optics when $dA = 0$. Points on the caustic surface for rays reflected from a mirror surface s are computed by setting $\mathscr{L}(s)$ in Eq. (22) equal to zero and solving the resulting quadratic equation for $r(s, S)$

$$\mathscr{L}(s) = L_0(1) + r(s, S)L_1(1) + r^2(s, S)L_2(1) = 0. \tag{35}$$

The two roots of Eq. (35) are labeled $[r_1(s),\ r_2(s)]$ and represent the distance from the point of reflection to one of the focal (caustic) points on the ray from *da*. Stavroudis and Fronczek (39) have shown that the caustic points of a wavefront are the principal radii of curvature of the wavefront. Therefore, $[r_1(s),\ r_2(s)]$ are the principal radii of curvature of the wavefront as it leaves *da* on surface s. Solving the quadratic equation (35) for the distance from *ds* to the caustic surface gives

$$r_1(s) = \frac{-L_1(1) + \sqrt{L_1^2(1) - 4L_0(1)L_2(1)}}{2L_2(1)} \tag{36}$$

$$r_2(s) = \frac{-L_1(1) - \sqrt{L_1^2(1) - 4L_0(1)L_2(1)}}{2L_2(1)} \tag{37}$$

where the $+(-)$ of the radical in Eqs. (36)–(37) was arbitrarily assigned to $r_1(r_2)$, respectively.

* See Appendix A for a discussion of some concepts from differential geometry which may be helpful to the reader in better understanding the physical meaning of the different curvatures of a surface discussed in this chapter.

Recognizing that $r(s, S)$ is the distance along ray path from da to dA, then it has been shown in Refs 32 and 40 that the principal radii of curvature of the wavefront as it reaches dA are given by

$$\begin{aligned} r_1(S) &= r_1(s) - r(s, S) \\ r_2(S) &= r_2(s) - r(s, S) \end{aligned} \tag{38}$$

where the optics sign convention (41) for radii of curvature* has been used in Eq. (38). Further, it has been shown in Refs 32 and 40 that

$$\frac{dW'(s)}{dW(S)} = \frac{r_1(s)r_2(s)}{r_1(S)r_2(S)}, \tag{39}$$

which permits Eq. (34) to be written in terms of wavefront principal radii of curvature:

$$E_{ds-dS} = \sigma(\mathbf{r})\tau(\mathbf{r}) \frac{\cos i \ \cos I}{\cos i'} \frac{r_1(s)r_2(s)}{r_1(S)r_2(S)}. \tag{40}$$

The generalized ray trace equations (32,33) are used to compute the principal radii of curvature of a reflected or refracted wavefront in terms of the curvatures and torsion of the incident wavefront and optical interface. References 32 and 40 provide additional details for using the flux flow equation to evaluate irradiance distributions over surfaces of an optical system.

B. Optical Design of Non-laser Illumination Systems

As noted earlier, when the input beam profile and irradiance over a receiver surface are given, the flux flow equation can be viewed as a second-order differential equation that can be solved for the contour of one optical surface. This approach to optical design will be discussed in Sec. II.B.1. Equivalently, the energy balance condition and ray trace equation can be used to obtain a first-order differential equation for the contour of one surface in the system, as discussed in Sec. II.B.2.

1. Using the Flux Flow Equation

As noted in Sec. II.A the flux flow equation (22) or Eq. (40) depends on the first and second derivatives of the surface equation $\mathbf{r}(u, v)$ of the

*The radius of curvature of a surface is positive if the center of curvature of the surface is located to the right of the vertex of the surface with respect to the optical axis, when the light is traveling from the left to the right. Otherwise, the radius of curvature of a surface is negative.

mirror as well as on the direction of incident radiation and $\sigma(\mathbf{r})$. Therefore, the flux flow equation may be considered as a differential equation for the shape of the mirror. If both the input and output beam profiles are given along with the geometrical surface parameters (boundary conditions), then the resulting differential equation can be solved for the shape of the mirror.

To illustrate this approach for design of a non-laser beam profile shaping system, consider a collimated beam with irradiance profile $\sigma(\mathbf{r})$ incident upon a rotationally symmetric mirror shown in Figs. 5 and 6.

The Fresnel reflection losses are not considered in this design approach, i.e., put $\tau(\mathbf{r}) = 1$ in the flux flow equation. Assume that the equation of the mirror surface can be written as

$$\mathbf{r}(r, \phi) = \hat{\mathbf{i}} r \cos\phi + \hat{\mathbf{j}} r \sin\phi + \hat{\mathbf{k}} z(r) \tag{41}$$

where (r, ϕ) are the polar coordinates in the x–y plane, and $\left(\hat{\mathbf{i}}, \hat{\mathbf{j}}, \hat{\mathbf{k}}\right)$ are the Cartesian unit vectors. The term $z(r)$ is an unknown function to be determined by solution of the flux flow differential equation to be written out below.

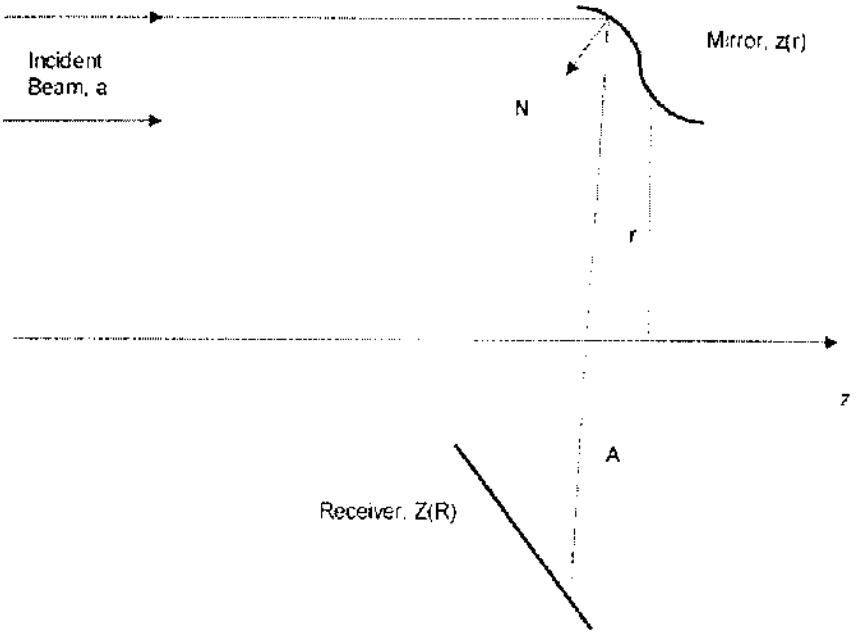


Figure 5 Collimated beam incident upon mirror and reflected to detector.

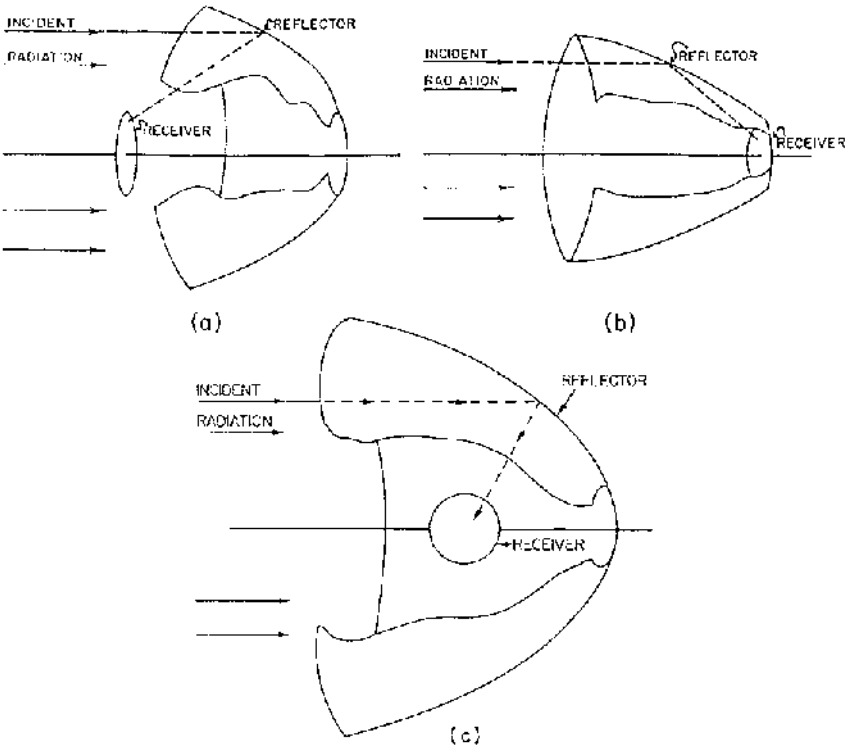


Figure 6 Rotationally symmetric solar collectors: (a) black-lighted configuration; (b) direct-lighted configuration; (c) spherical detector. (From Ref. 13.)

The metric coefficients of the mirror are computed from Eqs. (10)–(13):

$$\begin{aligned} g_{rr} &= 1 + z'^2 \\ g_{\phi\phi} &= r^2 \\ g_{r\phi} &= 0 \\ g &= r^2\left(1 + z'^2\right). \end{aligned} \tag{42}$$

The unit normal vector on the mirror surface is given by

$$\mathbf{n} = \frac{\mathbf{r}_r \times \mathbf{r}_\phi}{\sqrt{g}} = \frac{-\hat{\mathbf{i}}\, z' \cos\phi - \hat{\mathbf{j}}\, z' \sin\phi + \hat{\mathbf{k}}}{\sqrt{1 + z'^2}} \tag{43}$$

where $z' = (dz(r)/dr)$. Assuming the light incident upon the mirror is along the z-axis $\left(\mathbf{a} = \hat{\mathbf{k}}\right)$, then the cosine of the angle of incidence on the mirror is given by

$$\cos i = \mathbf{a} \cdot \mathbf{n} = \frac{1}{\sqrt{1 + z'^2}}. \tag{44}$$

The direction of reflected light from the mirror can be computed from Eq. (5):

$$\mathbf{A} = \frac{2z'\left(\hat{\mathbf{i}}\cos\phi + \hat{\mathbf{j}}\sin\phi\right) - \hat{\mathbf{k}}\left(1 - z'^2\right)}{\left(1 + z'^2\right)}. \tag{45}$$

The flux flow equation for reflection from a rotationally symmetric mirror can now be evaluated explicitly using Eq. (17) where the partial derivatives of $\mathbf{r}$ and $\mathbf{A}$ with respect to (r, ϕ) are explicitly evaluated using Eqs. (41) and (45). However, this direct method for evaluating the flux flow differential equation is very tedious and will not be discussed any further. Rather, the flux flow equation for reflection of collimated radiation from a mirror will be evaluated using Eq. (32), expressed in terms of the mean, Gaussian, and normal curvatures of the mirror, leading to a differential equation for the mirror surface.

From the theory of differential geometry of surfaces (36) and previous work (30,32), the mean, Gaussian, and normal curvature of a surface can be written in terms of the metric $g_{r\phi}$ and second fundamental form $b_{r\phi}$ coefficients of a surface by the following expressions:

$$H = \frac{\left(g_{rr}b_{\phi\phi} - 2g_{r\phi}b_{r\phi} + g_{rr}b_{\phi\phi}\right)}{2g}, \tag{46}$$

$$K = \frac{b_{rr}b_{\phi\phi} - \left(b_{r\phi}\right)^2}{g}, \tag{47}$$

$$K_n \sin^2 i = \sum_{i,j=r}^{\varphi} a^i a^j b_{ij} \tag{48}$$

where

$$b_{ij} = \frac{1}{\sqrt{g}}\left[\mathbf{r}_i \times \mathbf{r}_j \cdot \left(\frac{\partial^2 \mathbf{r}(r,\phi)}{\partial i \partial j}\right)\right] \quad \text{for } i, j = r \text{ or } \phi \tag{49}$$

$$a^i = g^{ir}\,\mathbf{a}\cdot\mathbf{r}_r + g^{i\phi}\,\mathbf{a}\cdot\mathbf{r}_\phi \tag{50}$$

$$g^{rr} = \frac{g_{\phi\phi}}{g}, \quad g^{\phi\phi} = \frac{g_{rr}}{g}, \quad g^{r\phi} = \frac{-g_{r\phi}}{g}, \tag{51}$$

or explicitly for the application shown in Figs. 5 and 6 using surface equation (41) and Eq. (42) for the metric coefficients,

$$b_{rr} = \frac{\left(\dfrac{\partial^2 z''}{\partial r^2}\right)}{\sqrt{1+z'^2}}, \quad b_{\phi\phi} = \frac{rz'}{\sqrt{1+z'^2}}, \quad b_{r\phi} = 0,$$

$$g^{rr} = \frac{1}{(1+z'^2)}, \quad g^{\phi\phi} = \frac{1}{r^2}, \quad g^{r\phi} = 0, \tag{52}$$

$$a^r = \frac{z'}{(1+z'^2)}, \quad a^\phi = 0. \tag{53}$$

where $z'' \equiv (\partial^2 z/\partial r^2)$. The mean, Gaussian, and normal curvatures of the surface are given by the following expressions:

$$H = \frac{1}{2}\left[\frac{z' + rz''}{r(1+z'^2)^{3/2}}\right], \tag{54}$$

$$K = \frac{z'z''}{r(1+z'^2)^2}, \tag{55}$$

$$K_n \sin^2 i = \frac{z'^2 z''}{(1+z'^2)^{5/2}}. \tag{56}$$

Then, the flux flow equation for collimated light reflected from a rotationally symmetric mirror can be explicitly written from Eq. (32) in the following form:

$$E_{ds-dS} = \frac{\sigma(\mathbf{r})\cos i \cos I}{\cos i\left[1 + 2r(s,S)\left(\dfrac{\left(\dfrac{z'}{r}\right) + z''}{\left(1+z'^2\right)^2}\right) + 4r^2(s,S)\left(\dfrac{\left(\dfrac{z'}{r}\right)z''}{(1+z'^2)^2}\right)\right]}, \tag{57}$$

or, after factoring the denominator,

$$E_{ds-dS} = \frac{\sigma(\mathbf{r}) \cos I}{\left[1 + \left(\frac{2r(s,S)z'}{r(1+z'^2)}\right)\right]\left[1 + \left(\frac{2r(s,S)z''}{(1+z'^2)}\right)\right]} \quad (58)$$

Using the following relationships,

$$r(s,S) = \frac{(R-r)}{A_r} = (R-r)\frac{(1+z'^2)}{2z'}, \quad (59)$$

$$\cos I = \mathbf{A}\cdot\mathbf{N} = \frac{\left[\left(\frac{2z'}{1+z'^2}\right)\left(\frac{dZ}{dR}\right) + \left(\frac{1-z'^2}{1+z'^2}\right)\right]}{\sqrt{1+\left(\frac{dZ}{dR}\right)^2}}, \quad (60)$$

where $\mathbf{N}$ is the unit normal vector to the receiver surface S, and $Z(R)$ specifies the shape of the receiver surface S. Then, the flux flow equation (58) can be written as a second-order differential equation for the mirror surface

$$\frac{z''}{z'} = \frac{1}{(R-r)}\left[\left(\frac{\sigma(\mathbf{r})}{E_{ds-dS}}\right)\left(\frac{r}{R}\right)\frac{\left[\left(\frac{2z'}{1+z'^2}\right)\left(\frac{dZ}{dR}\right) + \left(\frac{1-z'^2}{1+z'^2}\right)\right]}{\sqrt{1+\left(\frac{dZ}{dR}\right)^2}} - 1\right]. \quad (61)$$

Equation (61) is equivalent to Chapter 4, Eq. (98) and to the results of McDermit and Horton (14,42) [Eq. (3.14) of Ref. 14 and Eq. (13) of Ref. 42]. When appropriate boundary conditions are given, then Eq. (61) has been solved for the shape of the mirror surface that will illuminate the receiver surface S with a prescribed irradiance for a given source profile.

2. Using the Conservation of Energy Condition

Instead of using the second-order differential equation (61) for evaluating the contour of the beam shaping optics, the energy balance equation, Chapter 4, Eq. (10), can be integrated and combined with the ray trace equations to obtain a first-order differential equation for the reflecting surface. This approach has been used by Schruben (21) to design a mirror which illuminates its aperture with a specified distribution. This approach

is also equivalent to using the flux flow equation to obtain a second-order differential equation of the mirror surface, Eq. (61).

Consider the rotationally symmetric geometry shown in Figs. 5 and 6. The radiation is incident upon reflector surface s with equation $z = z(r)$. The equation of the receiver surface S is $Z = Z(R)$. The flux density $\sigma(r)$ is incident upon a circular ring about the z-axis of area $2\pi r\,dr$ and is reflected to a circular ring on the receiver surface S of area dA. The equation of the receiver surface S can be written in terms of the polar and radial coordinates $[\Phi, R]$

$$\mathbf{R}(\Phi, R) = \hat{\mathbf{i}} + R\cos\Phi + \hat{\mathbf{j}}R\sin\Phi + \hat{\mathbf{k}}Z(R). \tag{62}$$

Then, applying Eqs. (10) and (42) to the receiver surface S, a rotationally symmetric element of area on the receiver surface can be written as

$$dA = \int_0^{2\pi} G^{1/2}\,d\Phi\,dR = 2\pi R\sqrt{1 + (dZ/dR)^2}\,dR \tag{63}$$

where

$$\begin{aligned} G &= G_{RR}G_{\Phi\Phi} - G_{R\Phi}^2 = R^2\left[1 + \left(\frac{\partial Z(R)}{\partial R}\right)^2\right] \\ G_{RR} &= 1 + \left(\frac{\partial Z(R)}{\partial R}\right)^2 \\ G_{\Phi\Phi} &= R^2; \quad G_{R\Phi} = 0. \end{aligned} \tag{64}$$

The element of area dA in Eq. (63) represents a circular ring on the receiver surface as illustrated in Fig. 7.

When the receiver surface is a disk, as shown in Fig. 6a,b, $Z = \text{const.}$ and $dA = 2\pi R\,dR$. For a spherical receiver, as shown in Fig. 6c,

$$dA = 2\pi\left[1 + \left(\frac{R^2C^2}{1 - C^2R^2}\right)\right]^{1/2} dR = \frac{2\pi R\,dR}{\sqrt{1 - C^2R^2}} \tag{65}$$

when the equation of the spherical receiver surface is written in a form similar to that give in Chapter 4, Eq. (46),

$$\begin{aligned} Z(R) &= \frac{CR^2}{1 + \sqrt{1 - C^2R^2}} \\ R^2 &= X^2 + Y^2 \\ C &\equiv \frac{1}{\text{radius of spherical surface}} \end{aligned} \tag{66}$$

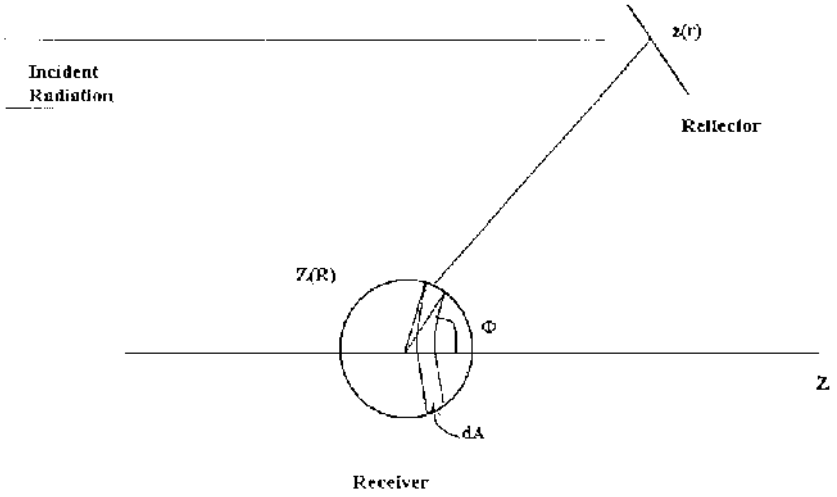


Figure 7 Rotationally symmetric beam shaping mirror and spherical detector.

The energy balance equation, Chapter 4, Eq. (10), is

$$ER\sqrt{1+\left(\frac{dZ}{dR}\right)^2}\,dR = \sigma\, r\, dr \tag{67}$$

where E gives the irradiance on the receiver surface. When the receiver surface equation is specified, then both sides of Eq. (67) can be integrated. E may be an arbitrary function of position on S, but it must have an adjustable parameter so that conservation of energy is satisfied between the input beam and receiver surface. For a flat receiver, $Z = Z_0$, conservation of energy between the input beam and receiver surface is given by

$$\int_{R_0}^{R} E\, R\, dR = \int_{r_0}^{r} \sigma\, r\, dr. \tag{68}$$

For uniform irradiance over a flat receiver with $R_0 = 0$, r_0, r_f, $R_f \neq 0$, integrating Eq. (68) gives

$$E_0 = \sigma \frac{\left(r_f^2 - r_0^2\right)}{R_f^2}. \tag{69}$$

This value of E_0 is substituted into the integrated form of Eq. (68) to yield a connection between r and R

$$R = \sqrt{\frac{\sigma\left(r^2 - r_0^2\right)}{E_0}}. \tag{70}$$

Equation (70) represents conservation of energy between the input beam and receiver surface and can now be used with the ray trace equations to write down a first-order differential equation of the reflector surface required to provide this illumination over the receiver.

A unit vector along the reflected light for this geometry is given by Eq. (45). Then the ray trace equation between reflector and receiver in the r–z plane is given by

$$\frac{(R - r)}{(Z_0 - z)} = \frac{2z'}{(z'^2 - 1)}. \tag{71}$$

Equation (71) can be solved as a quadratic equation for z' to give

$$z' = \left(\frac{z - Z_0}{r - R}\right) \pm \sqrt{1 + \left(\frac{z - Z_0}{r - R}\right)^2} = \left(\frac{z - Z_0}{r - \sqrt{\frac{\sigma\left(r^2 - r_0^2\right)}{E_0}}}\right) \pm \sqrt{1 + \left(\frac{z - Z_0}{r - \sqrt{\frac{\sigma\left(r^2 - r_0^2\right)}{E_0}}}\right)^2} \tag{72}$$

where the $+(-)$ sign is used when the concave side of mirror is oriented towards the positive (negative) z direction, and R as a function of r is given by Eq. (70). Equation (72) can be solved numerically to determine the shape of the reflector. The initial conditions are $z = z_0$ when $r = r_0$. Reference 13 contains several solar collectors designed by solution of Eq. (72) for different initial conditions.

In this section, two methods of designing non-laser beam profile shaping optical systems have been discussed. These methods are equivalent and are based on the application of conservation of energy between the input beam and receiver surface and on the ray trace equations between reflector and receiver. These methods for designing non-laser beam shaping systems are generally applicable to all forms of incident radiation. In the next section, the optical design method using the conservation of energy condition will be extended to include a point and line source of radiation for heating and illumination applications.

III. APPLICATION TO POINT AND LAMBERTIAN SOURCE

In the previous section, the principle of conservation of energy and the ray trace equations were used to obtain a first- and a second-order differential equation used in optical design of beam shaping systems to illuminate a receiver with a prescribed irradiance when the input beam was collimated. However, for many non-laser beam shaping applications associated with heating or illumination, it is important to consider the finite size and location of the source of radiation. Schruben (21) reported a differential-equation-based design of an illumination system for a reflector and small Lambertian source. Burkhard and Shealy (22) reported an optical design method for shaping a mirror to transform a point or line source of light into a prescribed irradiance over a receiver surface. In this section, the first-order differential equation* discussed in Sec. II.B will be revised to use with point or small line (Lambertian) source of light.

Consider the geometrical configuration of point source, mirror and receiver surface shown in Fig. 8, where the mirror and receiver surfaces have rotational symmetry around the z-axis.

The current analysis will take into account the fact that the incident radiation is not collimated, or its wavefront is not planar. This means that the direction of the incident light from the source will vary over the surface of the optics, which differs from the beam shaping applications addressed in Sec. II.B.2 and shown in Figs. 5 and 6 where the direction of incident light over optics was constant. A unit vector along an incident ray upon the mirror is given by

$$\mathbf{a} = \hat{\mathbf{r}} \sin\theta + \hat{\mathbf{k}} \cos\theta \tag{73}$$

where $\hat{\mathbf{r}} (= \hat{\mathbf{i}} \cos\phi + \hat{\mathbf{j}} \sin\phi)$ is the radial polar unit vector in the x–y plane. It will be helpful to understand clearly all the variables used in the present analysis. (ϕ, θ, ρ) are the conventional spherical coordinates, where the z-axis is also the optical (symmetry) axis. Since the beam shaping optics has rotational symmetry about the optical (z-) axis, it is convenient to use the r–z plane polar coordinates (ρ, θ) to solve for the shape of the mirror surface.

* Reference 19 contains a formula for the flux flow equation applied to reflection (or refraction) of point source light to illuminate a receiver surface. This formula for the flux flow equation could be used to derive a second-order differential equation for design of a mirror for beam shaping of point source light. However, it is more straightforward to use the two first-order differential equations resulting from application of conservation of energy and the ray trace equations than the second-order differential equation obtained from the flux flow equation as part of the optical design of a mirror used with point or extended source of light.

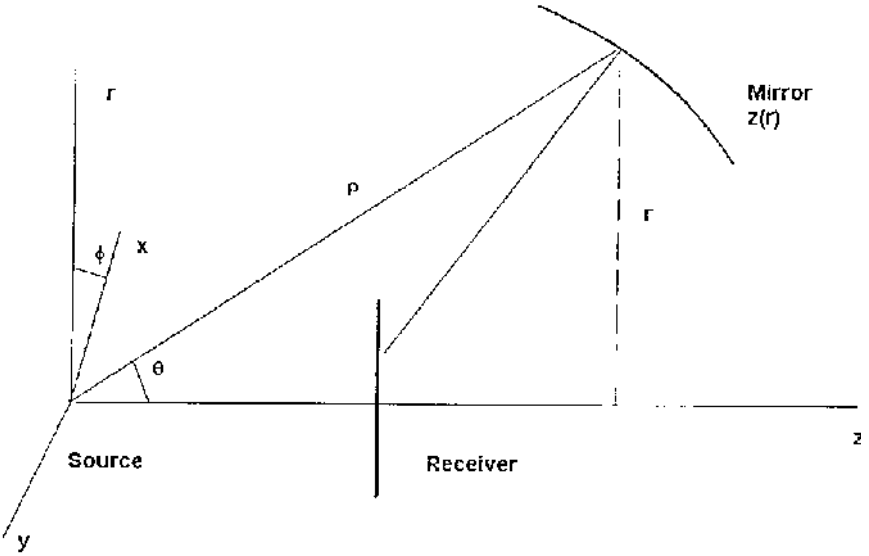


Figure 8 Geometrical configuration of point source, mirror and receiver surface.

The slope of the mirror in the r–z plane is given by

$$z' \equiv \frac{dz}{dr} = \frac{(\rho' \cos\theta - \rho \sin\theta)}{(\rho' \sin\theta + \rho \cos\theta)} \tag{74}$$

where $\rho' \equiv d\rho/d\theta$, $z = \rho\cos\theta, r = \rho\sin\theta$ as shown in Fig. 8. From Eq. (43), a unit normal vector of the mirror can be written in terms of the coordinates (ρ, θ) as

$$\mathbf{n} = \frac{-z'\hat{\mathbf{r}} + \hat{\mathbf{k}}}{\sqrt{1 + z'^2}} = \frac{-\hat{\mathbf{r}}(\rho' \cos\theta - \rho \sin\theta) + \hat{\mathbf{k}}(\rho' \sin\theta + \rho \cos\theta)}{\sqrt{\rho'^2 + \rho^2}}. \tag{75}$$

The direction of the reflected ray $\mathbf{A}$ can be evaluated from Eq. (5) to give

$$\mathbf{A} = \frac{\hat{\mathbf{r}}(\rho'^2 \sin\theta + 2\rho\rho' \cos\theta - \rho^2 \sin\theta)}{(\rho'^2 + \rho^2)} + \frac{\hat{\mathbf{k}}(\rho'^2 \cos\theta + 2\rho\rho' \sin\theta - \rho^2 \cos\theta)}{(\rho'^2 + \rho^2)}. \tag{76}$$

Then, the ray trace equation in the r–z plane, connecting the point (r, z) on reflector to the point (R, Z) on the receiver, is given by

$$\frac{(R-r)}{Z(R)-z(r)} = \frac{A_r}{A_z} = \frac{(\rho'^2 \sin\theta + 2\rho\rho' \cos\theta - \rho^2 \sin\theta)}{(\rho'^2 \cos\theta + 2\rho\rho' \sin\theta - \rho^2 \cos\theta)}. \tag{77}$$

For a planar receiver surface, $Z = Z_0$, a constant, Eq. (77) can be written as a first-order differential equation for the reflector surface

$$\rho' = \frac{\{\rho(R\sin\theta + Z_0\cos\theta - \rho) \pm \rho\sqrt{(-R\sin\theta - Z_0\cos\theta + \rho)^2 + (R\cos\theta - Z_0\sin\theta)^2}\}}{(\rho\cos\theta - Z_0\sin\theta)} \tag{78}$$

where the $+(-)$ sign in Eq. (78) is chosen to insure that ρ' is positive (negative) as required by the geometrical configuration of source, reflector, and receiver shown in Fig. 6. In subsequent calculations in this chapter, the positive root of Eq. (78) will be used.

In order to solve Eq. (78), the energy balance equation must used to obtain an expression for $R(\rho, \theta)$. For a flat receiver, conservation of energy condition [Chapter 4, Eq. (10)] becomes

$$\int_0^R 2\pi E(R) R\, dR = \int_{\theta_0}^{\theta} I(\theta) 2\pi r^2 \sin\theta\, d\theta \tag{79}$$

where $I(\theta)$ is the intensity of the source, and the reflectance of mirror has been assumed to be equal to unity. If the source is a small Lambertian source (41) along the z-axis, then $I(r, \theta) = I_0 \sin\theta / r^2$, where I_0 is a constant. If the optics is such that direct illumination from source to receiver can be ignored, then Eq. (79) can be integrated to obtain the constant E_0which insures conservation of energy for this system. For uniform irradiance over the receiver, $E = E_0$, integrating Eq. (79) over the full beam $[\theta \in (\theta_0, \theta_m)$ and $R \in (0, R_{\mathrm{m}})]$ leads to the following result:

$$E_0 = \left(\frac{I_0}{R_m^2}\right)\left[\theta_m - \theta_0 + \left(\frac{\sin 2\theta_0 - \sin 2\theta_m}{2}\right)\right] \tag{80}$$

where R_m and θ_m are the maximum values of R and θ. Similarly, for an isotropic point source, uniform irradiance of the receiver leads to the following expression for E:

$$E_0 = 2I_0 \frac{(\cos\theta_0 - \cos\theta_m)}{R_m^2}. \tag{81}$$

The constants I_0, θ_0, θ_m can be chosen to give the desired value of E_0.

For back-lighting configuration (Fig. 6a), integrating Eq. (79) leads to an expression for $R(\theta)$ which can be used to integrate Eq. (78) to determine

the shape of mirror required for a specific beam-shaping application. For a Lambertian source,

$$R = \sqrt{\frac{I_0}{2E_0}[2(\theta - \theta_0) - (\sin 2\theta - \sin 2\theta_0)]}, \tag{82}$$

and for an isotropic point source,

$$R = \sqrt{\frac{2I_0}{E_0}(\cos\theta_0 - \cos\theta)}. \tag{83}$$

Equation (78) can now be integrated to obtain the shape of reflector that will transform point source light into uniform irradiance on the back of a detector. Figure 9a is a scaled drawing of a point source, mirror, and back-lighted detector of an example solution of this differential equation optical design of non-laser beam shaping systems. Similar calculations could also be done for a Lambertian source.

When the optical configuration (Fig. 6b) allows direct illumination of the receiver from the source as well as from the reflector, the analysis leading to Eqs. (80)–(83) needs to take into account the direct illumination. The irradiance, E_{direct}, directly incident upon receiver from source, as shown in Fig. 6b, is given by

$$E_{\text{direct}} = \sigma \cos\chi \tag{84}$$

where σ is the flux density from source evaluated at the receiver and χ is the angle between ray and normal to receiver. For a Lambertian source (41)

$$\sigma = \frac{I_0 \sin\chi}{r_0^2} \tag{85}$$

where r_0 is the distance from the source to a point on the receiver. From the geometry of a direct illumination system, the following relations hold:

$$\begin{aligned} r_0 &= \sqrt{R^2 + Z_0^2} \\ \sin\chi &= \frac{R}{r_0} \\ \cos\chi &= \frac{Z_0}{r_0} \end{aligned} \tag{86}$$

Then, the direct irradiance on receiver is given by

$$E_{\text{direct}} = \frac{RZ_0 I_0}{[R^2 + Z_0^2]^2}. \tag{87}$$

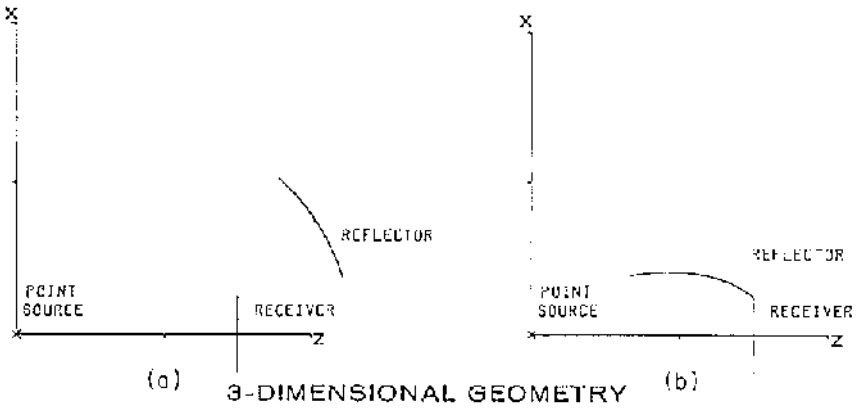


Figure 9 Uniform illumination of a flat disk by a rotationally symmetric reflector for point source light incident upon (a) back lighted and (b) direct lighted configurations. (From Ref. 13.)

To obtain a uniform irradiance (E_0) over receiver surface when both direct and reflected light are considered, note that the total irradiance $E_{\text{total}}(R)$ at the receiver is the sum of the reflected light $E_{\text{reflected}}(R)$ plus the direct light $E_{\text{direct}}(R)$ from the source:

$$E_{\text{total}}(R) = E_{\text{direct}}(R) + E_{\text{reflected}}(R). \tag{88}$$

If $E_{\text{total}}(R)$ is to be constant, then $E_{\text{reflected}}(R)$, which appears on the left-hand side of Eq. (79), will be given by

$$E_{\text{reflected}} = E_{\text{total}} - E_{\text{direct}}, \tag{89}$$

and the constant E_{total} needs to satisfy the following integral equation

$$\int_0^{R_m} 2\pi \left[E_{\text{total}} - \frac{RZ_0 I_0}{[R^2 + Z_0^2]^2} \right] R\, dR = \int_{\theta_0}^{\theta_m} I_0 2\pi \sin^2 \theta\, d\theta. \tag{90}$$

After carrying out integrals and solving for E_{total} in Eq. (90), one has

$$E_{\text{total}} = \frac{I_0}{R_m^2}\left\{(\theta_m - \theta_0) - \frac{1}{2}(\sin 2\theta_m - \sin 2\theta_0) + \tan^{-1}\left(\frac{R_m}{Z_0}\right) - \frac{\left(\frac{R_m}{Z_0}\right)}{1 + \left(\frac{R_m}{Z_0}\right)^2}\right\}. \tag{91}$$

Equation (91) expresses the total energy leaving the sources that is intercepted by the mirror and receiver. The value of E_{total} from Eq. (91) is put into Eq. (90) with upper limits of R and θ to solve for $R(\rho, \theta)$ and subsequently solve numerically the differential equation (78) for $\rho(\theta)$ the mirror surface shape (13,21). Figure 9b is a scaled drawing of point source and reflector which will uniformly illuminate receiver plane when taking both direct and reflected light into account. This optical design method has been extended in Ref. 22 to include multiple point sources and continuous line sources along the symmetry axis when solving for the shape of a mirror which will uniformly illuminate a detector.

IV. CONCLUSION

The design and analysis of non-laser beam shaping systems has been discussed in this chapter. The ray trace equations and the principle of conservation of energy within a bundle of rays have been used to derive several alternate forms of the flux flow equation, Eqs. (22) and (40). Equation (22) is useful when computing the irradiance distribution for a collimated incident beam. Equation (40) is useful when computing the irradiance distribution for cases when the incident beam wavefront is not planar and, subsequently, when the incident beam wavefront curvatures vary over the beam shaping optics, which happens for point or extended sources near the optics.

The flux flow equation can be used to monitor the irradiance along a ray path as it propagates through the optical system. When the input and output beam profiles are known, the shape of a single surface is determined by a differential energy balance equation. Specific examples of using this optical design method are presented for collimated, point and Lambertian sources of radiation.

REFERENCES

1. AC Claus. On Archimedes' burning glass. Appl Opt 12.10: A14, 1973.
2. ON Stavroudis. Comments on: on Archimedes' burning glass. Appl Opt 12.10: A16, 1973.
3. MH Cobble. Theoretical concentration for solar furnaces. Solar Energy 5.2: 61, 1961.
4. O Kamada. Theoretical concentration and attainable temperature in solar furnaces. Solar Energy 9.1: 39–47, 1965.
5. K-E Hassan, MF El-Refaie. Theoretical performance of cylindrical parabolic solar concentrators. Solar Energy 15: 219–244, 1973.
6. RA Zakhidov, AA Vainer. Distribution of radiation produced by a paraboloidal concentrator. Geliotekhnika (Applied Solar Energy) 10.3: 34–40, 1974. The Russian to English translation of this article was part of the Allerton Press journal program.
7. WS Duff, GF Lameiro, GOG Löf. Parametric performance and cost models for solar concentrators. Solar Energy 17: 47–58, 1975.
8. GL Strobel, DG Burkhard. Irradiance for skew rays incident upon trough-like solar collector of arbitrary shape. Solar Energy 20: 25–27, 1978.
9. RF Jones Jr. Collection properties of generalized light concentrators. J Opt Soc Am 67.11 1594–1598, 1977.
10. WT Welford, R Winston. High Collection Nonimaging Optics. New York: Academic Press, 1989.
11. R Winston. Principles of solar concentrators of a novel design. Solar Energy 16: 89–95, 1974.
12. R Winston, H Hinterberger. Principles of cylindrical concentrators for solar energy. Solar Energy 17: 255–258, 1975.
13. DG Burkhard, DL Shealy. Design of reflectors which will distribute sunlight in a specified manner. Solar Energy 17.1: 221–227, 1975.
14. JH McDermit. Curved reflective surfaces for obtaining prescribed irradiation distributions. PhD dissertation, University of Mississippi, Oxford, MS, 1972.
15. TE Horton, JH McDermit. Optical design of solar concentrators. J Energy 4.1: 4–9, 1980.
16. DL Shealy, HM Berg. Simulation of optical coupling from surface emitting LEDs. Appl Opt 22.11: 1722–1730, 1983.
17. DR Gabardi, DL Shealy. Coupling of domed light-emitting diodes with a multimode step-index optical fiber. Appl Opt 25.19: 3435–3442, 1986.
18. JH McDermit, TE Horton. Reflective optics for obtaining prescribed irradiative distributions from collimated sources. Appl Opt 13.6: 1444–1450, 1974.
19. DG Burkhard, DL Shealy, RU Sexl. Specular reflection of heat radiation from an arbitrary reflector surface to an arbitrary receiver surface. Int J Heat Mass Transfer 16: 271–280, 1973.
20. DG Burkhard, DL Shealy. View function in generalized curvilinear coordinates for specular reflection of radiation from a curved surface. Int J Heat Mass Transfer 16: 1492–1496, 1973.

21. JS Schruben. Analysis of rotationally symmetric reflectors for illuminating systems. J Opt Soc Am 64.1: 55–58, 1974.
22. DG Burkhard, DL Shealy. Specular aspheric surface to obtain a specified irradiance from discrete or continuour line source radiation: design. Appl Opt 14.6: 1279–1284, 1975.
23. DG Burkhard, DL Shealy. A different approach to lighting and imaging: formulas for flux density, exact lens and mirror equations, and caustic surfaces in terms of the differential geometry of surfaces. Proc SPIE 692: 248–272, 1986.
24. JM Gordon, A Rabl. Reflectors for uniform far-field irradiance: fundamental limits and example of an axisymmetric solution. Appl Opt 37.1: 44–47, 1998.
25. BS Westcott, AP Norris. Reflector synthesis for generalized far-fields. J Phys A: Math Gen 8.4: 521–532, 1975.
26. F Brickell, BS Westcott. Reflector design for two-variable beam shaping in the hyperbolic case. J Phys A: Math Gen 9.1: 113–128, 1976.
27. P Mouroulis, J Macdonald. Geometrical Optics and Optical Design. New York: Oxford University Press, 1997.
28. DG Kock. Simplified irradiance/illuminance calculations in optical systems. Proc SPIE 1780:226–240, 1993. Presented at the International Symposium on Optical Design, Berlin, September 14, 1992.
29. DL Shealy, DG Burkhard. Analytical illuminance calculation in a multi-interface optical system. Optica Acta 22.6: 485–501, 1975.
30. DL Shealy, DG Burkhard. Flux density for ray propagation in discrete index media expressed in terms of the intrinsic geometry of the deflecting surface. Optica Acta 20.4: 287–301, 1973.
31. M Born, E Wolf. Principles of Optics. 5th ed. New York: Pergamon Press, 1975, p. 115.
32. DG Burkhard, DL Shealy. Simplified formula for the illuminance in an optical system. Appl Opt 20.5: 897–909, 1981.
33. ON Stavroudis. The Optics of Rays, Wavefronts, and Caustics. New York: Academic Press, 1972, p. 157.
34. DL Shealy, DG Burkhard. Heat flux contours on a plane for parallel radiation specularly reflected from a cone, a hemisphere and a paraboloid. Int J Heat Mass Transfer 16: 281–291, 1973.
35. DG Burkhard, DL Shealy. Flux density for ray propagation in geometrical optics. J Opt Soc Am 63.3: 299–304, 1973.
36. E Kreyszig. Introduction to Differential Geometry. Math Exp 16. Ontario: University of Toronto Press, 1968.
37. VA Fock. Electromagnetic Diffraction and Propagation Problems. New York: Pergamon Press, 1965, p. 164.
38. M Born, E Wolf. Principles of Optics. 5th ed. New York: Pergamon Press, 1975, pp. 126–127.
39. ON Stavroudis, RC Fronczek, J Opt Soc Am 66: 795, 1976.
40. DG Burkhard, DL Shealy. A different approach to lighting and imaging: formulas for flux density, exact lens and mirror equations, and caustic surfaces in terms of the differential geometry of surfaces. Proc SPIE 692: 48–72, 1986.

41. MV Klein. Optics. New York: John Wiley & Sons, Inc., 1970, p. 124.
42. JH McDermit, TE Horton. Reflective optics for obtaining prescribed irradiance distribution from collimated sources. Appl Opt 13.6: 1444–1450, 1974.

APPENDIX A

Summary of Some Concepts and Results from Differential Geometry*

In general the equation of a surface is a constraint equation between three coordinates such that when two coordinates are given the third coordinate will indeed lie on the surface. The equation of the surface can be represented by the following symbolic equation

$$w = w(u, v) \tag{A-1}$$

where $[u, v, w]$ are curvilinear coordinates which themselves are defined in terms of the Cartesian coordinates $[x, y, z]$

$$\begin{aligned} x &= x(u, v, w) \\ y &= y(u, v, w). \\ z &= z(u, v, w) \end{aligned} \tag{A-2}$$

Combining the equation of the surface (A-1) and the relationship between Cartesian coordinates $[x, y, z]$ of a point and the curvilinear coordinates $[u, v, w]$ of the same point, Eq. (A-2), one obtains a parametric representation of the surface:

$$\begin{aligned} x &= x(u, v) \\ y &= y(u, v), \\ z &= z(u, v) \end{aligned} \tag{A-3}$$

where $[u, v]$ are the curvilinear coordinates of the surface which can also be considered as surface parameters. It will be convenient to write Eq. (A-3) as the vector equation

$$\mathbf{r} = \mathbf{r}(u, v) \tag{A-4}$$

where the Cartesian components of the vector $\mathbf{r}(\equiv \hat{\mathbf{i}}x + \hat{\mathbf{j}}y + \hat{\mathbf{k}}z)$ are given by Eq. (A-3).

* For a more complete discussion of differential geometry, the reader is referred to one or more of the comprehensive books in the literature on this topic, such as Ref. 36.

The two vectors $\mathbf{r}_u \equiv \partial\mathbf{r}/\partial u$ and $\mathbf{r}_v \equiv \partial\mathbf{r}/\partial v$, which are tangent to the u-parameter curve and the v-parameter curve on the surface, are two linearly independent vectors in the tangent plane of the surface at a specified point on the surface and thus, any vector in the tangent plane can be written as a linear combination of the vectors $[\mathbf{r}_u, \mathbf{r}_v]$. The unit normal vector to the surface can then be written as the vector cross product of $[\mathbf{r}_u, \mathbf{r}_v]$:

$$\mathbf{n} = \frac{\mathbf{r}_u \times \mathbf{r}_v}{|\mathbf{r}_u \times \mathbf{r}_v|}. \tag{A-5}$$

The measurement of lengths on the surface is conveniently expressed in terms of the coefficients:

$$\begin{aligned}
g_{uu} &\equiv \mathbf{r}_u \cdot \mathbf{r}_u = \left(\frac{\partial x}{\partial u}\right)^2 + \left(\frac{\partial y}{\partial u}\right)^2 + \left(\frac{\partial z}{\partial u}\right)^2 \\
g_{vv} &\equiv \mathbf{r}_v \cdot \mathbf{r}_v = \left(\frac{\partial x}{\partial v}\right)^2 + \left(\frac{\partial y}{\partial v}\right)^2 + \left(\frac{\partial z}{\partial v}\right)^2 \\
g_{uv} &\equiv \mathbf{r}_u \cdot \mathbf{r}_v = \left(\frac{\partial x}{\partial u}\right)\left(\frac{\partial x}{\partial v}\right) + \left(\frac{\partial y}{\partial u}\right)\left(\frac{\partial y}{\partial v}\right) + \left(\frac{\partial z}{\partial u}\right)\left(\frac{\partial z}{\partial v}\right) \\
g_{vu} &\equiv \mathbf{r}_v \cdot \mathbf{r}_u = g_{uv}
\end{aligned} \tag{A-6}$$

or by

$$g_{jk} = \mathbf{r}_j \cdot \mathbf{r}_k \tag{A-7}$$

where it is understood in Eq. (A-7) that (j, k) may each take on the value of u or v.

The coefficients g_{uv} transform like a second rank symmetric tensor and are called the metric coefficients of the surface or the coefficients of the first fundamental form of the surface. The first fundamental form of the surface is a quadratic expression for the differential arc-length of a curve on the surface and is given by

$$\begin{aligned}
ds^2 \equiv d\mathbf{r} \cdot d\mathbf{r} &= \left(\frac{\partial \mathbf{r}}{\partial u}\, du + \frac{\partial \mathbf{r}}{\partial v}\, dv\right) \cdot \left(\frac{\partial \mathbf{r}}{\partial u}\, du + \frac{\partial \mathbf{r}}{\partial v}\, dv\right) \\
&= g_{uu}(du)^2 + 2g_{uv}\, du\, dv + g_{vv}(dv)^2 \\
&= \sum_{j,k=u}^{v} g_{jk}\, d(j)\, d(k) = g_{jk}\, d(j)\, d(k).
\end{aligned} \tag{A-8}$$

where the simplified (last) expression uses the summation convention which means that if a given letter appears twice on the same side of an equation, then the summation of that letter must be carried out. For surfaces, the summation will be over the two curvilinear coordinates u and v.

The determinant of the metric coefficients is given by

$$g \equiv \begin{vmatrix} g_{uu} & g_{uv} \\ g_{uv} & g_{vv} \end{vmatrix} = g_{uu}g_{vv} - (g_{uv})^2. \tag{A-9}$$

By direct computation one sees that the determinant g is identically equal to the magnitude square of the normal vector $\mathbf{r}_u \times \mathbf{r}_v$:

$$g = |\mathbf{r}_u \times \mathbf{r}_v|^2. \tag{A-10}$$

In terms of g one can write the Cartesian components of $\mathbf{n}$:

$$\begin{aligned} n_x &= g^{-1/2}\left(\frac{\partial y}{\partial u}\frac{\partial z}{\partial v} - \frac{\partial y}{\partial v}\frac{\partial z}{\partial u}\right) \\ n_y &= g^{-1/2}\left(\frac{\partial z}{\partial u}\frac{\partial x}{\partial v} - \frac{\partial z}{\partial v}\frac{\partial x}{\partial u}\right) \\ n_z &= g^{-1/2}\left(\frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}\right) \end{aligned} \tag{A-11}$$

An expression for the element of area, da, on the surface in terms of the differentials du, dv is found by separately varying the position vector $\mathbf{r}(u, v)$ of a point on the surface by an amount du and dv, and then take the cross product of the resulting differential vectors:

$$da = \left|\frac{d\mathbf{r}}{du}\,du \times \frac{d\mathbf{r}}{dv}\,dv\right| = |\mathbf{r}_u \times \mathbf{r}_v|\,du\,dv = g^{1/2}\,du\,dv. \tag{A-12}$$

We have seen that a knowledge of the metric coefficients g_{jk} is sufficient for calculating lengths and areas on a surface; however, they do not uniquely determine a surface. In order to fully characterize a surface in terms of radii of curvature, for example, it is necessary to introduce another quadratic form in the coordinate differentials du, dv which is usually referred to as the second fundamental form of a surface and is given by

$$\begin{aligned} &b_{uu}(du)^2 + 2b_{uv}\,du\,dv + b_{vv}(dv)^2 \\ &\text{or simply by} \\ &b_{jk}d(j)d(k) \end{aligned} \tag{A-13}$$

where the summation convention is implied. The second fundamental form coefficients, b_{jk}, will transform like a second rank symmetric tensor and are

given by

$$
\begin{aligned}
b_{uu} &= \mathbf{r}_{uu}\cdot\mathbf{n} = n_x\,\frac{\partial^2 x}{\partial u^2} + n_y\,\frac{\partial^2 y}{\partial u^2} + n_z\,\frac{\partial^2 z}{\partial u^2};\\
b_{uv} &= \mathbf{r}_{uv}\cdot\mathbf{n} = n_x\,\frac{\partial^2 x}{\partial u\partial v} + n_y\,\frac{\partial^2 y}{\partial u\partial v} + n_z\,\frac{\partial^2 z}{\partial u\partial v};\\
b_{vu} &= \mathbf{r}_{vu}\cdot\mathbf{n} = b_{uv};\\
b_{vv} &= \mathbf{r}_{vv}\cdot\mathbf{n} = n_x\,\frac{\partial^2 x}{\partial v^2} + n_y\,\frac{\partial^2 y}{\partial v^2} + n_z\,\frac{\partial^2 z}{\partial v^2};
\end{aligned}
\tag{A-14}
$$

or when $\mathbf{r}_{jk} \equiv \partial^2 r/\partial(j)\,\partial(k)$

$$
b_{jk} = \mathbf{r}_{jk}\cdot\mathbf{n}; \quad (j,k = u,v). \tag{A-15}
$$

The determinant of the coefficients b_{jk} is given by

$$
b = b_{uu}b_{vv} - (b_{uv})^2. \tag{A-16}
$$

Since the unit normal $\mathbf{n}$ to the surface is given by

$$
\mathbf{n} = \frac{\mathbf{r}_u \times \mathbf{r}_v}{\sqrt{g}}, \tag{A-17}
$$

the second fundamental form coefficients, b_{jk}, can be written explicitly as

$$
\begin{aligned}
b_{uu} &= \frac{\mathbf{r}_u \times \mathbf{r}_v\cdot\mathbf{r}_{uu}}{\sqrt{g}};\\
b_{uv} &= \frac{\mathbf{r}_u \times \mathbf{r}_v\cdot\mathbf{r}_{uv}}{\sqrt{g}};\\
b_{vv} &= \frac{\mathbf{r}_u \times \mathbf{r}_v\cdot\mathbf{r}_{vv}}{\sqrt{g}};
\end{aligned}
\tag{A-18}
$$

or by

$$
b_{jk} = \frac{\mathbf{r}_u \times \mathbf{r}_v\cdot\mathbf{r}_{jk}}{\sqrt{g}}. \tag{A-19}
$$

Since $\mathbf{r}_j$ is a vector in the tangent plane,

$$
\mathbf{r}_j\cdot\mathbf{n} = 0. \tag{A-20}
$$

And therefore, one has after taking the partial derivative of Eq. (A-20) with respect to the variable k ($= u$ or v)

$$
\mathbf{r}_{jk}\cdot\mathbf{n} + \mathbf{r}_j\cdot\mathbf{n}_k = 0. \tag{A-21}
$$

Hence, the second fundamental form coefficients b_{jk} given by Eq. (A-15) can also be written as

$$b_{jk} = -\mathbf{r}_j \cdot \mathbf{n}_k. \tag{A-22}$$

One can now write the second fundamental form as

$$b_{jk}\, d(j)\, d(k) = -d\mathbf{r} \cdot d\mathbf{n}. \tag{A-23}$$

One can associate the second fundamental form with the distance from the tangent plane of a point on the surface to an adjacent point on the surface. This can be verified if one draws a tangent plane at a point P on the surface whose position vector is $\mathbf{r}(u, v)$. Then, the distance δ to an adjacent point P′ whose position vector is $r(u + \Delta u, v + \Delta v)$ will be given by

$$\delta = \Delta\mathbf{r} \cdot \mathbf{n} \tag{A-24}$$

where $\mathbf{n}$ is the unit normal to the surface at the point P and $\Delta\mathbf{r}$ is given by

$$\Delta\mathbf{r} = \mathbf{r}(u + \Delta u, v + \Delta v) - \mathbf{r}(u, v)$$

or, expanding in a Taylor series,

$$\Delta\mathbf{r} = (\mathbf{r}_u \Delta u + \mathbf{r}_v \Delta v) + \tfrac{1}{2}[\mathbf{r}_{uu}(\Delta u)^2 + 2\mathbf{r}_{uv}\Delta u \Delta v + \mathbf{r}_{vv}(\Delta v)^2 + \cdots]. \tag{A-25}$$

Hence, one has to second order in displacement,

$$\begin{aligned} \delta &= \Delta\mathbf{r} \cdot \mathbf{n} \\ &= \tfrac{1}{2}[\mathbf{r}_{uu} \cdot \mathbf{n}(\Delta u)^2 + 2\mathbf{r}_{uv} \cdot \mathbf{n}\Delta u \Delta v + \mathbf{r}_{vv} \cdot \mathbf{n}(\Delta v)^2] \end{aligned}$$

or in the limit of small displacement

$$\begin{aligned} \delta &= \tfrac{1}{2}[b_{uu}(du)^2 + 2b_{uv}\, du\, dv + b_{vv}(dv)^2] \\ &= \tfrac{1}{2} b_{jk} d(j) d(k), \quad (j, k \text{ summed over } u, v) \end{aligned} \tag{A-26}$$

It is interesting to note that the coefficients b_{jk} are strictly functionally dependent upon the properties of the surface at the point P, whereas the location of the adjacent point P′ is uniquely characterized by the displacement du, dv. Thus, the interpretation of the second fundamental form as being proportional to the distance between P′ and the tangent plane to the surface at P seems reasonable.

We shall now be interested in deriving expressions for suitable measures of the curvature of a surface. The concept of curvature of a surface is given meaning in terms of the curvature of an arbitrary curve C on the surface S which is represented by

$$u = u(s), \quad v = v(s) \tag{A-27}$$

where s is the arc-length of C. For such a curve C on S, one could not expect the unit principal normal $\mathbf{p}$ of the curve C, which is given by

$$\mathbf{p} = \frac{1}{K}\frac{d^2\mathbf{r}(u(s), v(s))}{ds^2} \tag{A-28}$$

where K is the curvature of C, to lie along the unit normal to the surface. On the contrary, $\mathbf{p}$ and $\mathbf{n}$ will make a non-zero angle η which will be a function of both the curve C and the surface S. The cosine of η is given by

$$\cos\eta = \mathbf{p}\cdot\mathbf{n}. \tag{A-29}$$

The differentiation of $\mathbf{r}(u(s), v(s))$ with respect to s appearing in Eq. (A-28) is given by

$$\begin{aligned}
\frac{d\mathbf{r}}{ds} &= \frac{\partial\mathbf{r}}{\partial u}\frac{du}{ds} + \frac{\partial\mathbf{r}}{\partial v}\frac{dv}{ds} = \mathbf{r}_j\frac{d(\,j)}{ds} \\
\frac{d^2\mathbf{r}}{ds^2} &= \frac{d}{ds}\left(\frac{\partial\mathbf{r}}{\partial u}\right)\frac{du}{ds} + \frac{\partial\mathbf{r}}{\partial u}\frac{d^2u}{ds^2} + \frac{d}{ds}\left(\frac{\partial\mathbf{r}}{\partial v}\right)\frac{dv}{ds} + \frac{\partial\mathbf{r}}{\partial v}\frac{d^2v}{ds^2} \\
&= \left[\frac{\partial^2\mathbf{r}}{\partial u^2}\frac{du}{ds} + \frac{\partial^2\mathbf{r}}{\partial u\partial v}\frac{dv}{ds}\right]\frac{du}{ds} + \frac{\partial\mathbf{r}}{\partial u}\frac{d^2u}{ds^2} \\
&\quad + \left[\frac{\partial^2\mathbf{r}}{\partial u\partial v}\frac{du}{ds} + \frac{\partial^2\mathbf{r}}{\partial v^2}\frac{dv}{ds}\right]\frac{dv}{ds} + \frac{\partial\mathbf{r}}{\partial v}\frac{d^2v}{ds^2} \\
&= \mathbf{r}_{jk}\frac{d(\,j)}{ds}\frac{d(k)}{ds} + \mathbf{r}_j\frac{d^2(\,j)}{ds^2}
\end{aligned} \tag{A-30}$$

Combining Eqs. (A-28), (A-29), and (A-30) and making use of Eq. (A-30), one finds

$$K\cos\eta = \mathbf{r}_{jk}\cdot\mathbf{n}\frac{d(\,j)}{ds}\frac{d(k)}{ds}. \tag{A-31}$$

However, we have already defined $\mathbf{r}_{jk}\cdot\mathbf{n}$ as being the coefficients b_{jk} of the second fundamental form, and from Eq. (A-8) we have identified ds^2 with the first fundamental form $g_{lm}d(l)d(m)$. Therefore, Eq. (A-31) becomes

$$K\cos\eta = \frac{b_{jk}\,d(\,j)\,d(k)}{g_{lm}\,d(l)\,d(m)}. \tag{A-32}$$

It is interesting to note that the right-hand side of Eq. (A-32) is only a function of the point (u, v) on the surface and the direction (du/dv) of the curve passing through that point. Thus, at a given point P on the surface S, if we fix the tangent to the curve, then the right-hand side of Eq. (A-32) is a constant which we shall denote by K_n

$$K_n = \frac{b_{jk}\, d(\,j)\, d(k)}{g_{lm}\, d(l)\, d(m)}. \tag{A-33}$$

K_n is called the normal curvature of the surface S at the point P. From Eqs. (A-32) and (A-33)

$$K \cos \eta = K_n. \tag{A-34}$$

If $\eta = 0,\ K = K_n$ and if $\eta = \pi,\ K = -K_n$. Hence $|K_n|$ is the curvature of the intersection of the surface S and the plane which passes through both the tangent to the curve on the surface and the normal to the surface at P. Such a curve will be called a normal section of S. One may introduce the idea of the radii of curvature of a normal section by putting $R = 1/K_n$ in Eq. (A-33).

We are now interested in obtaining an expression for suitable measures of the curvature of a surface in terms of the coefficients g_{jk} and b_{jk} . We shall see that these measures can be expressed in terms of the two primary curvatures of a surface. Writing out Eq. (A-33) explicitly one has

$$K_n = \frac{b_{uu}(du)^2 + 2b_{uv}\, du\, dv + b_{vv}(dv)^2}{g_{uu}(du)^2 + 2g_{uv}\, du\, dv + g_{vv}(dv)^2}$$

$$\text{or, in terms of the direction } q \equiv du/dv, \tag{A-35}$$

$$K_n = \frac{b_{uu}q^2 + 2b_{uv}q + b_{vv}}{g_{uu}q^2 + 2g_{uv}q + g}$$

which can be written as

$$(b_{uu} - K_n g_{uu})q^2 + 2(b_{uv} - K_n g_{uv})q + (b_{vv} - K_n g_{vv}) = 0. \tag{A-36}$$

The curvature K_n in Eq. (A-36) is a function of q. Therefore, if one differentiates Eq. (A-36) with respect to q and makes use of the condition for an extremum value of K_n, namely, $dK_n/dq = 0$, one obtains

$$(b_{uu} - K_n g_{uu})q + (b_{uv} - K_n g_{uv}) = 0. \tag{A-37}$$

In order to solve for the explicit values of the extremum values of the normal curvature, one must eliminate q between Eqs. (A-36), (A-37) and solve for K_n. One obtains

$$(b_{vv} - K_n g_{vv})(b_{uu} - K_n g_{uu}) - (b_{uv} - K_n g_{uv})^2 = 0$$

or

$$K_n^2 + \frac{1}{g}(g_{uu}b_{vv} - 2g_{uv}b_{uv} + g_{vv}b_{uu})K_n + \frac{b}{g} = 0. \tag{A-38}$$

The principal curvatures K_1, K_2, which are the extremum values of K_n, will be solutions of Eq. (A-38). It follows then that

$$(K_n - K_1)(K_n - K_2) = K_n^2 - (K_1 + K_2)K_n + K_1K_2 = 0 \tag{A-39}$$

where the coefficient of K_n in Eq. (A-38) is put equal to $(K_1 + K_2)$. It is conventional to denote the product of K_1K_2 by K, and by comparing Eqs. (A-38) and (A-39), one can write

$$K \equiv K_1K_2 = \frac{b}{g}. \tag{A-40}$$

K is called the Gaussian curvature. In spite of the fact that the Gaussian curvature given by Eq. (A-40) appears to depend on both the first and second fundamental forms, it can be shown that K depends only on the first fundamental form coefficients and their first and second derivatives.

The arithmetic mean $(K_1 + K_2)/2$ of the principal curvatures is called the mean curvature of the surface and is denoted by H. From Eqs. (A-38) and (A-39) one sees

$$H = \tfrac{1}{2}(K_1 + K_2) = \frac{1}{2g}(g_{uu}b_{vv} - 2g_{uv}b_{uv} + g_{vv}b_{uu}). \tag{A-41}$$

The Gaussian curvature and mean curvature are useful expressions of the curvature of a surface in terms of the coefficients g_{jk}, b_{lm}. As seen from Eqs. (A-40) and (A-41), a knowledge of H and K determines the principal curvatures K_1, K_2 of a surface which themselves are extremum values of the normal curvature K_n. Furthermore, by the Euler theorem of differential geometry, one can express the curvature of a normal section in an arbitrary direction in terms of the two principal curvatures K_1, K_2 and the angle between the direction of the curve and the direction of the curve with the curvature K_1.

This brief discussion of the theory of surfaces is intended only to highlight some of the key ideas of differential geometry used in Chapter 8 as well as to present an accessible reference to derivations of some of the results that we have used in this chapter. For a more complete discussion of differential geometry with application to the theory of surfaces, see Ref. 36, for example.

9

Current Technology of Beam Profile Measurements

Carlos B. Roundy
Spiricon, Inc., Logan, Utah

I. INTRODUCTION

As explained in earlier chapters of this book, laser beam shaping is a process whereby the irradiance of the laser beam is changed along its cross section. In some cases the laser beam is shaped so that it is uniform or flat top. In other cases it is given a different shape such as a Gaussian or super Gaussian. In order for this laser beam shaping to be effective, it is necessary to be able to measure the degree to which the irradiance pattern or beam profile has been modified by the shaping medium. In some cases the beam shaping requires a specific input beam. For example, in many cases the input beam must be Gaussian in order for the shaping to create an undistorted flat top beam. In this case the beam profile of the input beam must be measured to assure that the proper input beam is being used. If the input beam does not have the proper profile, then measurements will tell the user that adjustments to the source beam must be made before attempting to perform the laser beam shaping. Therefore, laser beam profile analysis becomes an essential part of effective laser beam shaping.

This chapter describes the general state of the art of laser beam profile analysis (1–14). It introduces the general need for beam profile analysis, methods for measuring the laser beam profile, a description of instrumentation that is used in beam profile measurement, a discussion of the information that can be obtained simply by viewing the beam profile, and finally, how quantitative measurements are made on laser beam profiles, and the significance of those quantitative measurements.

II. LASER BEAM PROPERTIES

A. Unique Laser Beam Characteristics

Laser beams produce light with many characteristics that are unique to this type of light. Some of the things that make laser beams unique are listed in Table 1. For example, the monochromatic nature of a laser beam means that it is typically a single narrow wavelength with very little light at wavelengths other than the central peak. The temporal nature of a laser beam enables it to vary from a continuous wave to an extremely short pulse providing very high power densities. The coherence of a laser enables it to travel in a narrow beam with a small and well-defined divergence or spread. This allows a user to define exactly the area illuminated by the laser beam. Because of this coherence a laser beam can also be focused to a very small and intense spot in a highly concentrated area. This concentration makes the laser beam useful for many applications in physics, chemistry, the medical industry, and in industrial applications. Finally, a laser beam has a unique irradiance profile that gives it very significant characteristics. The beam profile is a unique pattern of irradiance distribution across the beam.

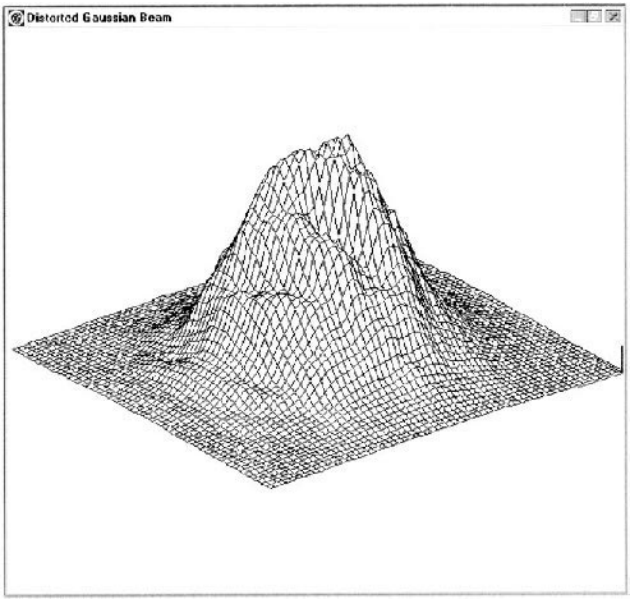
B. Significance of the Beam Profile

The significance of the beam profile is that the energy density, the concentration, and the collimation of the light are all affected by it. Also the propagation of the beam through space is significantly affected by the beam profile. Figure 1 shows a number of typical laser beam profiles illustrating the variety that can exist. Since such a variety exists in laser beam profiles, it is essential to measure the profile in any application if the energy distribution affects the performance of the laser or its intended purpose.

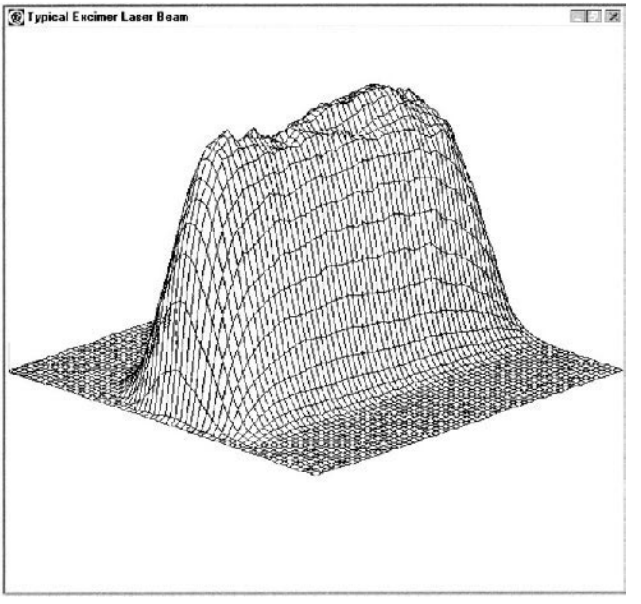
Examples of two different types of ideal laser beams for different purposes are a Gaussian and a flat top beam. A Gaussian beam allows the highest concentration of focused light, whereas a flat top beam allows

Table 1 Unique characteristics of a laser beam

Monochromatic (single wavelength)
Temporal (continuous wave to femtosecond pulses)
Coherence (consistent phase between all light elements)
Highly concentrated (focusable to extremely small spots)
Beam irradiance profile (unique spatial power or energy distribution)

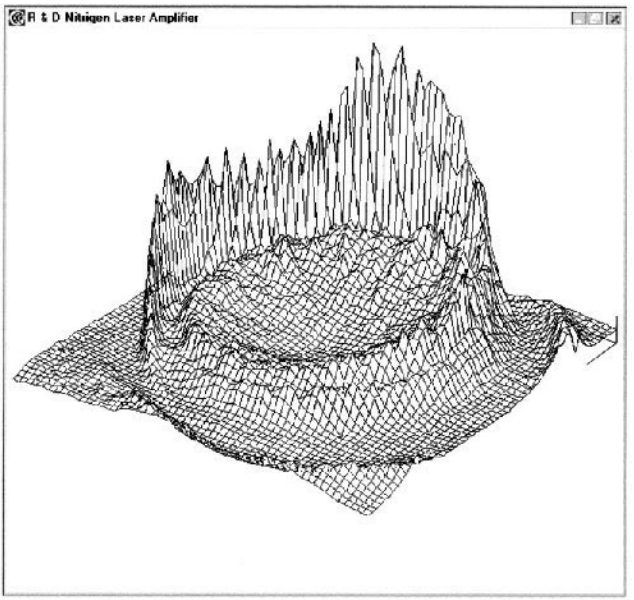


(a)



(b)

Figure 1 Various laser beam profiles: (a) HeNe, (b) Excimer, and (c) nitrogen ring laser.



(c)

Figure 1 (*Continued*)

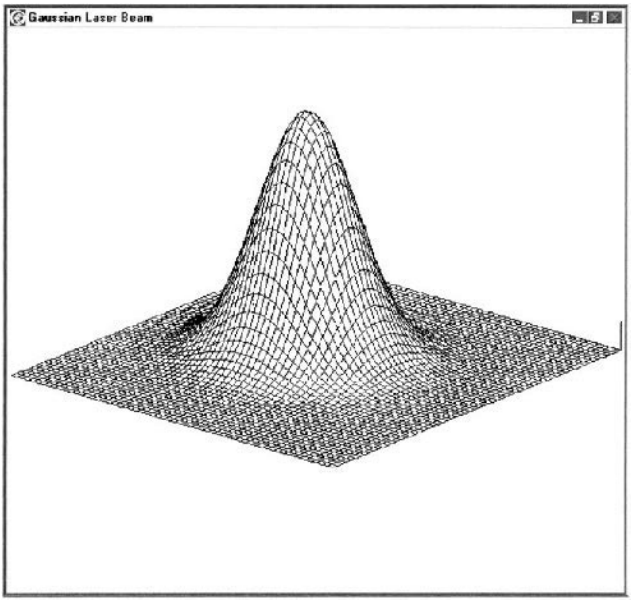
for very uniform distribution of the energy across a given area. These two idealized beams are shown in Fig. 2.

C. Effects of Distorted Beam Profiles

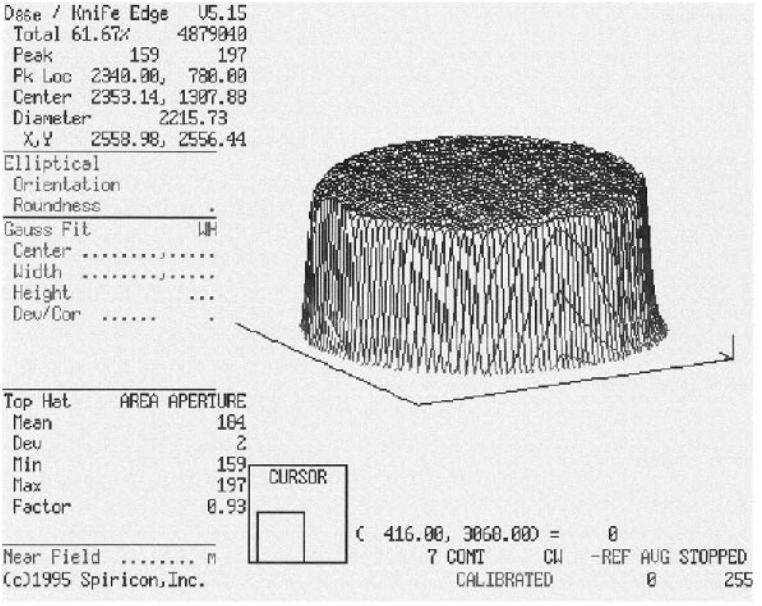
However, lasers rarely exhibit the most uniform irradiance profile. Sometimes Gaussian beams are highly structured, and often-intended flat top beams are non-uniform across the top, or may be tilted in energy from one side to the other. Figure 3 illustrates some real world examples of distorted beam profiles. For example, in Fig. 3a the highly structured beam would not focus nearly as well as the ideal Gaussian beam. The tilted flat top beam of Fig. 3b would not give uniform illumination as intended, and could cause severe distortion in the process for which it is being applied.

1. Scientific Applications

The significance of distorted beam profiles varies with the application. In scientific applications nonlinear processes are typically proportional to the irradiance squared or cubed. Thus a non-Gaussian profile may have a peak energy as low as 50% of what a Gaussian beam would have under the same

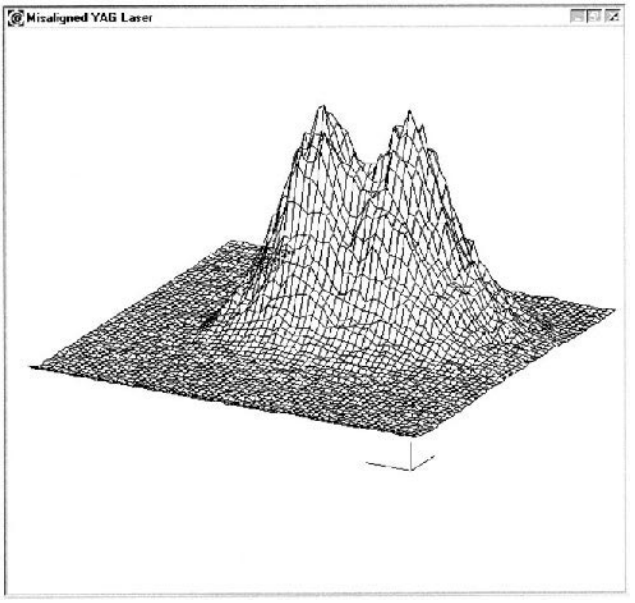


(a)

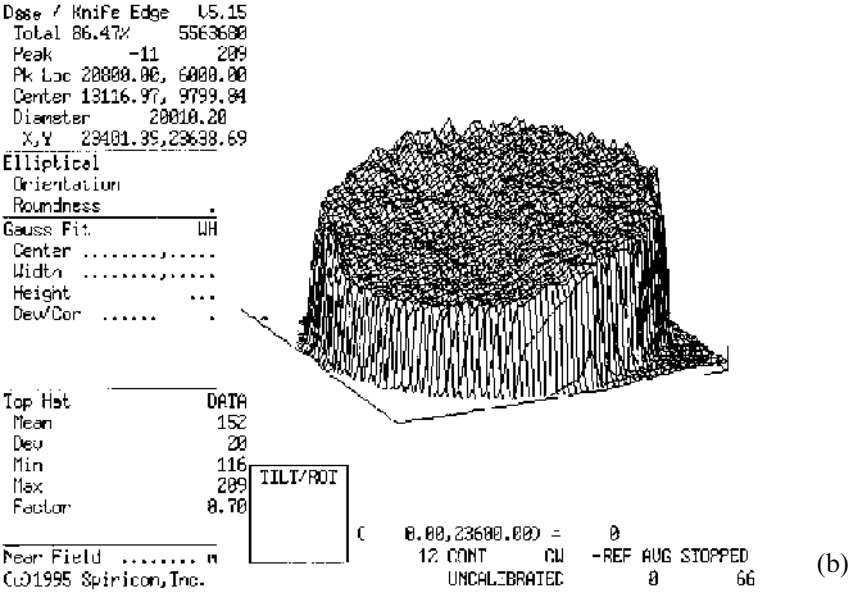


(b)

Figure 2 (a) Ideal Gaussian beam for highest concentration of energy; (b) ideal flat top beam for uniform laser illumination.



(a)



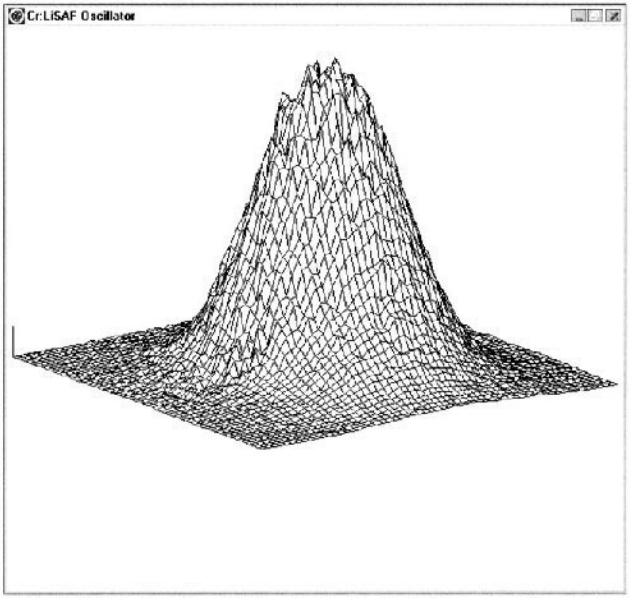
(b)

Figure 3 (a) Highly structured would-be Gaussian beam; (b) Tilted or non-uniform flat top beam.

conditions of total power or energy. Therefore the nonlinear process may deteriorate to 25% or 12% of what is expected. This is a 300% to 700% error on an experiment that should be accurate to within ±5%. Figure 4 shows the beam profiles of a Cr:LiSAF oscillator with subsequent amplifier outputs when the amplifier is properly aligned and when it is not.

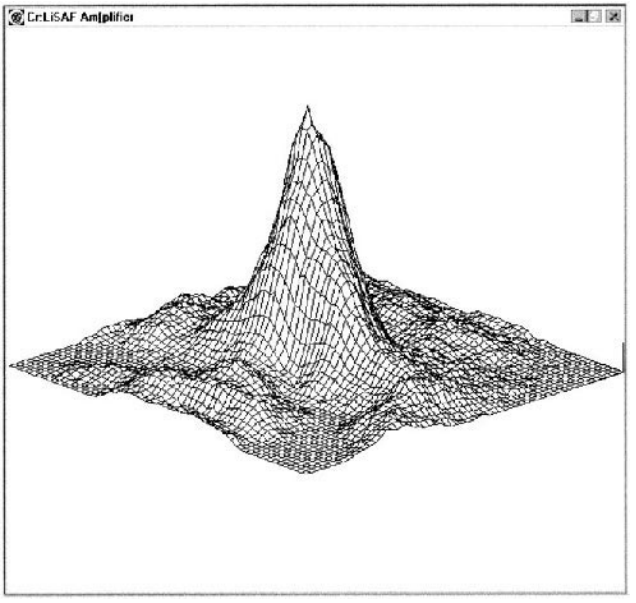
2. Instruments Using Lasers

Instruments using lasers, such as printers, fiber optics, communications, etc. require a high degree of control of the laser light to accomplish the intended task. The uniformity, pointing direction and stability, and mode pattern of a typical laser diode used in instruments can be dramatically deteriorated by misalignment of the collimating optics or mounting, causing the instrument not to perform as expected. For example, Fig. 5 illustrates a collimated laser diode beam being focused into a single mode fiber optic. In Fig. 5b the Z-axis of the focused laser diode is poorly aligned to the fiber, and much of the energy is fed into the cladding rather than to the inner fiber. Thus much of the energy does not come out of the central lobe. In Fig. 5c the Z-axis is

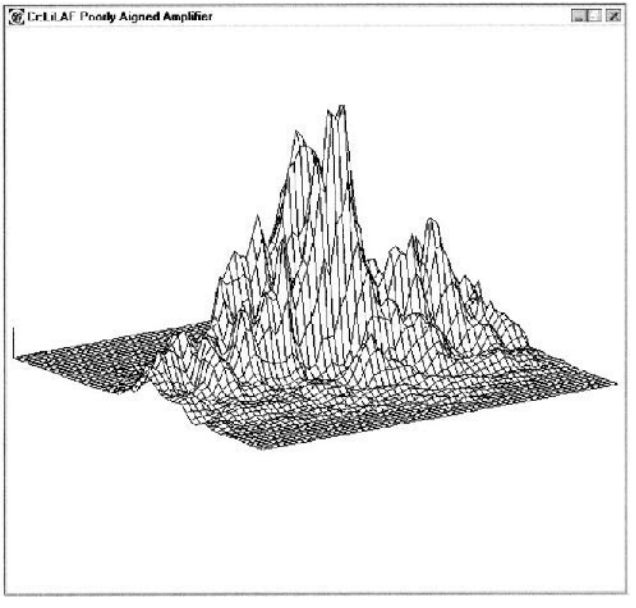


(a)

Figure 4 (a) Cr:LiSAF laser oscillator; (b) Cr:LiSAF laser with amplifier well aligned; (c) Cr:LiSAF Laser with amplifier misaligned.

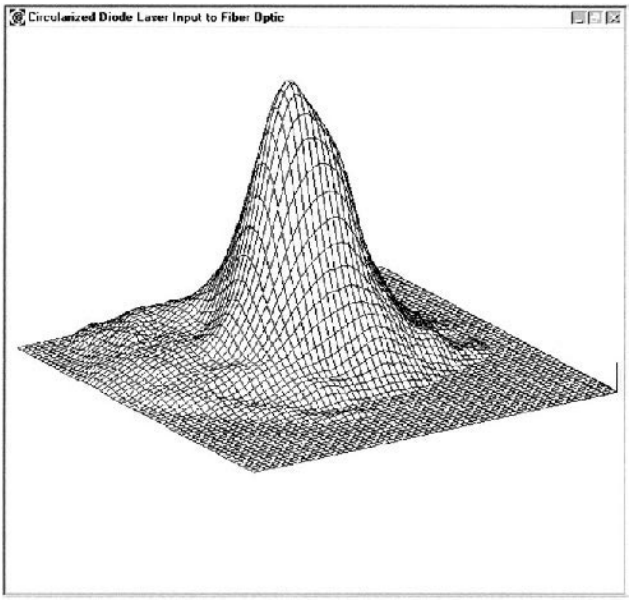


(b)

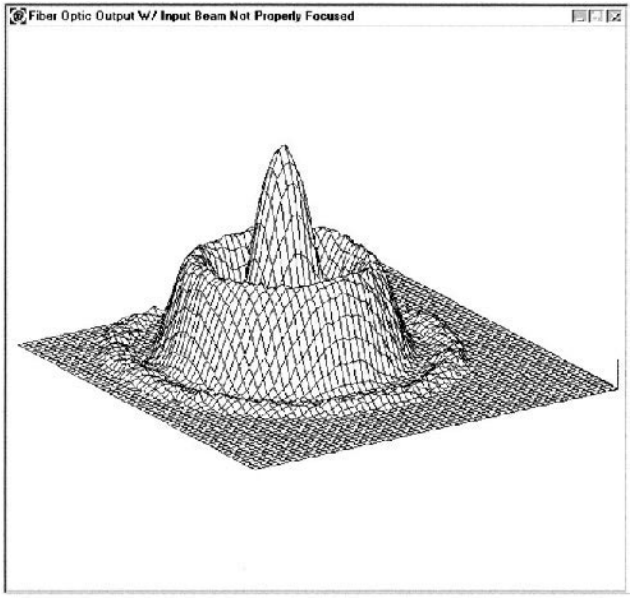


(c)

Figure 4 (*Continued*)

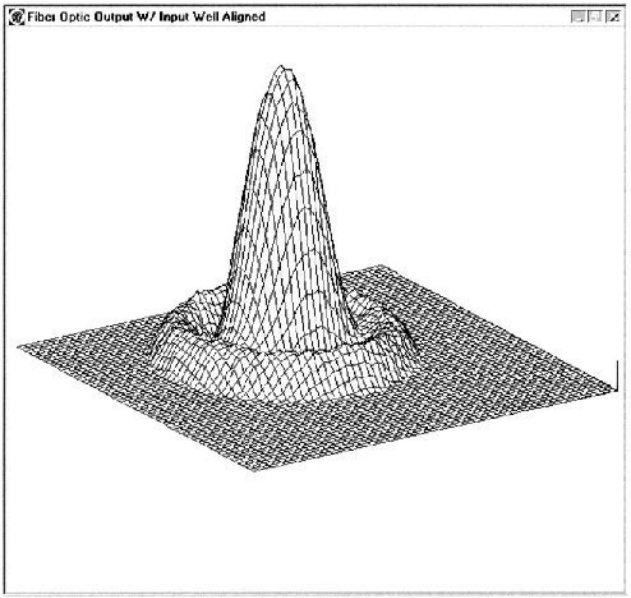


(a)



(b)

Figure 5 (a) Collimated laser diode beam; (b) Fiber output with diode poorly coupled into fiber optic (c) Fiber output with diode well-coupled into fiber optic.



(c)

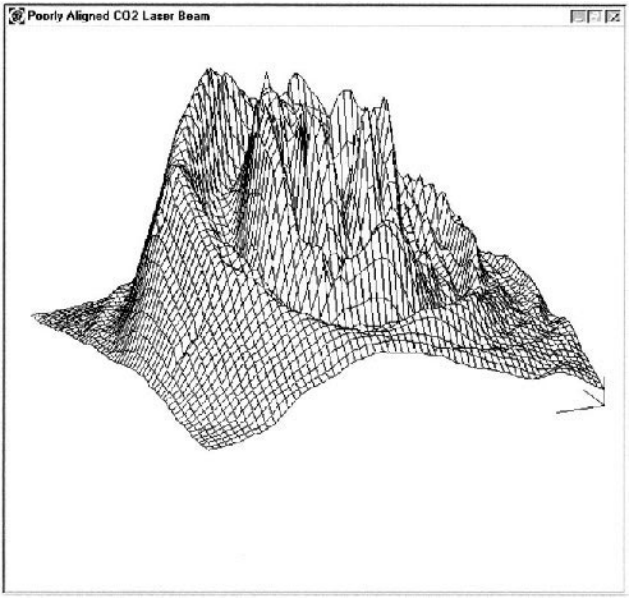
Figure 5 (*Continued*)

adjusted slightly, and the major portion of the laser beam is coupled into the propagating mode of the fiber optic.

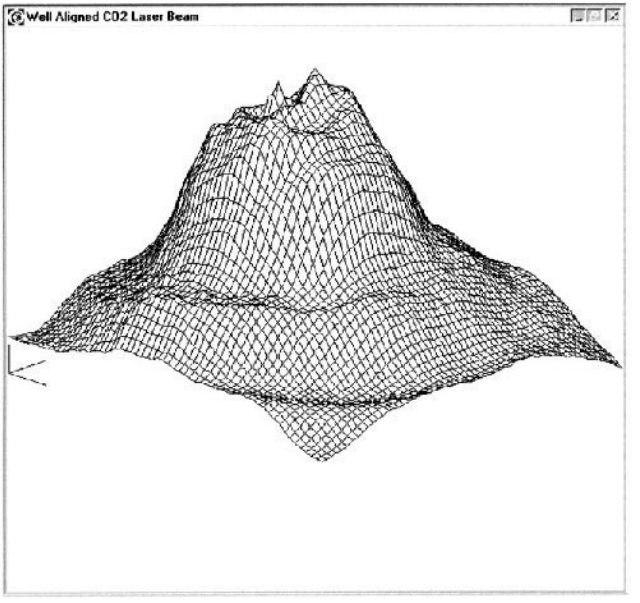
3. Medical Applications

There are many medical applications of lasers (15). One of these is photorefractive keratotomy (16), in which a flat top beam is used to make vision corrections. If the homogenizer producing the flat top is out of alignment and there is a 50% tilt in the flat top, the correction to the eye may be 4 diopters on one side of an iris, with only 2 diopters on the opposite side. The flat top beam in Fig. 2b would give expected results, whereas the tilted beam in Fig. 3b would cause severe difficulty. This could account for some PRK operations causing the patient to end up with non-correctable vision after the operation. The flatness of the laser beam is also critical in the removal of port wine stains and in other cosmetic surgery.

Tissue cutting and welding requires an extremely well-controlled irradiance density to accomplish the task properly. Finally, many medical applications, such as photodynamic therapy, use fiber optics delivery systems, and the efficiency of these systems is strongly affected by the initial alignment of the laser beam into the fiber, as shown in Fig. 5.



(a)



(b)

Figure 6 (a) Poorly aligned CO_2 laser cavity; (b) well-aligned CO_2 laser cavity.

4. Industrial Applications

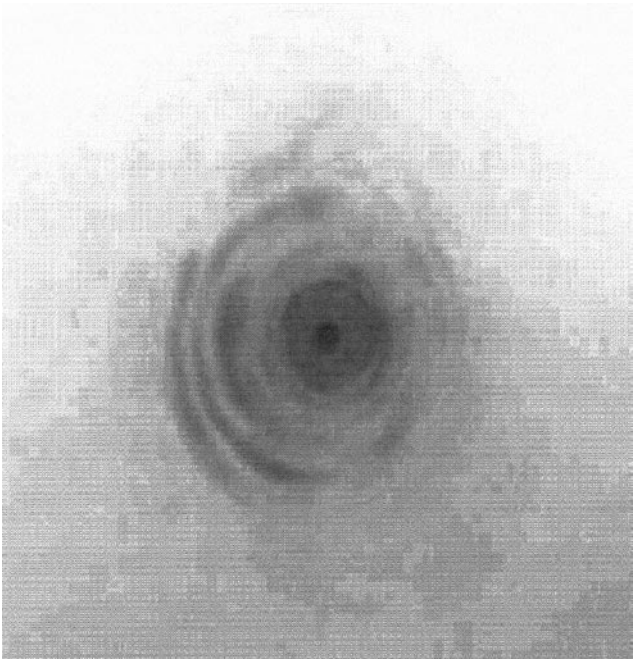
In industrial laser applications (17–19) most high-power Nd:YAG lasers and some CO2 lasers run multimode. The cutting, welding, and drilling efficiencies of these lasers are directly related to the beam profile. For example an Nd:YAG laser with a double peak can cause one cut width in the X direction, and a different cut width in the Y direction. Also, a beam with a poor profile can result in hole drilling of a different size than expected, and welding that is not as strong as necessary.

Figure 6 shows the beam profile of a poorly aligned and well-aligned CO_2 laser cavity. An industrial laser shop was using CO_2 lasers for scribing of ceramic wafers before breaking them into individual pieces. Most of the lasers in the machine shop gave extremely good results. However, one laser gave very inconsistent results, which caused poor scribing, and therefore very erratic breaking of the ceramic. The laser had been measured by non-electronic mode burns in wooden tongue depressors, which gave the appearance that the laser had a uniform output. However, as soon as the laser was measured with an electronic pyroelectric camera, the high structure in the laser beam of Fig. 6a became immediately evident. As soon as the technicians recognized this high structure, they began adjusting the laser cavity mirrors, and within a short time the beam was improved to that of Fig. 6b. Incidentally, the beam profile of Fig. 6b was similar to that of the other lasers that were operating well in this shop. Once the laser was tuned to the beam profile of Fig. 6b, it gave consistent results in scribing the ceramic.

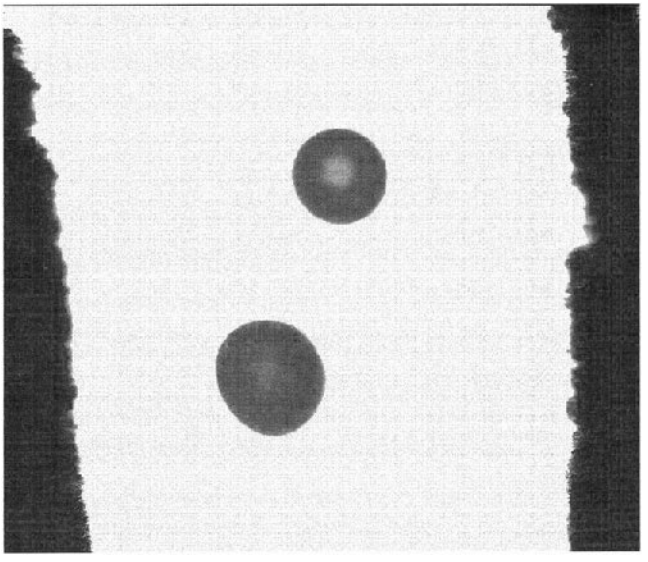
III. LASER BEAM PROFILE MEASUREMENT METHODS

A. Non-Electronic Methods

There are a number of non-electronic methods of laser beam profile measurement that have been used ever since lasers were invented. The first of these is observance of a laser beam reflected from a wall or other object. This is by far the simplest and least expensive method of measuring or observing a laser beam profile. The problem with this method is that the human eye is logarithmic, and can see many orders of magnitude difference in light irradiance. Even though it is logarithmic, the eye can only distinguish 8–12 shades of gray. Thus it is nearly impossible for a visual inspection of a laser beam to provide anything even approaching a quantitative measurement of the beam size and shape. Thus the beam width measurement by eye may have as much as 100% error. Figure 7a is a photograph of an HeNe laser beam being reflected off the wall. While photographic film has even less



(a)



(b)

Figure 7 (a) Reflected laser beam; (b) laser beam burn spots.

dynamic range than the human eye, the figure shows a very intense beam at the center, but a very large amount of structure far out from the center. This structure, which one might mistake as part of the laser beam, is less than 1% of the total energy in the beam. Yet the eye and the film are clearly able to discern this low energy. In addition, the eye cannot really distinguish structure in a laser beam with less than 2 to 1 magnitude variation.

Burn paper and Polaroid film are often used for making beam profile measurements. Figure 7b illustrates thermal paper having been illuminated by a laser beam. The burn paper typically has a dynamic range of only 3, unburned paper, blackened paper, and paper turned to ash. Sometimes skilled operators can distinguish between levels, and give a dynamic range of 5. The main objection to this manual method is that the spot size is highly subjective to the integration time on the paper. With longer exposures the center may not change, but the width of the darkened area could change $\pm 50\%$ or more.

Wooden tongue depressors and burn spots on metal plates are used in similar methods to those for burn paper. Sometimes the depth of the burn gives additional insight into the laser irradiance. Sometimes operators learn from experience with these burn spots, such as which beam tuning gives a specific result in a specific application. This might be tuning of one burn spot for cutting, and a different one for drilling holes. However, this measurement system is archaic, crude, and non-quantitative, subjec to the capability and experience of the operator, and therefore quite unreliable.

Fluorescing plates contribute to being able to see laser beams by converting UV and IR laser beams into the visible, where they can be seen by the human eye. These fluorescing plates have limited dynamic range, which adds to the dynamic range problem already described when viewing the reflected beam.

As shown in Fig. 8 Acrylic mode burns provide quite representative beam profiles of CO_2 lasers. The depth of the acrylic burns clearly shows the irradiance profile of the beam, and it is often possible to see mode structure. This gives an excellent visual interpretation of the beam profile. However, the acrylic mode burns are not real time, which makes it very cumbersome to tune the laser. They also do not enable one to see if there are short-term fluctuations in the laser beam, which is quite common in CO_2 lasers. It is possible that fumes from the burning acrylic may form a plasma at the center of the hole, which blocks the incoming CO_2 beam. Unless care is taken to have a fan blowing the fumes away, the acrylic mode burn will contain a hole in the center of the image that is non-existent in the beam. An additional problem is that the fumes from burning acrylic are toxic to humans, and care must be taken to exhaust these fumes outside of the work area.

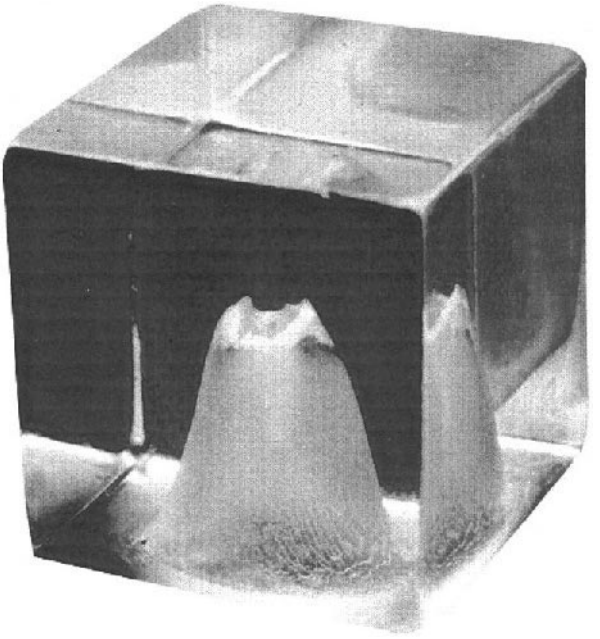


Figure 8 Laser beam acrylic mode burn.

B. Electronic Measurement Methods

For electronic laser beam profile analysis, it is nearly always necessary to attenuate the laser beam, at least to some degree, before measuring the beam with an electronic instrument. The degree of attenuation required depends on two factors. The first is the irradiance of the laser beam being measured. The second is the sensitivity of the beam profile sensor. Figure 9 shows a typical setup for the case where maximum amount of attenuation is required before the sensor measures the beam.

Typically when measuring the beam profile of a high-power laser, i.e., in excess of 50–100 W, the beam has enough energy to burn up most sensors that might be placed in the beam path. Therefore, the first element of Fig. 9, the beam sampling assembly, is typically used regardless of the beam profiling sensor. It should be noted, however, that there are some beam profiling sensors, to be described later, which can be placed directly into the path of a high-power beam of 10 kW and greater.

For mechanical scanning instruments, the beam sampling assembly is usually sufficient to reduce the signal from high-power lasers down to the level that is acceptable by such instruments. Also, if the original laser beam

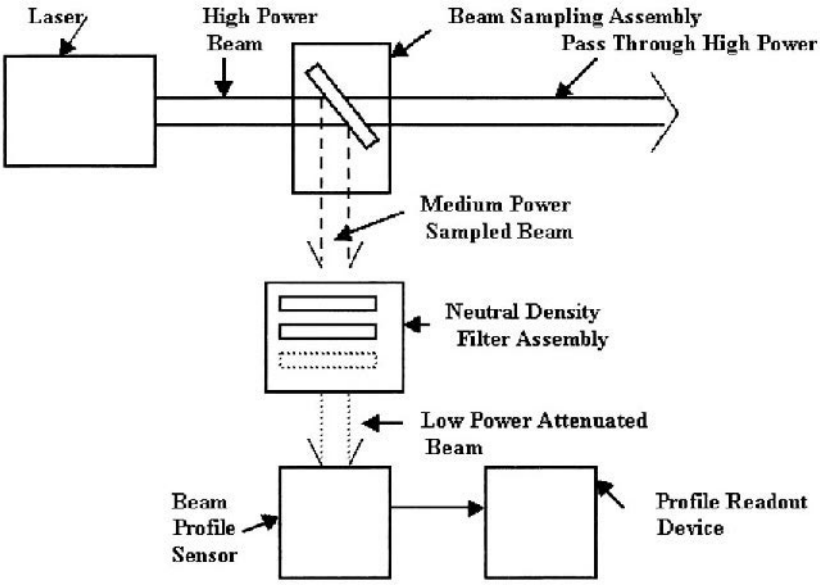


Figure 9 Optical setup for electronic laser beam analysis.

is in the range of 50 W or less, then often mechanical scanning instruments can measure the beam directly without using the beam sampling assembly. The reason that mechanical scanning instruments are able to be placed in the path of medium power beams is that they usually consist of a single-element detector with a rotating drum blocking the light from the sensor during most of the duty cycle of the sensor. The rotating drum either absorbs or reflects the incident laser beam during a significant part of the time, and thus high power is not placed on the sensing element.

For camera-based beam analyzers, the beam sampling assembly does not perform sufficient attenuation to reduce the beam power low enough for the camera sensor. In this case, usually a set of neutral density filters is placed in the beam path to reduce the power to the level acceptable by the camera. In some cases the beam power, even after reflection from one sampling surface, is too high, and would burn up neutral density filters. In this case a second reflecting surface is used to further reduce the incident power before impinging on the ND filter set. This is described in more detail in the Sec. IV, "Camera Based Instrument Description."

The neutral density filter assembly can be adjusted over a very wide dynamic range, as much as from ND0 to ND10 (or transmission of 1 to 10^{-10}) to reduce the beam power to a level acceptable by typical camera-

based systems. A detailed description of these ND filter assemblies is contained in Sec. IV.E.2.

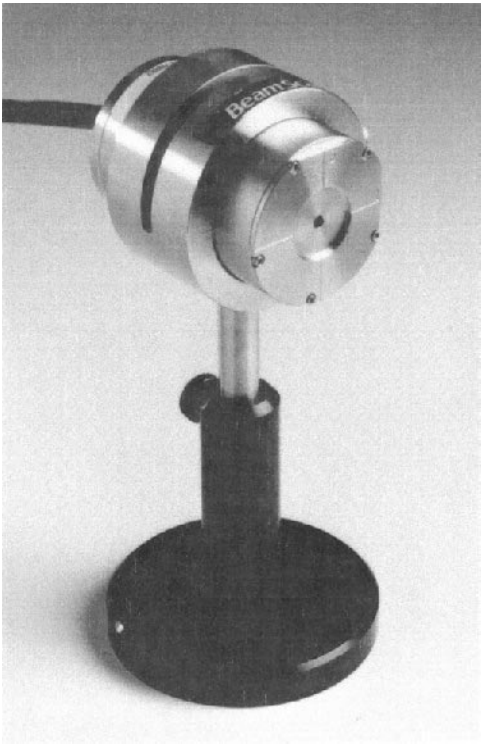
The fourth item in Fig. 9, the beam profile sensor, can be either one of the mechanical scanning devices described in the next section, or a CCD or other type of camera described in the following section. The beam profile readout device either consists of a dedicated monitor for reading out beam profile information, or in camera-based systems, a PC style computer and monitor.

1. Mechanical Scanning Instruments

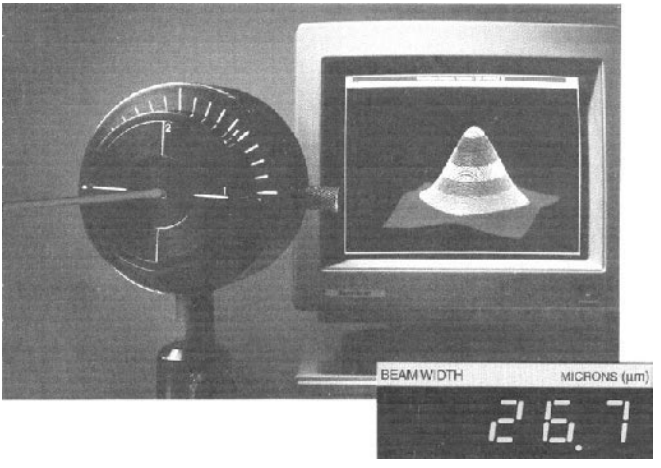
One of the earliest methods of measuring laser beams electronically was using a mechanical scanning device. This usually consists of a rotating drum containing a knife-edge, slit, or pinhole that moves in front of a single-element detector. This method provides excellent resolution, sometimes to less than 1 μm. The limit of resolution is set by diffraction from the edge of the knife-edge or slit, and roughly 1 μm is the lower limit set by this diffraction. These devices can be used directly in the beam of medium power lasers with little or no attenuation because only a small part of the beam is impinging on the detector element at any one time. Most of the time the scanning drum is reflecting the beam away from the detector.

However, mechanical scanning methods work only on CW lasers, not on pulsed lasers. Also they have a limited number of axes for measurement, usually two, and integrate the beam along those axes. Thus they do not give detailed information about the structure of the beam perpendicular to the direction of travel of the edge. However, there are rotating drum systems with knife-edges on seven different axes, which then provides multiple axis cuts to the beam. This can assist somewhat in obtaining more detailed information about the beam along the various axes. These beam profile instruments are adaptable for work in the visible, UV, and infrared by using different types of single element detectors for the sensor. In addition, software has been developed which provides illuminating beam profile displays, as well as fairly detailed quantitative measurements from the scanning system. This software now exists in the PC-based Windows operating system for easy use.

Figure 10a illustrates a commercial version of the knife-edge scanning slit beam profiling instrument. Figure 10b shows a typical Windows computer readout. Even though Fig. 10b shows a 3D representation from a two-axis scanning system, it could be misleading, since it does not really give information about the structure off-axis. Figure 10c illustrates a typical mechanical diagram of a scanning slit beam profiler.

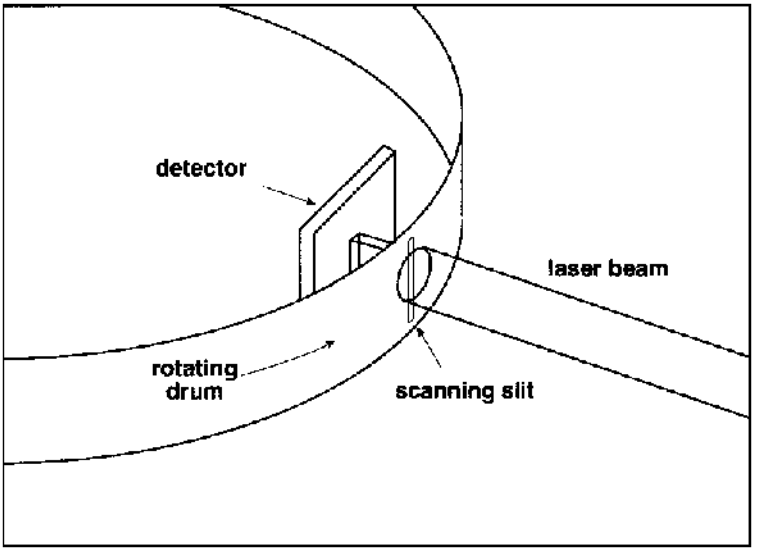


(a)



(b)

Figure 10 (a) Scanning knife-edge beam profiler; (b) Windows readout of scanning knife-edge system; (c) Mechanical diagram of scanning slit or knife-edge beam profiler.



(c)

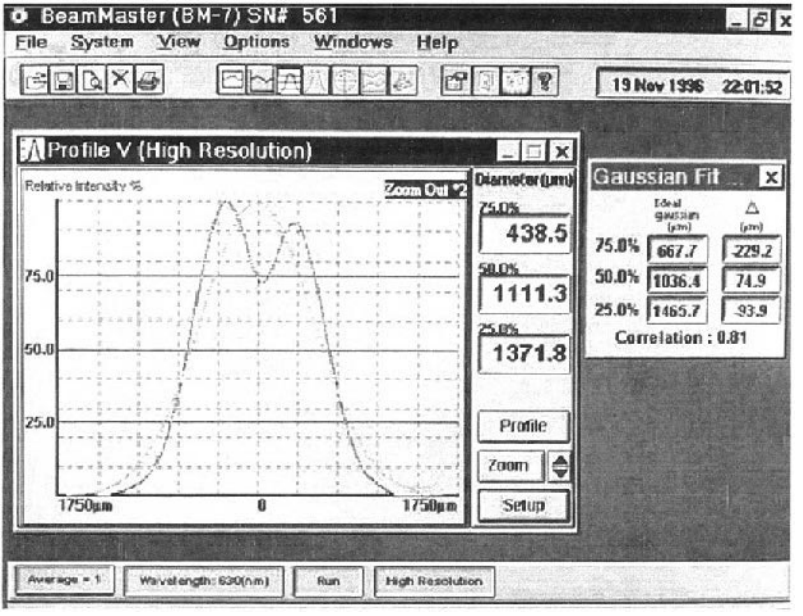
Figure 10 (*Continued*)

Figure 11a is a photograph of a seven-axis scanning blade system. Figure 11b illustrates a typical Windows readout from this seven-axis system. The mechanical layout of the multi-axis profiler is similar to Fig. 10c except that knife-edge apertures are used, and the angles of the knife-edges are varied so that the beam is scanned across multiple axes.

A variation of the rotating drum system includes a lens mounted in front of the drum (20). The lens is mounted on a moving axis, and thus focuses the beam to the detector at the rear side of the drum. By moving the lens in the beam, a series of measurements can be made by the single element detector that enables calculation of M^2. (A more detailed discussion of M^2 will be provided in a later section.) A photograph of this M^2 measuring instrument and readout is given in Fig. 12a. A mechanical layout of the instrument is shown in Fig. 12b.

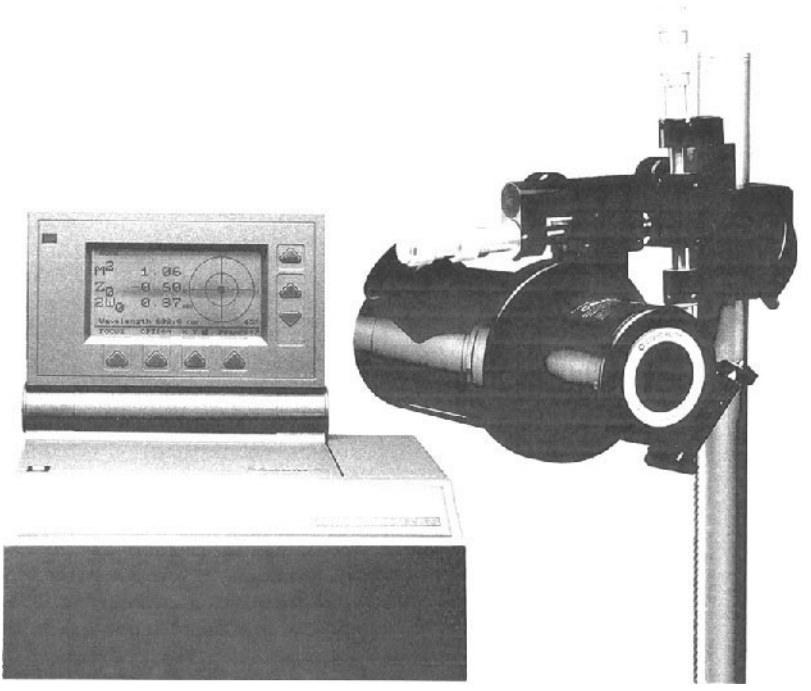
Another mechanical scanning system consists of a rotating needle that is placed directly in the beam. This needle has a small opening allowing a very small portion of the beam to enter the needle. A 45° mirror at the bottom of the needle reflects the sampled part of the beam to a single-element detector. The needle is both rotated in the beam, and axially moved in and out of the beam to sample it in a complete two-dimensional manner. The advantage of a rotating needle system is that it can be placed directly in the beam of high-power industrial lasers, both Nd:YAG and

(a)

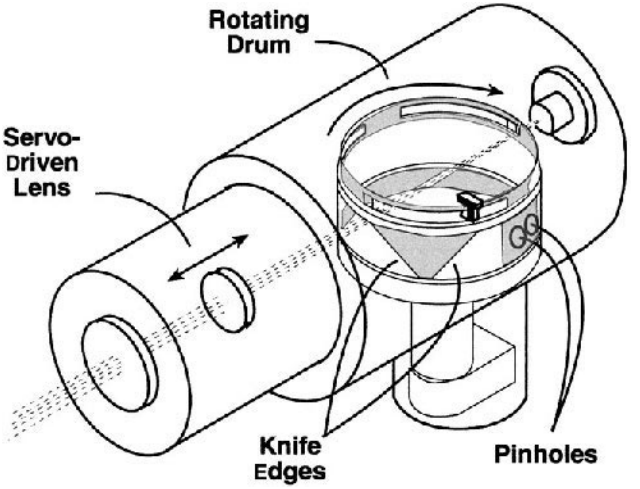


(b)

Figure 11 (a) Seven-axis knife-edge instrument; (b) typical readout from seven-axis system.

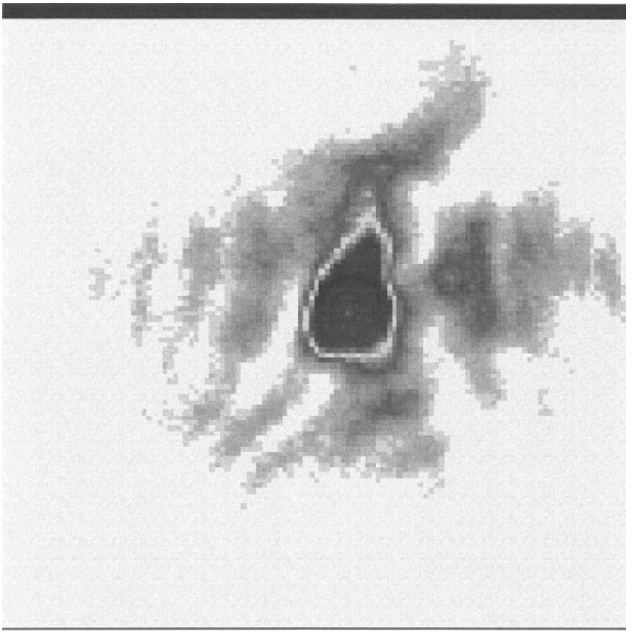


(a)

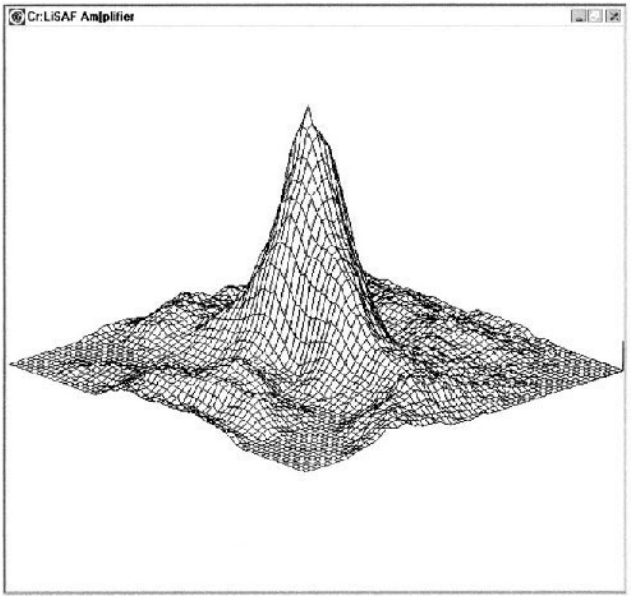


(b)

Figure 12 (a) M^2 measuring instrument and readout; (b) Mechanical diagram of M^2 measuring instrument.



(a)



(b)

Figure 13 Highly structured laser beam measured with a CCD camera and shown in both 2D and 3D views.

CO_2. It introduces a very small distortion in the beam, and so can be used in the beam while on-line. In addition, the translation of the needle can be made extremely small for focused spots, or large for unfocused beams. It has, however, some characteristics of the rotating drum mentioned above. Specifically, it is not very useful for pulsed lasers because of the synchronization problem between the position of the needle and the timing of the laser pulse. These rotating needle systems also have extensive computer processing of the signal with beam displays and quantitative calculations in Windows-based systems.

2. Camera-based Systems

Cameras are used to provide simultaneous, whole two-dimensional laser beam measurements. They work with both pulsed and CW lasers. There are silicon-based cameras that operate in the UV region to the near IR at 1.1 μm, which are useful with a great majority of lasers. In addition, there are other types of cameras that operate in the X-ray and UV regions, and other cameras that cover the infrared from 1.1 μm to greater then 400 μm. A drawback of cameras is that the resolution is limited to approximately the size of the pixels. In CCD cameras this is roughly 10 μm, and for most infrared cameras the size is somewhat larger. However, a focused spot can be re-imaged with lenses to provide a larger waist for viewing on the camera, which provides a resolution down to approximately 1 μm. Again the resolution for the camera system is limited by diffraction in the optics. Cameras have been interfaced to digitizers to connect the signal into a computer. Current computer software provides very illuminating two- and thre-dimensional beam displays as shown in Fig. 13. They also provide very sophisticated numerical analysis on the beam profile. A drawback of camera-based systems is that the cameras are extremely sensitive relative to the energy in laser beams. Thus nearly all lasers must be significantly attenuated before the cameras can be used for beam profile analysis.

IV. CAMERA-BASED INSTRUMENT DESCRIPTION

Complete instrumentation for a camera-based beam profiling system includes a modern computer, a framegrabber card to digitize the signal, and software for controlling the framegrabber card, displaying beam profiles and making quantitative calculations. A camera such as a CCD is used for the visible, or a pyroelectric or other camera is used for other wavelengths. Attenuation is almost always needed to either split off part of the beam, or at least attenuate the beam before going into the camera. Often the

beam is either too large or too small, and beam sizing optics or other techniques must be used to size the beam for the camera.

A. Computers

The computer used with beam profilers is now typically an advanced Pentium II IBM PC type with Windows 95, 98, or NT operating system. These Pentium II computers with processor speeds, at this writing, of up to 400 MHz, are so fast that a 512X480 pixel beam profile can now be displayed at the camera frame rate of 30 Hz. Just a few years ago, in 1996 when 100 MHz Pentium computers were the state of the art, a 128X120 beam profile could barely be displayed at 10 Hz. Thus modern computers becoming much faster and more powerful are having a significant impact on the performance of laser beam analyzers, especially in terms of being able to provide real-time displays.

The performance of these Pentium computers is enhanced by the modern PCI bus, in which the framegrabber is typically installed. In addition, the very large capacity of random access memory in excess of 32 or 64 MB assists the computer in rapid beam analysis processing. Very large hard disks in excess of 1 GB enable storage of high-resolution profiles, in large quantities, for analysis of time variation in lasers.

B. Frame Grabbers

The frame grabber is an analog to digital converter, normally placed inside the computer, which digitizes the signal from the camera, and presents it to the computer for signal processing. There are many brands of commercial framegrabbers available that have characteristics that are useful for beam profile analysis.

However, measuring parameters of a laser beam presents many unique challenges that do not exist in measuring parameters of other items such as are common in machine vision. As Tom Johnston of Coherent once said, "Measuring the width of a laser beam is like trying to measure the size of a cotton ball with a caliper." This difficulty in measuring laser beam width comes from the fact that laser beams never cut off at the baseline to zero, but almost always have energy that extends out into the wings. Thus processing of very low-level signals becomes critical in measurement of laser beams. Therefore it is essential that the framegrabber used for beam profile analysis has the capability for very accurate baseline adjustments, and the capability for special processing of the noise in the baseline of the camera. The section on quantitative measurements discusses the details of accurate baseline con-

trol, and sophisticated noise processing in order to make accurate beam profile measurements.

C. Beam Analysis Software

Laser beam analysis software now typically runs on Windows 95, 98, and NT. It provides extremely sophisticated signal processing, very detailed views of the laser beam, and highly accurate quantitative beam profile measurements. This software also must control the framegrabber and be able to control the baseline and noise of the signal coming into the framegrabber as described in the above paragraph. In addition to providing beam displays and calculations, software also provides the capability for versatile file and data transfer and management, either to the hard disk or to other computers, as a means of logging the laser beam characteristics. The capabilities of the software for very intuitive beam profile displays is given in more detail in Sec. V, and the quantitative calculations made by the software are provided in Sec. VI.

D. Cameras Used in Beam Profile Measurement

There are currently many types of cameras that are used in beam profile analysis. Each of these has advantages and disadvantages for various applications. The most common type of camera used for laser beam diagnostics incorporates a silicon-based sensor. These cameras consist of two types, charge injection devices, CID, and charge coupled devices, CCD. Silicon-based cameras cover the wavelength range from 190 nm to 1.1 μm when the normal glass window is removed, which would otherwise attenuate the UV. These cameras are fairly inexpensive, and since they cover the visible region, which includes many lasers, they are the most common cameras used. At slightly longer wavelengths lead sulfide vidicon tubes cover the region from the visible to 2.2 μm. Also a new solid state camera, InGaAs, covers the range from visible to about 2 μm. The most common cameras for infrared lasers of wavelengths longer than 1 μm are pyroelectric solid state cameras. These are fairly low resolution at 124×124 pixels and 100 μm elements, but cover the wavelength range all the way from 1 μm to beyond 400 μm, as well as work in the UV range from 190 nm to 350 nm. Formerly pyroelectric vidicon tubes were used for the visible to 12 μm range. However, their characteristics were never very satisfactory, and they are now seldom used for laser beam profiling. Finally, there are many cooled infrared cameras that can be used for the wavelength range from 1 μm to about 12 μm. These are discussed in the concluding paragraph of the next section.

1. Characteristics of Cameras

One of the initial camera technologies used for beam profile analysis was a charge injection device, CID. CID cameras are very versatile in that they have an X/Y readout rather than a sequential readout, and can thus be programmed to read out only a part of the camera matrix. This enables them to operate at high frame rates. In addition, CID cameras are being coated with phosphors that make them especially sensitive for UV radiation, and even X-ray radiation. The drawback of CID cameras is that it is difficult to synchronize the readout of the camera with the triggering of a pulsed laser. If the laser can be triggered from the camera, then it's a simple matter. But if the laser is free running, then distortion of the beam can occur if the laser pulse does not arrive at the optimum part of the scan cycle of the camera.

CCD cameras are the most common type of cameras used in beam profile analysis. Because they are used so extensively, a more detailed analysis of CCD cameras is provided in the paragraphs below. In addition to standard CCD cameras, there are CCD cameras with coolers attached to the sensor. This significantly reduces the noise of the CCD camera, and allows it a much greater signal-to-noise ratio in measuring laser beams. A typical uncooled CCD has about an 8-bit dynamic range, whereas cooled CCDs can be obtained with 10-, 12-, and even 16-bit signal-to-noise ratio. However, these cooled CCDs are a factor of 5–20 times more expensive than uncooled CCDs.

Lead sulfide vidicon cameras, PbS–PbO, are commonly used with near infrared lasers from 1 μm to 2.2 μm. The advantage of these vidicon tubes is that they are relatively inexpensive, for coverage at this wavelength range where silicon-based CID and CCD cameras do not function. One drawback of these vidicon cameras is that they have a long image lag of up to 1 s, and thus cannot be effectively used to track pulses occurring at faster than 1 Hz. In addition, they typically have a very high offset shading which reduces the dynamic range when it corrected by background subtraction.

Pyroelectric vidicons have been in existence for many years, and have been used for beam profile analysis, particularly of CO_2 lasers. They have the advantage of providing relatively high resolution for a CO_2 beam in the far infrared at 10.6 μm. However, they have many of the disadvantages of lead sulfide vidicons, in that they have a slight image lag in the readout, some nonlinearity, and some baseline shading.

Solid-state pyroelectric arrays have been provided for many years. Recently high-resolution models have become available. Because these cameras, as with CCD cameras, are commonly used in beam profile analysis, more details will be given in a later paragraph.

A new camera that is particularly useful for beam profile analysis of lasers in the 1–2 μm region are solid-state InGaAs. These cameras have the advantage of being solid state, and thus do not suffer from the problems of vidicon tube cameras. Probably their only drawback is that they are relatively expensive compared to CCD cameras, although they are in the same price range as the solid-state pyroelectric cameras mentioned in the preceding paragraph.

Another type of solid-state camera useful for far infrared is a camera containing an infrared bolometer array. These infrared bolometers are designed primarily to operate in the 8–12 μm range for thermal imaging, but nevertheless could be useful for long-wavelength laser beam analysis. In addition to infrared bolometers there is an infrared ferroelectric type of camera. This camera is designed for thermal imaging in the 8–12 μm region, but potentially useful for laser beam analysis. Both of these cameras provide higher spatial resolution than the solid-state pyroelectric camera mentioned above. However, the technology is still being developed to make them sufficiently stable to make accurate laser beam profile measurements.

Finally, there are cooled infrared cameras that could be used for laser beam analysis. This includes cameras made from material such as indium antimonide, InSb, and mercury cadmium telluride, HgCdTe. These cooled infrared cameras, however, require significantly more cooling than cooled CCDs, and typically use liquid nitrogen as the cooling mechanism. In addition, they are extremely expensive, costing between 20 and 50 times as much as CCD cameras, and 2–3 times as much as solid-state pyroelectric cameras. Another drawback of these cooled IR cameras is that they are made for thermal imaging, and thus are extremely sensitive. This requires additional attenuation over and above what would be required for uncooled, solid-state pyroelectric cameras.

2. Characteristics of Cameras Relevant to Beam Profile Analysis

There are a number of characteristics to evaluate in choosing one camera over another, or in specifying a given type of camera. The most significant characteristic is the wavelength response of the camera. This was alluded to in the section above. For example, CCD and CID cameras are the most useful cameras in the visible and near IR wavelength. A second essential factor is that the sensor on the camera be windowless to eliminate interference fringes between the two surfaces. Alternatively, if a window is required, then the window should be configured to minimize these interference effects. This can either be done by AR coating the window for the wavelength of use, or by making the window a bulk absorbing ND filter, which attenuates the reflection from the second surface going back and interfering with the

incoming irradiation of the first surface. The dynamic range of the camera is another factor for serious consideration. This subject is discussed in more detail when describing the CCD and pyroelectric cameras below.

Another useful feature to consider in choosing a camera is that some CCD cameras have electronic shutters. This enables the CCD to integrate only during a short time, for example, 1/1000 s. This enables the camera to select a single laser pulse out of a kiloherz pulse train, and display that single pulse. Integration is another useful feature, just the opposite of shutters, which enables very low light level signal to be accumulated on the camera, and obtain a higher signal-to-noise ratio.

The scan rate of the camera should be considered. In the US the standard for driving monitors is RS-170 at 60 Hz. In Europe and most of the rest of the world, the standard is CCIR for 50 Hz operation. Neither of these factors is very significant in measuring beam profiles because the framegrabber card is not affected by whether it is running at the US 60 Hz environment or the CCIR 50 Hz environment. However, if the camera is also going to be used for direct display on a monitor, then users should choose the type of camera for the part of the world in which they live.

Another factor that a user should be aware of, is that in nearly all cameras a single frame of the camera output, at say 30 Hz in the US, consists of two fields that are interlaced together. The two fields are provided at twice the frame rate. In some cameras these fields can be combined together for pulsed lasers, whereas in others, only one field of the camera is active during a single laser pulse.

Fill factor should be considered in the choice of a camera. Normally CCD and most other cameras have a relatively high fill factor, and thus do not lose signal in between the active parts of the pixel.

Shading can be a serious factor in some cameras. Shading is defined as a non-flat baseline offset. This means, for example, that a center region of the camera could have an offset of 1 or 2 digital counts out of 256, whereas perhaps one edge of the camera could have a baseline offset of as much as 5–10 counts. Even more seriously, some vidicon tube cameras have shading all the way around the edges, sometimes as much as 50 counts higher offset than the baseline offset in the center of the camera. Most CCD cameras have very little shading, but do have enough that it needs attention in terms of signal processing.

As mentioned above, image lag can be a serious problem with vidicon type cameras. Nearly all solid-state cameras completely read out the signal during each frame, and thus have no image lag.

Linearity of the camera output is another factor to be considered. Most solid-state cameras have nonlinearity of less than 1% over the specified dynamic range of the camera, which enables accurate profile measure-

ments. Lead sulfide vidicons are nonlinear, but have a typically known gamma. The gamma factor means that for a given change in the input signal magnitude there is not a proportional change in the output magnitude. If the gamma is 1, then the output is linear relative to the input. However, many vidicons have a typical gamma of 0.6, meaning that for a 10 times increase in input signal, for example, only a 6 times increase in the voltage output would occur. Beam profile software and hardware correct for this gamma, and convert the nonlinear signal from the vidicon into a linear signal.

A useful feature of some cameras is that they can be triggered externally. This enables a trigger pulse from the laser to synchronize the camera to the laser. A more common feature is that the camera is free running, and synchronization must be obtained by triggering the laser from the camera. When this is impossible, and the laser and camera run asynchronously, the user takes a slight chance that the camera will be in a reset mode when the laser pulse arrives. However, this typically occurs less than 1% of the time with most CCD type cameras.

The sensitivity of the relative types of cameras is another consideration because of laser attenuation needed. Almost all of the silicon-based cameras, such as CCD, CID, etc. have very similar sensitivity. The solid-state uncooled pyroelectric cameras are about six orders of magnitude less sensitive than CCDs, and thus require less attenuation than CCD cameras. Many of the other cameras mentioned under Sec. IV.D.1 ''Characteristics of Cameras'' have sensitivity somewhere between the solid-state pyroelectric and the CCD cameras, and thus the attenuation requirement falls somewhere between these two.

3. CCD Type Cameras

CCD cameras are the most common type of cameras used in beam profile analysis. There are very inexpensive CCDs typically used in camcorder and consumer type applications. These CCDs typically have a very large proportion of bad pixels, as well as a poor signal-to-noise ratio, and thus are not very suitable for laser beam profile analysis. Industrial grade CCD cameras have fewer bad pixels, and the electronics in the camera are typically designed to mask any bad pixels that do exist, which makes it easier for beam analysis software to process the signal.

The specifications given by the manufacturers for the signal-to-noise ratio of CCD cameras must be understood in order to properly measure laser beams. A typical specification is a signal-to-noise ratio of 50–60 db. This specification is arrived at by dividing the saturated signal level by the RMS noise. Thus 50–60 db implies a signal-to-noise ratio of 300–

1000. However, the RMS noise of a camera is equivalent to only one standard deviation, with significant noise peaks out to three standard deviations. When you consider both positive and negative noise transitions, the peak-to-peak noise becomes six times the RMS. Thus a CCD camera with a specification of 50–60 db signal-to-noise ratio typically has a peak to peak signal-to-noise ratio of only about 50 to 180. Thus 8-bit digitizers with 256 counts of resolution have typically been sufficient framegrabbers for beam analysis. It may be noted that with a 256 count digitizer, and a 50–180 signal-to-noise ratio, the bottom 2–6 counts from the digitizer will be noisy. There are cases when a 10- or 12-bit digitizer can provide an advantage over 8 bits, in beam profile analysis, in that the noise of the cameras is thus divided into smaller increments. This enables the software to provide signal summing and averaging to a greater accuracy in order to improve the signal-to-noise ratio from the beam profile measurement.

There are basically two types of CCD camera technology currently in use. One is called frame transfer, and the other is called interline transfer. In frame transfer cameras there is only one sensor site for both fields of the signal frame. Thus on a pulsed laser, since there is only one cell, this cell is read out during the first field, and no signal remains for the second field. Thus frame transfer cameras have only one half the resolution for pulsed lasers that they do for CW lasers. Some frame transfer CCD cameras have been shown to have signal response beyond the normal 1.1 μm cutoff of silicon sensors, out to 1.3 μm, even though the sensitivity is typically 1000 times less than it is at 0.9 μm. This slightly reduces the dynamic range of the camera when used in this wavelength range.

Interline transfer sensors have individual pixels for each field of the camera frame. Thus they maintain twice the resolution of frame transfer cameras with pulsed lasers. An interline transfer camera can pick out a single pulse from pulse rates up to 10 kHz with a 1/10 000 s shutter speed. Interline transfer cameras typically have higher speed shutters than with frame transfer cameras. However, a problem exists with interline transfer cameras in that the readout electronics are typically on the rear of the silicon wafer behind the sensor cells. For infrared lasers with wavelengths approaching 1.06 μm, the absorption of all the radiation does not occur in the sensor cells on the front, and some of the radiation is absorbed in the transfer electronics on the rear of the cell. This absorption of radiation creates a ghost image in the beam, which distorts the view of the beam profile. Even more significantly, it greatly distorts measurements on the beam, since this ghost image appears as a low-level energy in the wings of the beam. Thus interline transfer cameras are recommended for pulsed lasers wherein the wavelength is less than 1 μm. Frame transfer

cameras are recommended for YAG lasers at 1.06 μm, even though on pulsed lasers they have only half the resolution as interline transfer cameras.

4. Pyroelectric Solid State Cameras

Pyroelectric solid-state cameras (21) have been developed that cover the wavelength range from 1.1 μm to greater than 400 μm. These cameras are solid state with a very reliable and linear output. They have a flat baseline with no excessive shading and no image lag (as is present in pyroelectric and PbO–PbS vidicon tubes). Also, pyroelectric cameras interface to computers and software much as do CCD cameras, and provide the same viewing and numerical capability. However, pyroelectric cameras have a lower resolution of 100 μm per pixel, and only a 124 × 124 matrix. Figure 14 shows a pyroelectric camera with the output of a CO_2 laser displayed on a VGA monitor, provided directly from the camera microprocessor.

Pyroelectric solid-state cameras work well with pulsed laser radiation. However, it is necessary that the camera be triggered from the laser to synchronize the scanning. The reason for this is that the pyroelectric sensor is a thermal sensor, and after the signal is read out from the heating radia-

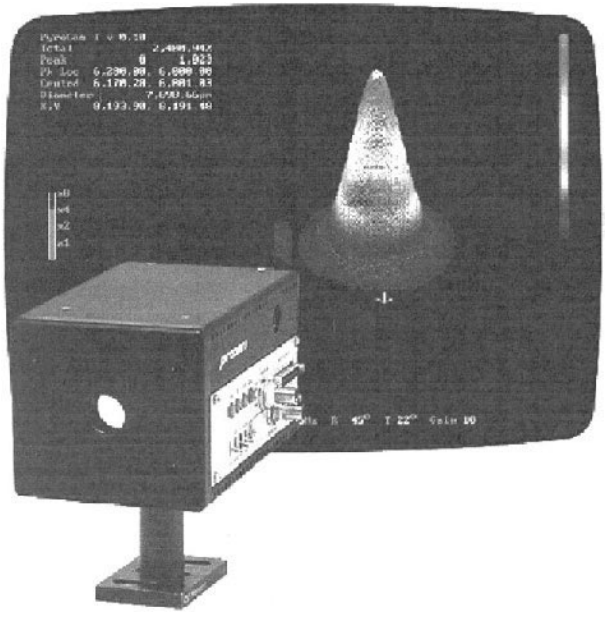


Figure 14 Pyroelectric camera and readout onto a VGA monitor.

tion pulse, the heated area of the sensor cools and generates a signal of the opposite polarity. It is necessary to read out this negative polarity signal and reset the sensor before the next laser pulse. This is done by having the camera sense a short series of pulses, calculate the interval between pulses, and then adjust the resetting scan to occur just before the next laser pulse. Depending on the pulse rate of the laser, whether it be single shot, very low frequency below 5 Hz, intermediate frequency between 5 and 50 Hz, or high frequency between 50 Hz and 1 kHz, this resetting operation is performed differently in the pyroelectric camera.

For CW operation the sensor must be mechanically chopped to provide an alternating heating and cooling cycle. This typically is done with a 50% duty cycle between heating and cooling, and is normally performed by a rotating chopper blade. The chopper blade is usually incorporated into the camera such that the blade crosses the camera sensor from top to bottom. The camera readout is then triggered to read out each row from the camera just as the blade crosses that row of pixels. In this manner every row of pixels in the pyroelectric sensor has the same integration time, is read out immediately after being covered or uncovered, and thus gives optimum uniformity of the signal. If chopping synchronization is not performed properly, the pyroelectric sensor can give a very distorted output signal.

Because the pyroelectric sensor is a thermal device, there is some distortion that occurs from thermal and electrical crosstalk. Thermal crosstalk occurs when some heat spreads from one element into adjacent elements. This is minimized by either chopping at a relatively high frequency, and with pulsed lasers by resetting the array just prior to an incoming laser pulse. With chopped radiation the maximum distortion occurs in the corners of the array, because the chopper blade is not parallel to the row on the top and bottom edges. Even so, the non-uniformity in the corners is only about 5%. At the same time, the linearity and uniformity of the corner areas usually has the least impact on measuring a beam. There is some distortion from electrical crosstalk coupling between pyroelectric detector elements. However, with high-resolution elements in current pyroelectric solid-state arrays, this electrical crosstalk is typically less than 5%. Since laser beams do not typically have very sharp irradiance changes from one element to the next, this crosstalk is seldom a problem in obtaining relatively accurate analysis of the structure in a laser beam.

E. Laser Beam Attenuation

Laser beam profile measurements are made on lasers that vary from less than 1 mW to greater than 10 kW average power. This typically corresponds to a power density of less than 10^{-1} W/cm^2 to greater than 10^5 W/cm^2. A

CCD camera typically saturates at a power density in the range of 10^{-7} W/cm^2. Solid state pyroelectric cameras typically saturate at approximately 1 W/cm^2. Thus the necessary attenuation arranged for CCD cameras varies from 10^5 to 10^{12}. For pyroelectric cameras the attenuation range is a little more modest at about 10^4. This laser beam attenuation is usually performed by one of two methods. The first is using a beam splitter to pick off a small percentage of a beam, allowing the main part to pass through the beam splitter. The second method is inline attenuation in which the beam is reduced in power by the absorption of neutral density filters.

1. Beam Pickoff

The first step in attenuating a high-power laser beam is to pick off or sample a small percentage of the beam from the main beam, without affecting the beam profile of the sampled beam. There are basically three ways to perform this pickoff. The most common is to have a beam splitter that is mostly transmitting and partially reflecting. The beam splitter is typically put in the beam at 45°, so that a small percentage of the beam is reflected at 90° to the incident beam. However, this beam sampling surface can be placed at any angle, and there is an advantage to placing the pickoff surface nearly perpendicular to the beam so that the reflection becomes less polarization sensitive.

The other type of pickoff is to use a mostly reflecting and partially transmitting surface. In this case the surface is placed in the beam at an angle to reflect the majority of the beam, and then transmit a small part through the surface to be measured by the beam analyzer.

A third method of beam pickoff is to use a diffraction grating. This can either be a reflecting or a transmitting type. In the transmitting type of diffraction grating the beam is typically incident upon the grating perpendicular to the surface, and most of the beam passes directly through the diffraction grating. However, a small percentage of the beam is transmitted at an angle offset from the output angle of the main pass through beam. The portion diffracted typically has multiple modes, whereas for example, 1% of the beam may transmit at, for example, 15° from the emitting main beam. Second-order diffraction may be 0.01% at 30°, and even a third-order beam may be 10^{-6} of the input beam at 45°. The angle and the diffraction percent depend upon the manufacturing characteristics of the diffraction grating, as well as the wavelength of the beam incident upon the grating.

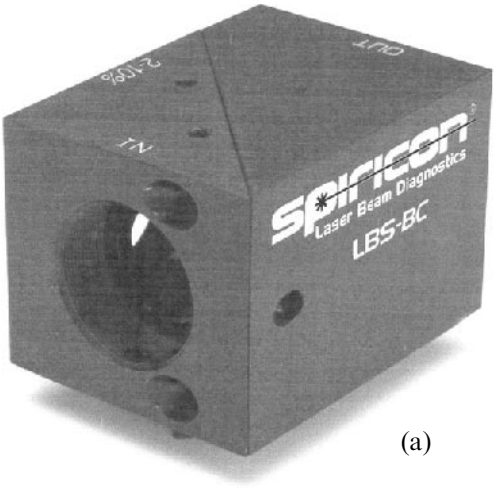
A reflection type of diffraction grating works in a similar manner, except that the beam incident on the diffraction grating is at an angle other than perpendicular. For example, instead of at 90° to the plane of the grating, it may be 30° from normal incidence. The main reflected beam is

then reflected at 30° from normal incidence in the opposite direction. Now the first-order, second-order, third-order beams are reflected at angles other than the angle of the main reflection. As in the case of the transmitted reflection beam, these refracted and attenuated beams may be 5°, 10°, or 15° away from the main reflected beam, and are typically 1%, 0.01%, or 10^{-6} of the main beam, etc. The diffracted low-intensity beam maintains all of the beam profile characteristics of the main beam. This small intensity beam can then be used for beam analysis.

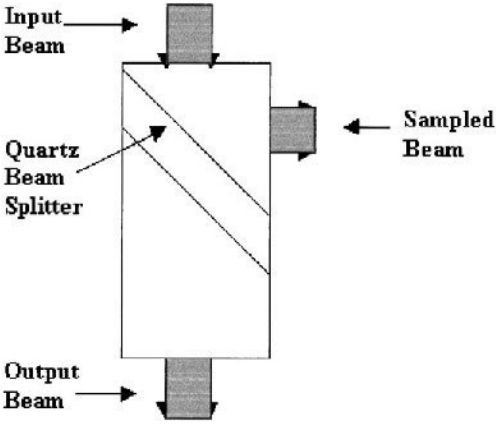
The most commonly used beam pickoff surface is quartz, usually used at a 45° angle to the incoming beam. If the quartz is not AR coated, it reflects an average of 4% of the beam per surface. However, at 45° the quartz becomes polarization sensitive, and one polarization is reflected at about 2%, and the other can be as high as 8–10%. Thus the reflected sample beam does not truly represent the incoming laser beam. This problem can be solved by placing a second quartz surface in the path of the initially sampled beam, but angled in a perpendicular plane that reflects the two polarizations in the opposite way to the first surface. (That is, the first surface may reflect the beam 90° in the horizontal, and the second surface reflects 90° in the vertical.) After two such reflections the sampled beam once again has the same characteristics as the initial beam.

The quartz sampling plates have two configurations. One is a wedge, so that the back surface of the quartz reflects at a different angle than the front surface, and keeps the beams from the two surfaces from interfering from each other. The other configuration uses a very thick, flat quartz plate such that the reflection from the back surface is displaced sufficiently far from the front surface reflection, that it does not overlap. Flat pickoffs have the advantage that the throughput beam, while being slightly displaced in position from the input beam, exits the quartz flat at the same angle as the entrance, and is not distorted. Figure 15a shows a commercially available quartz reflecting device using a thick flat as the reflector. The mechanical layout is shown in Fig. 15b. With the wedge, the exit beam is displaced in position and angle, as well as being slightly elongated. Thus if a beam pickoff were to be used in process and left permanently in place, then the flat would have superior characteristics to a wedge. Figure 16a shows a commercially available attenuation device using a wedge as the reflecting mechanism. Figure 16b shows the mechanical layout of the device, with indents for mounting the wedge on the right side and slots for ND filters on the left side. For infrared lasers ZnSe commonly replaces the quartz as the reflecting material. ZnSe can be AR coated for the specific wavelength of interest, and achieve reflection lower than 1% per surface.

An advantage of reflecting gratings is that they can be made from metal, and then the rear cooled with water to enable them to withstand



(a)

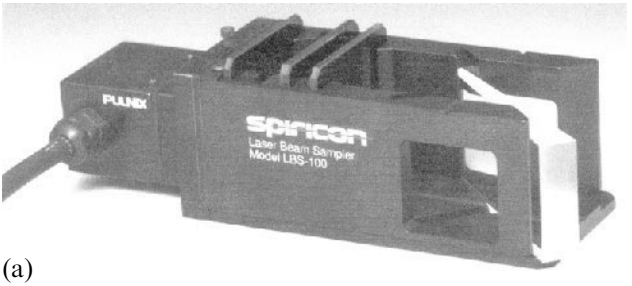


(b)

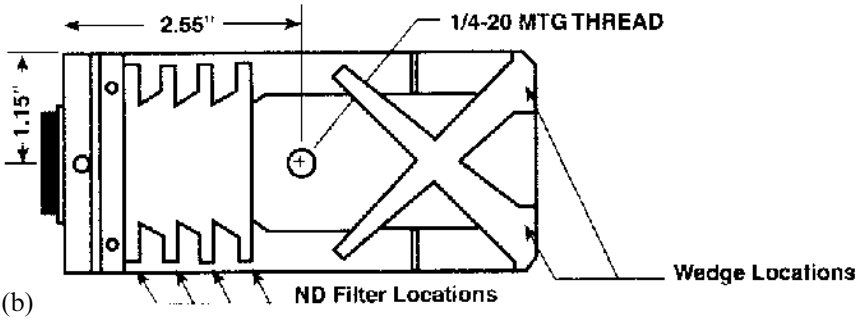
Figure 15 (a) Beam cube reflecting beam splitter; (b) mechanical layout of beam cube.

very high powers. Some gratings are of a transmitting type, either from quartz for visible radiation, or from ZnSe for infrared radiation, and have the advantage that the main beam continues along the same path as its entrance.

Sometimes a thin pellicle, of 10–50 μm thick, is used for a beam sampler. This is so that the rear reflection is so close to the front side that the interference effects can sometimes be negligible. However, a pellicle as thin as 10 μm can still cause interference fringes that could be seen with a 10 μm pixel camera.



(a)



(b)

Figure 16 (a) Combined beam splitter and ND filter holder; (b) mechanical diagram of combined beam splitter and ND filter holder.

Finally, for industrial Nd:YAG lasers a good pickoff scheme is the use of the dichroic mirror that is normally employed as a turning mirror for the laser. This dichroic mirror is typically made of quartz or fused silica with an AR coating and placed at 45° to reflect nearly the entire 1.06 μm beam at 90° from the input. The dichroic mirror is configured so that visible light passes through the filter, so an operator can either see through the filter to the work surface, or a camera can be mounted behind the filter to monitor the industrial process being performed. These dichroic filters transmit a small percentage of the YAG laser beam directly through the filter, so that it is used as a sampling mechanism. Dichroic filters used in this manner are also very polarization sensitive, and so once again, two filters must be placed at 90° to each other in order to obtain a true representation of the input beam.

2. In-line Attenuation

There are a number of methods of further attenuating a laser beam once reflection has reduced the power or energy low enough that it does not

damage the in-line attenuators. One type of in-line attenuators consists of glass or quartz with a broad band neutral density reflecting surface coating. When these neutral density filters are made of quartz, they are particularly useful for ultraviolet radiation. However, when the multiple surface reflecting neutral density filters are used in conjunction with each other, there is the danger of causing interference between the multiple surfaces. This interference can cause interference fringes, which completely distort the ability to measure the beam profile. This problem can be somewhat alleviated by tilting the filters so that the reflected beam bounces away from the camera sensor.

Another type of surface-reflecting ND filter consists of circular variable filters, in which the attenuation varies around the surface of a circular disk. This type of filter is very useful for single-element detectors, but is not very useful for beam profilers, in that the attenuation is continuously varying, and therefore will attenuate one side of the beam more than the other.

The more common in-line attenuation filters for beam profile analysis consist of bulk absorbing neutral density filters. Bulk absorbing filters are usually made of BK7 glass impregnated with an absorbing material. The range of attenuation achievable with these filters varies from a neutral density, ND, of 0.1 to an ND of greater than 4. The ND number is defined by

$$\mathrm{ND} = \log(1/T) \tag{1}$$

where T is the transmission ratio of output divided by input.

Since the absorption is within the material, there is very little danger of reflection from one surface bouncing back and interfering with the reflection from the other surface. Nevertheless, when two filters are stacked together, the back surface of one, and the front surface of the next, need to be slightly angled so that interference does not occur in this region where there is no attenuation between the surfaces. Bulk absorbing filters are very useful for the entire visible and near infrared. However, they cut off at about 350 nm in the UV, and are not useful for excimer lasers. They also begin to change their attenuation characteristics at 1 μm in the infrared, and then cut off completely between 2 and 2.5 μm.

Bulk absorbing ND filters are commercially available in a number of forms. One common form is simply a flat plate, 2 inches (50 mm) square, which can be stacked one after another in any mechanism to hold them in place. Figure 16 shows 2 inch square ND filter flat plates used to attenuate the beam. This same instrument could also accommodate surface reflecting filters for the UV. A second commercially available type is to have individual round filters mounted on a wheel so that the wheel can be turned, enabling the user to change attenuation simply by rotating the wheel.

Often these wheels can be stacked one behind another, so that multiple ND filters can be selected. Typical filters have an attenuation range between 0 and 10^{-10}. Figure 17 shows a commercially available rotating wheel neutral density filter set.

A third type of bulk absorbing ND filter consists of two filters made in the form of wedges. An individual wedge would be like a circular variable filter, and attenuate more on one side than another. However, an opposing wedge is placed behind the first wedge, and the entire beam passes through the same amount of attenuating material. These wedges enable a user to make continuous changes in attenuation in small increments, which can be very convenient. However, in some instances beam distortion has been observed from these filters.

None of the in-line filters discussed thus far are useful for infrared lasers beyond 2 μm. It turns out that for CO_2 lasers, the use of CaF_2 flats is very useful. A 1 mm thick CaF_2 plate absorbs roughly 50% of 10.6 μm radiation impinging upon it. Thus by stacking CaF_2 flats, CO_2 lasers can be

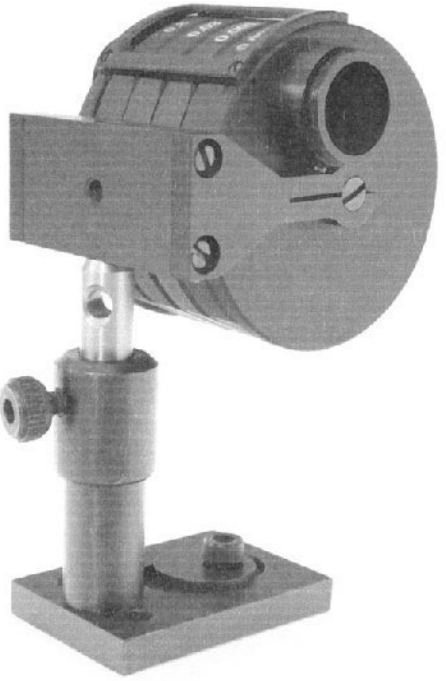


Figure 17 Rotating wheel ND filter set.

attenuated so that the signal is reduced in fine increments to within the range of the infrared camera.

Crossed polarizers are popular for in-line attenuation of laser beams. However, it would be difficult to assure that the cross polarizers are attenuating each polarization of the beam identically. Therefore, they are not commonly used in beam profile analysis, even though they work very well to attenuate beams for single-element power measurement.

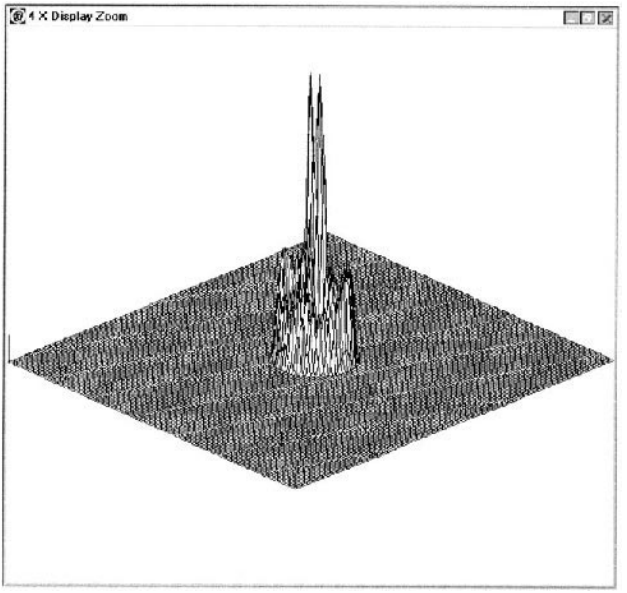
Finally, the last method of attenuating a laser beam is to allow the beam to impinge upon a scattering surface. The beam must first be attenuated by beam sampling, so it does not burn or damage the scattering surface. Once the beam impinges upon the scattering surface, the camera can use a lens to image the reflection of this surface. The image reflection is typically very representative of the beam profile. A problem that can exist is that speckle always occurs from scattering surfaces. Speckle is a situation in which the roughness of the scattering causes interference to create both bright and dark spots in the image reflection. Having the scattering surface move at a rate faster than the camera integration frame rate can solve this problem. There exists a commercial product called a "speckle eater," which is simply a scattering surface mounted to a small vibrating motor. An advantage of imaging scattered beam reflection is that the iris in the camera lens can be used for attenuation to achieve a fine degree of beam irradiance reduction.

F. Beam Size

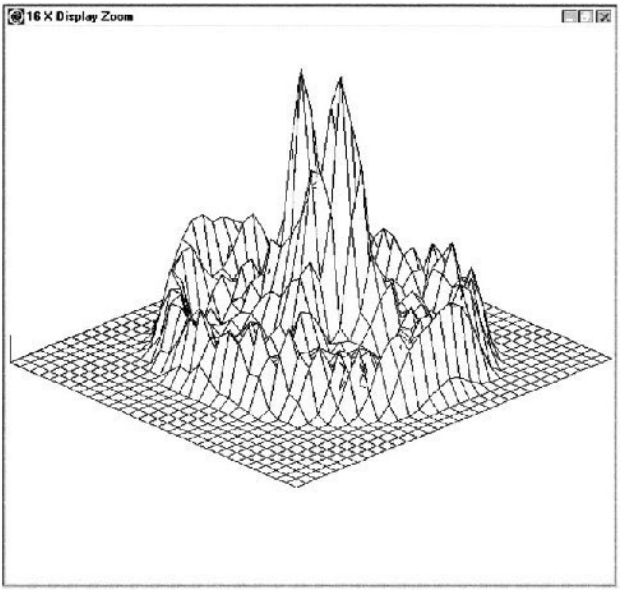
Laser beams typically vary from 1 mm in beam width to 40 mm, and in many applications, much larger. Focused laser beam spots can be as small as 1 μm in width. Since camera pixels, at the smallest, are approximately 8 μm, cameras are not very useful for measuring focused spots. In addition, typical commercial grade cameras have an overall sensitive area of roughly 6 mm, with 10 mm being the size of 1000×1000 large area CCD cameras. Thus there are many cases when the beam is much too small to be measured with the camera pixels, and other cases when the beam is much too large to fit onto the camera sensor.

When the beam is too small for the pixels on a camera system, one of the most straightforward solutions is to use one of the mechanical scanning devices instead. Both the rotating drum and the rotating needle systems can measure small beams. However, these systems are still plagued with the problem that they do not work with pulsed lasers, and do not give instantaneous whole beam analysis.

A second solution that is used especially when looking at a laser emitter such as a laser diode is to use a microscope objective and focus



(a)



(b)

Figure 18 (a) Focused laser diode beam shown at a 4× computer magnification and (b) 16× magnification.

the camera onto the emitter surface. This images the laser output aperture on the array.

An indirect method of measuring a small focused spot is to allow the beam to go through focus, then use another lens to collimate the beam. A third lens with a long focal length then refocusews the beam to a much larger waist that can be resolved by the camera pixels. If the beam is not a tightly focused spot, but rather a long waist, then a beam expander can perform the same function to increase the size of the beam. Finally, a small focused spot can be scattered from either a reflecting or transmitting surface, and imaged with a camera lens. Difficulty with this technique is that if the spot is very small, it is difficult to obtain scattering surfaces with structure small enough to accurately scatter the beam, rather than simply reflect off one of the facets of the scattering surface. Magnification of the beam can also be achieved in the computer, however at the loss of resolution. An example of the magnification capability in the computer software is shown in Fig. 18.

When the beam is too large for the camera, the first solution is to use a beam expander in reverse. A beam expander can typically give beam reductions in the order of 10 to 1. Thus a 5 cm beam could be reduced to 5 mm, which would fit nicely on a CCD camera. A second method with large beams is to use large area sensors. This is limited to approximately 1 cm^2 for silicon type sensors.

Finally, the most common method of viewing very large beams is to reflect the beam from a scattering surface and image the beam with a lens. A scale can be used to calibrate the pixel pitch of the lens/camera system. This is the same technique as is used for beam attenuation, but now the primary purpose is to be able to image a large area beam, rather than attenuate a large energy. All the same techniques described above must be used to minimize speckle and other problems.

V. VIEWING BEAM PROFILES

A tremendous amount of information can be gained about the beam profile simply by being able to clearly see it on a computer screen. Mode structure and distortion of the beam are immediately recognized. Examples are a Gaussian beam distorted into an elliptical shape, or the introduction of spurious multimode beams into the main beam. The beam splitting up into multiple spots or clipping of the beam on an edge is immediately seen. In top hat beams an electronic display can show hot and cold spots in the flat top, as well as distortion in the ideal vertical sides of the beam.

A. Two-dimensional Beam Profile Displays

A two-dimensional (2D) view of the beam enables the user to see the entire beam simultaneously. A false color or a grayscale plot is given which enables the user to tell intuitively which are the hot and cold spots in the beam. Cross sections through the beam, set either manually or automatically at some part of the beam, introduce displays of beam irradiance in the vertical axis, which help interpret the 2D display. Shown in Fig. 19 is the 2D display of a beam profile with the cross section vertical displays drawn through the peak of the beam. The cross section profiles can be drawn at any other part of the beam as well, or rotated from the X/Y-axis to the major/minor axis of an elliptical beam.

Sometimes color can have a significant effect in providing intuitive beam profile information. Other times a grayscale image can show informa-

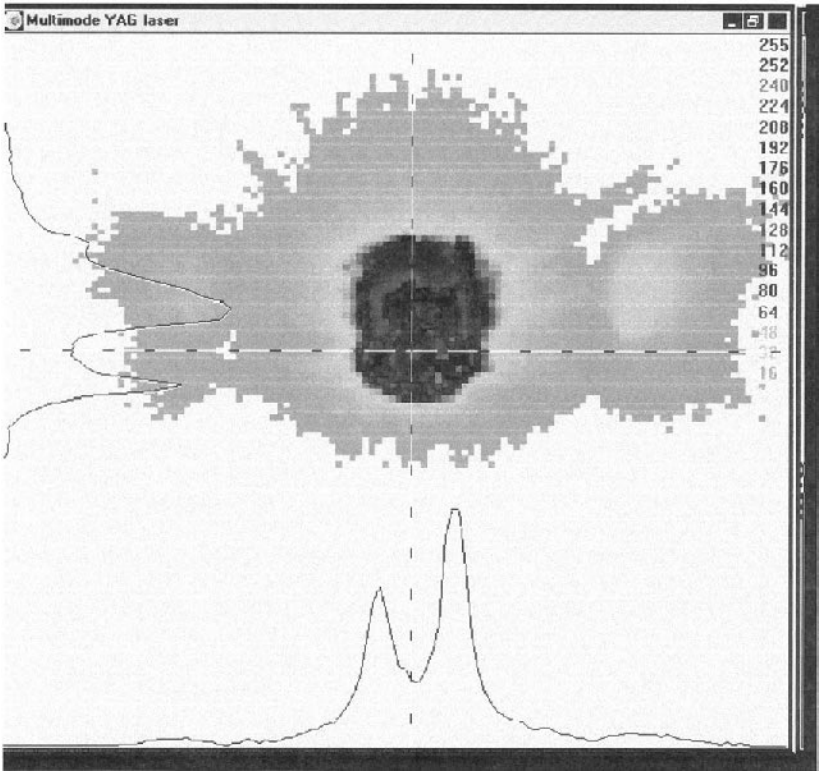


Figure 19 2D Beam profile display with cross section on the X/Y axis.

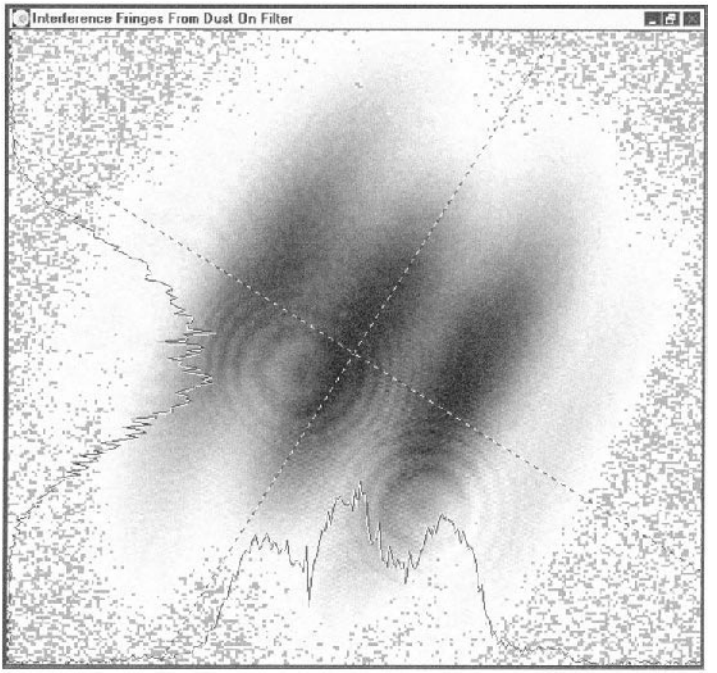
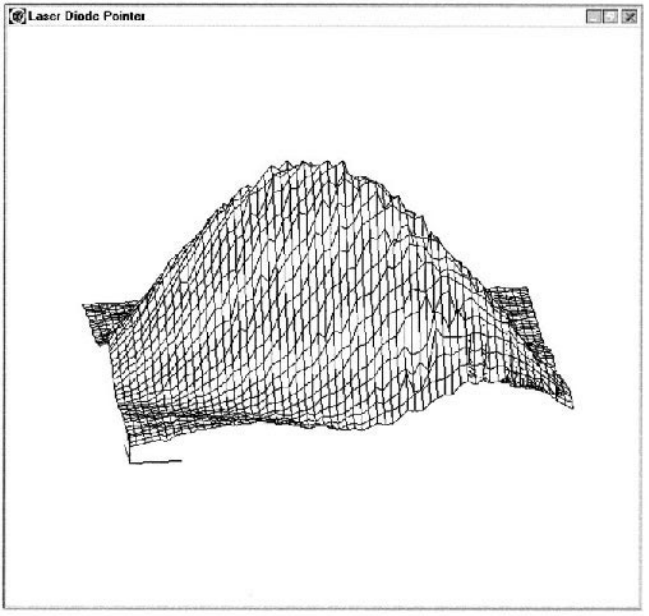


Figure 20 2D gray scale beam profile display showing interference fringes from dust on a neutral density filter used to attenuate the beam.

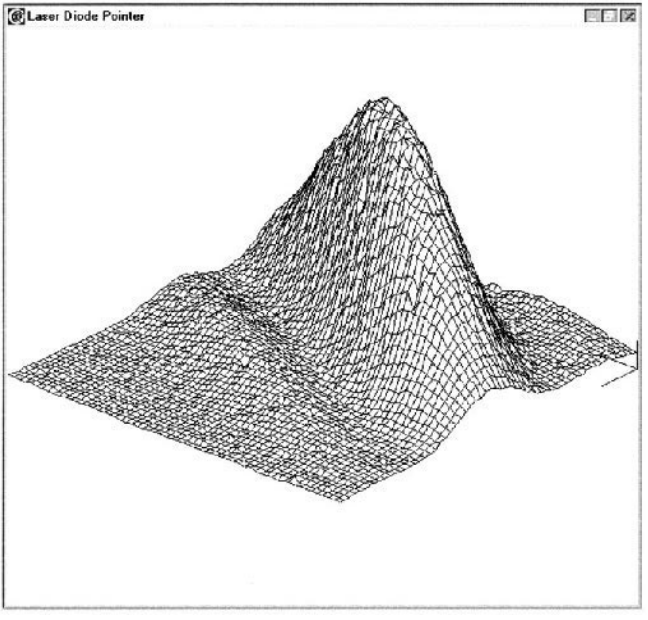
tion that color does not. Figure 20 shows a 2D beam profile display created as a grayscale image. Notice the interference rings that show up dramatically in the shades of gray. If this were in color, one might not even notice the interference rings. These interference rings are reflections from small specks of dust on one of the neutral density filters.

B. Three-dimensional Beam Profile Displays

3D views of the beam profile give a higher level of intuition of what the beam profile really looks like. The user has the option of rotating and tilting the beam, changing the resolution and color, etc. to maximize his ability to obtain intuitive information from the beam display. However, while 2D displays give all the beam profile information simultaneously, 3D displays hide the rear of the beam. Nevertheless, the 3D view is still often useful in gaining greater intuition from the beam profile. Shown in Fig. 21 is a 3D



(a)



(b)

Figure 21 3D Beam profile shown at two rotation angles: (a) 135°; (b) 225°.

beam profile at different angles of rotation, illustrating how the beam looks from different sides.

There are many view processing features now available that assist in enhancing the intuition gained from seeing the beam. These include choices of resolution in the 3D display; that is, the number of lines displayed. Also adjacent pixel summing and convolution are available to reduce the visual signal-to-noise ratio, and thus enable a user to more clearly see the major features of the beam, especially in the presence of noise.

VI. QUANTITATIVE MEASUREMENTS

One of the most important features of modern beam profilers is the ability to make very accurate measurements of the beam characteristics. Two important laser characteristics, the wavelength of the laser, and the temporal pulse width, are unrelated to beam profile measurement; and are easily measured by instruments other than beam profile measuring instruments. Nearly all other qualities of a laser beam are related to the beam profile. One additional measurement that is not measured directly by beam profile instruments is the total power or energy, which also must be measured by a separate instrument. However, the total power or energy can be measured with a power or energy meter at the same time that the sampled beam profile is measured. When this is done then the beam profiling system can be calibrated to the total power or energy measurement, and from then on the beam profiler is able to track the total power or energy, being calibrated to the power meter reading.

Characteristics of a laser that are directly related to beam profile measurements include the pulse-to-pulse relative power or energy, as discussed above, the peak power or energy, the location of the peak, the location of the centroid of the power or energy, and the beam width. The beam width can be measured either on an X/Y-axis, or for an elliptical beam, can be measured along the major and minor axis. Each of these characteristics is discussed below.

A. Relative Beam Power or Energy

Cameras are seldom able to give a direct measurement of the total energy or power in a laser beam. Two reasons for this are as follows. First, the camera follows a long chain of attenuation so that it does not see the total beam directly. Since this attenuation is put in place for the purpose of getting the energy down to the level of the camera, and can be as much as a factor of 10^{11}, it is not practical to calibrate each element of attenuation. Thus the

absolute power fed to the camera is unknown relative to the total power of the beam. Secondly, cameras do not have uniform wavelength absorption. Therefore they would have a different calibration factor for every wavelength of laser that is used. It would be impractical to attempt to calibrate the camera as a function of wavelength.

Nevertheless, as described above, a power meter can be used in the direct beam path to measure the total power while the split off portion is sent to the camera. After correcting for the power lost to the first beam splitter, the total energy or power measured by the meter can then be entered into the software of the beam analysis instruments. From then on the camera can give a readout of the total power or energy. This is especially useful because cameras see the entire two- dimensional beam distribution, and thus give a relative total beam measurement as accurately as the power meter.

B. Peak Power or Energy

Peak power or energy is a relatively easy measurement that is derived from the total power. Since the total power on a camera is a summation of the irradiance on each pixel, it becomes relatively easy to determine what part of this total power is contained within each pixel, and thus the energy on the pixel with the highest power is derived in software. This is a useful measurement in that it tells whether there are hot spots in the beam, and what is the magnitude of these hot spots. This can be particularly useful when the laser power or energy is approaching the damage threshold of optics through which the beam must pass. A hot spot in the beam could cause damage even when power averaged over the area of the beam is well below the damage threshold.

C. Peak Pixel Location

When the software in the beam analyzer finds the magnitude of the pixel with the highest irradiance, it can also provide the location of this pixel. This may be useful to track the stability of the hot spot or peak irradiance, and determine whether or not this highest irradiance is stable or is moving back and forth across the beam. The actual peak irradiance location is seldom useful in telling where the majority of the energy of the beam is located, however.

D. Beam Centroid Location

Quite often, more significant than the peak pixel location is the location of the centroid of the beam. The centroid is defined as the center of mass or first moment of the laser beam, and is described in Eq. 2:

$$X = \int\int xE(x,y,z)\,dx\,dy \Big/ \int\int E(x,y,z)\,dx\,dy \tag{2a}$$

$$Y = \int\int yE(x,y,z)\,dx\,dy \Big/ \int\int E(x,y,z)\,dx\,dy \tag{2b}$$

The centroid of the beam can be more significant than the peak pixel because it is independent of hot spots in the beam. This is where the energy center is located. Usually the pointing stability of a beam is measured by doing statistical analysis on the centroid rather than on the peak pixel. This pointing stability provides significance in showing the stability of the laser beam position.

The significance of the beam centroid can be very important in alignment of laser beams. This is true in optical trains, on research tables, and in industrial laser applications where it is important to know that the beam is positioned correctly in the optics. It is also significant in aligning lenses to laser diodes to collimate the beam. The beam centroid must also be accurately known when aligning beams into fiber optics. Many beam-shaping systems require alignment of the beam, usually the centroid, to the shaping optics.

E. Beam Width

One of the most fundamental laser profile measurements is the beam width. It is a measurement of primary significance because it affects many other beam parameters. For example, the beam width gives the size of the beam at the point where measured. This can be significant in terms of the size of the elements that are in the optical train. Measurement of beam width is also a part of measuring divergence of laser beams, which is significant in predicting what size the beam will be at some other point in the optical train. The beam width is critical for the performance of most non-integrating beam shaping systems. Statistical measurement of the width of the beam is also a significant factor in determining the stability of the laser output. Finally, measurement of the beam width is essential in calculating the M^2 of the laser. This is an important characteristic of laser beams that will be discussed later in this section. Even though fundamental and important,

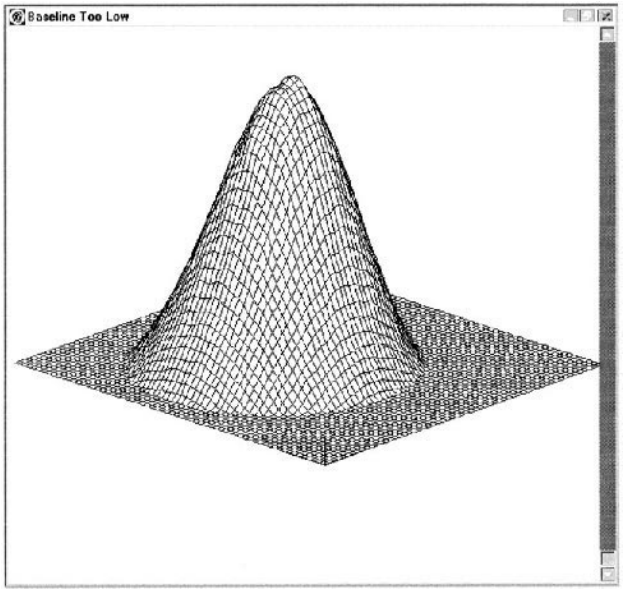
this beam width is sometimes a very difficult measurement to perform accurately.

1. Considerations in Accurate Beam Width Measurement

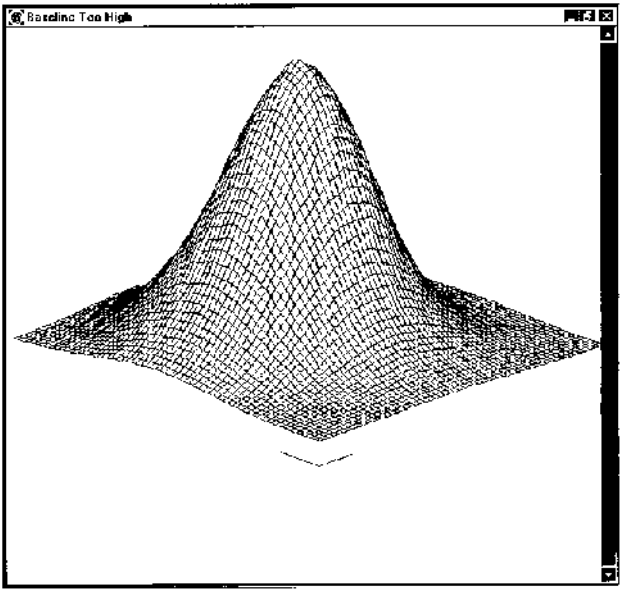
A number of considerations in the characteristics of a camera used for beam profile measurement must be carefully considered and accounted for in accurately measuring beam width. Among these considerations is the signal-to-noise ratio, i.e., the magnitude of the beam relative to the background noise in the camera. The amount of attenuation used for the camera is usually adjusted to enable the peak pixel in the camera to be as near to saturation as possible without overdriving the camera. Also if the beam is of a very small size in a very large field of camera pixels, this may be a very small amount of signal compared to the random noise of all the pixels. Proper treatment of this noise is discussed below.

The camera baseline offset is another factor that must be accurately controlled. Because the energy of a laser does not abruptly go to zero, but trails off to a width roughly four times the standard deviation, or twice the $1/e^2$ width, there is a lot of low-power energy that must be accounted for in accurately measuring the width of the beam (22–24). (The proportion of energy in a Gaussian beam is 68% in $\pm 1\sigma$, 95% in $\pm 2\sigma$, and 99.7% in $\pm 3\sigma$. Nevertheless, experiments performed by the author have shown that as an aperture cuts off the beam at less than $\pm 4\sigma$ the measured beam width begins to decrease.) Correct and incorrect baseline control is illustrated in Figs. 22a, 22b, and 22c. In Fig. 22a the baseline is set too low, and the digitizer cuts off all the energy in the wings of the beam. The beam is seen to rise out of a flat, noiseless baseline. This means that without the wings of the laser beam, a measurement would report a width much too small. In Fig. 22b the baseline offset is too high, as seen by observing the beam baseline relative to the small corner defining mark. In this case the software will interpret the baseline as part of the laser beam. A calculation of beam width will be much too large. In Fig. 22c the baseline is set precisely at zero. Both positive and negative noise components are retained out beyond the wings of the beam where there is no beam energy. The software will interpret the average of the positive and negative signal as nearly zero.

Because the low power energy in the wings of a laser beam can have a significant effect on the width measurement, it becomes necessary to be able to characterize the noise in the wings of the beam. Both the noise components that are above and the noise components below the average noise in the baseline must be considered. The noise below the average baseline will hereafter be called negative noise.

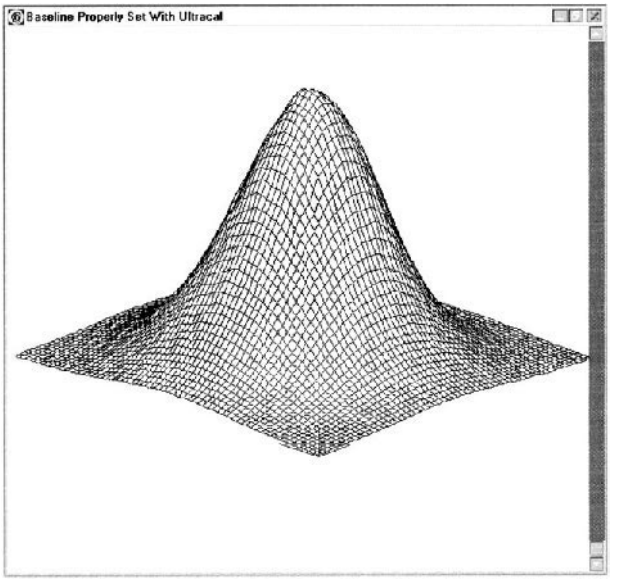


(a)



(b)

Figure 22 Camera baseline set (a) too low, (b) too high, and (c) properly. (Low baseline shows beam rising out of a flat background, which would cause a beam width calculation too small. High baseline would cause a beam width measurement too large.)



(c)

Figure 22 (*Continued*)

Since the size of the beam measurement is affected by the total amount of laser beam energy relative to the noise of the camera, it has been found that software apertures placed around the beam can have a very strong effect in improving the signal-to-noise ratio. For a non-refracted beam, an aperture approximately twice the $1/e^2$ width of the beam can be placed around the beam, and all noise outside the aperture can be set to zero in the calculation. This greatly improves the relative signal-to-noise ratio when small beams are being measured in a large camera field. Finally, the measurement algorithm that is used to measure the beam width can have a notable effect on the accuracy and significance of the measurement.

2. Beam Width Definitions

There are various traditional definitions of beam width, which may or may not contribute to knowing what the beam will do when focused or propagated into space. Some of these include measurements of the width at some percentace of the peak, full width/half maximum, which would be 50% of peak, a percentage of energy, or the $1/e^2$ width. Software equivalent knife-edge measurements are also used as means of determining the beam width. Finally, a more recent definition of beam width is called the second moment (25).

The software equivalent knife-edge measurement and the second moment measurement are becoming the most widely accepted means of measuring laser beams. Both measurements are independent of holes or structure within a beam. When the knife-edge measurement is performed correctly, it does an excellent job of approximating a second moment measurement (26). The knife-edge measurement with a camera is simply a software algorithm simulating the motion of an actual moving knife-edge. One advantage of cameras over actual mechanical scanning knife-edges is that the software can quickly find the major and minor axis of an elliptical beam, and perform the knife-edge measurement along these axes without having to actually reposition the mechanical device.

3. Second Moment Beam Width Measurements

Recent ISO standards (26–29) have defined a second moment beam width, abbreviated $D4\sigma$, which, for many cases, gives the most realistic measure of the actual beam width. The equation for the second moment beam width is given in Eq. (3). Equation (3) is an integral of the irradiance of the beam multiplied by the square of the distance from the centroid of the beam, and then divided by the integrated irradiance of the beam. This equation is called the second moment because of the analogy to the second moment of mechanics, and is abbreviated D4σ because it is the diameter at $\pm 2\sigma$ which is $\pm 1/e^2$ for Gaussian beams. This second moment definition of a beam width enables a user to accurately predict what will happen to the beam as it propagates, what is its real divergence, and the size of the spot when the beam is focused.

$$D4\sigma_x = 4\left(\iint (x - X)^2 E(x, y, z)\, dx\, dy \Big/ \iint E(x, y, z)\, dx\, dy \right)^{1/2} \quad (3a)$$

$$D4\sigma_y = 4\left(\iint (y - Y)^2 E(x, y, z)\, dx\, dy \Big/ \iint (E(x, y, z)\, dx\, dy \right)^{1/2} \quad (3b)$$

where

$(x - X)$ and $(y - Y)$ are the distances to the centroid coordinates X and Y.

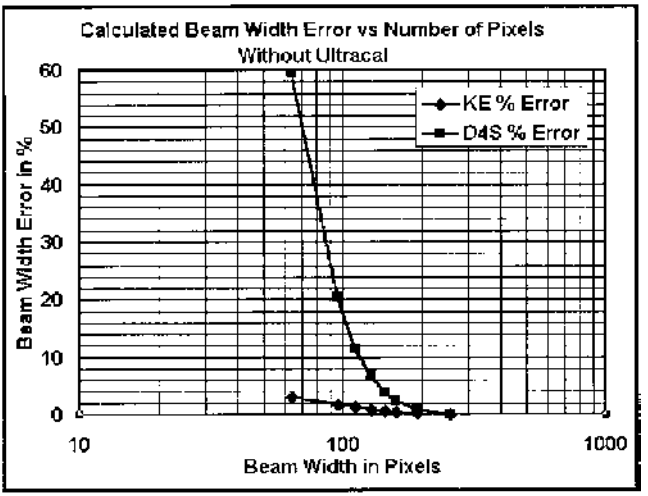
Sometimes there are conditions of laser beams wherein the second moment measurement is not an appropriate measurement to make. This is particularly true when there are optical elements in the beam smaller than twice the $1/e^2$ width that cause diffraction of part of the energy in the beam. This diffraction will put energy further out into the wings of the beam, which when measured by the second moment method, will

cause a measurement of the beam width much larger than is significant for the central portion of the beam. By Eq. (3), the $(x - X)^2$ term overemphasizes small signals far from the centroid. This requires judgment on the part of users as to whether or not measurement of this diffracted energy is significant for their application. If the diffracted energy, which typically diverges more rapidly than the central lobe, is not significant, it is possible to place a physical or software aperture around the main lobe of the beam and make second moment measurements only within this aperture, and disregard the energy in the wings. However, if the application is dependent on the total amount of energy, and it is important to know that part of this energy is diffracted, then one would want to place this aperture such that it includes all the beam energy in making the calculation.

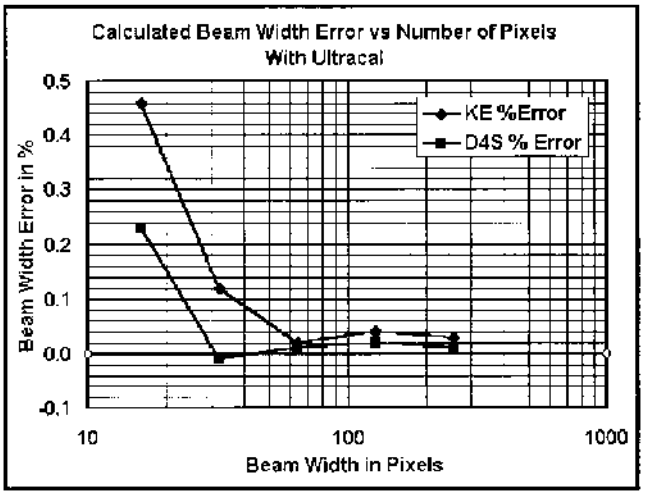
Second moment measurement, however, is very difficult to make with CCD cameras because the high camera noise out in the wings of the beam is multiplied by $(x - X)^2$ producing a large error component. Also any offset or shading of the camera in the wings of the beam causes very large errors because these small energy numbers are multiplied by $(x - X)^2$. For example, Figs. 23a and 23b illustrate the difficulty of making second moment measurements. These figures are from theoretical calculations based on creating a perfect Gaussian beam, adding random noise to the mathematically derived beam, then using beam width measurement algorithms to calculate the beam size. In Fig. 23a it is seen that a knife-edge measurement can measure a beam of 64 pixels in a 512 field with only 3% error. However, using second moment measurement and random camera noise, the beam width error rises to over 60%. For this reason, a few years ago, theoreticians believed that is was not possible to make an accurate second moment beam width measurement with a commercial grade CCD camera. However, as shown in Fig. 23b, using a knife-edge can initially calculate a relatively accurate beam width. Then by placing a 2× software aperture around the beam, the second moment measurement can make very accurate beam width calculations down to a beam containing as few as 13 pixels.

In the following comparisons, Figs. 24 to 26, measurements were made to determine the effect on beam width measurement accuracy of various parameters. Since there is no "traceable standard beam width" the beam was first measured under the most ideal conditions. This includes a large beam of high intensity, and using 2× apertures and negative noise components. Then as measurement conditions are changed the "error" is calculated as the percentage change in measured beam width from the measurement made under the ideal conditions. All measurements in Figs. 24–26 were made on the same beam and in the same time frame.

Figures 24a and 24b illustrate the measured experimental accuracy of making second moment beam width measurements with and without a 2×



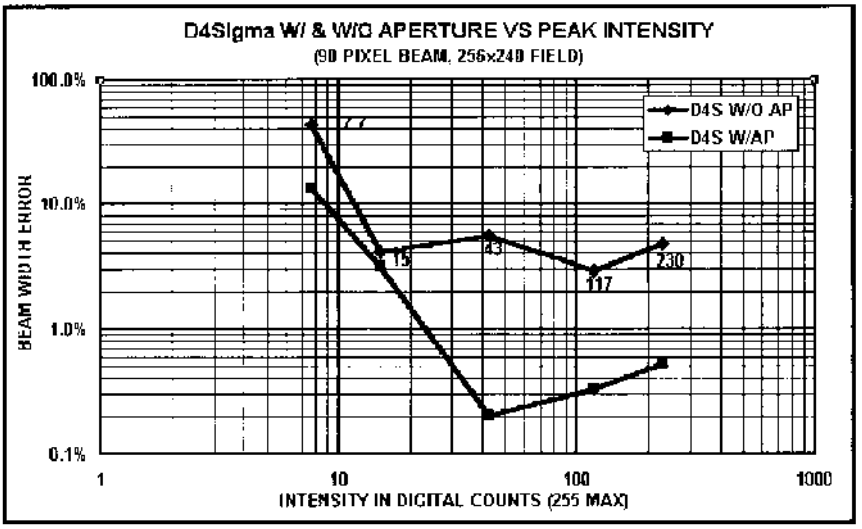
(a)



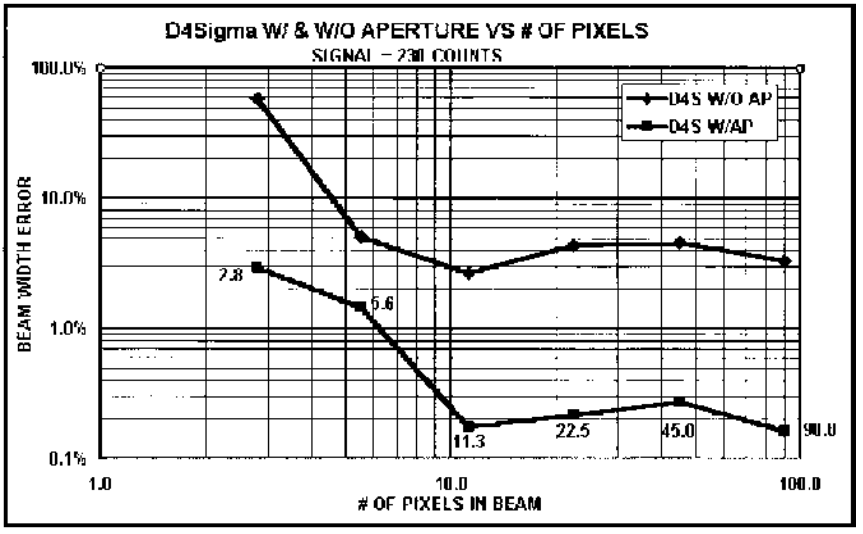
(b)

Figure 23 (a) Simulated beam width error versus number of pixels without a 2× aperture; (b) simulated beam width error versus number of pixels with a 2× aperture.

software aperture. Figure 24a illustrates the accuracy versus the irradiance of the peak pixel on the camera. Notice that with the 2× software aperture around the beam, the irradiance can be reduced to as low as 16 counts out of 256, or roughly 5% of saturation, and the beam width measurement error is still only about 3%. Without an aperture the beam width measurement error is in the 3–5% range, regardless of the irradiance of the beam. In Fig. 24b it is shown that the number of pixels in the beam can be reduced to about

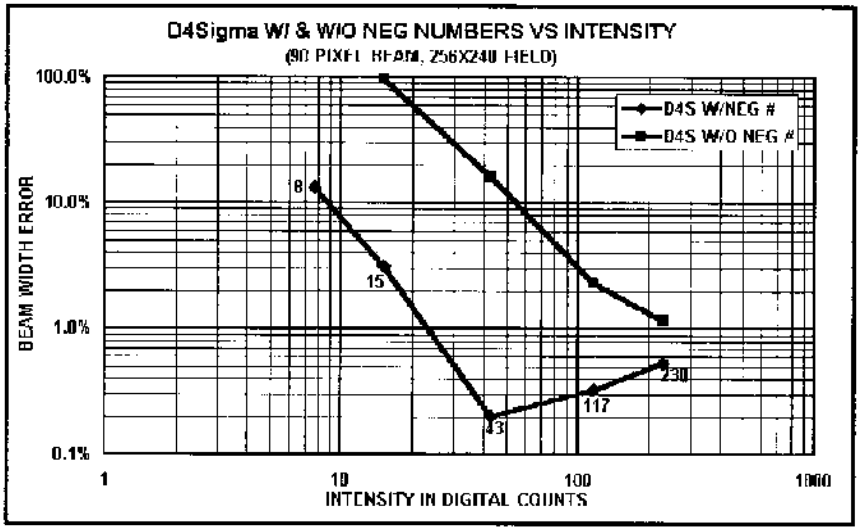


(a)

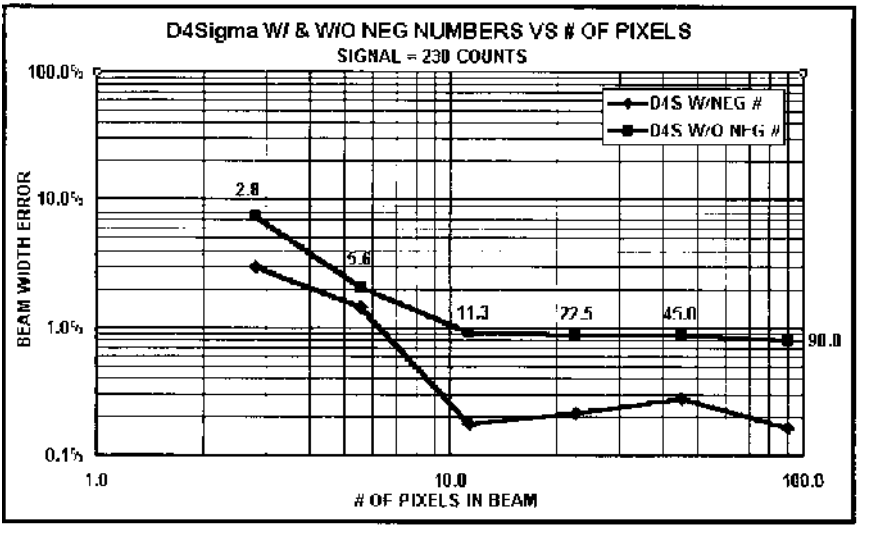


(b)

Figure 24 (a) Measured beam width error versus irradiance with and without a 2× aperture; (b) measured beam width error versus number of pixels with and without a 2× aperture.

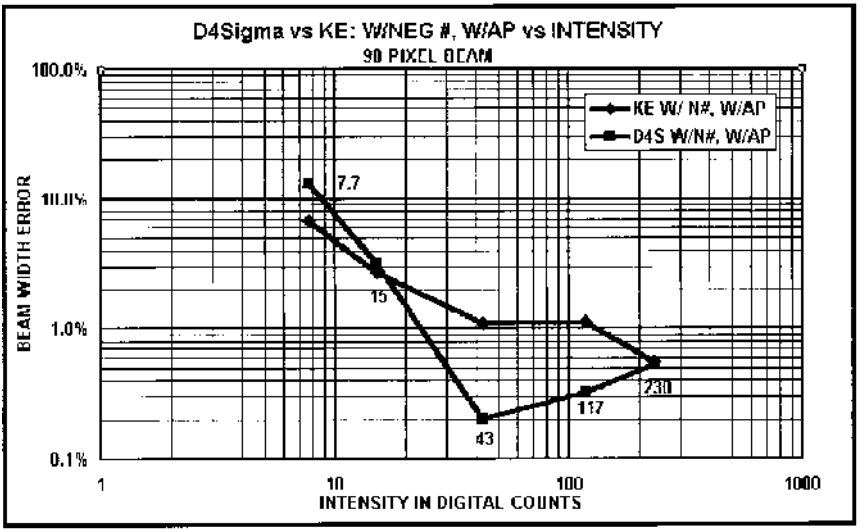


(a)

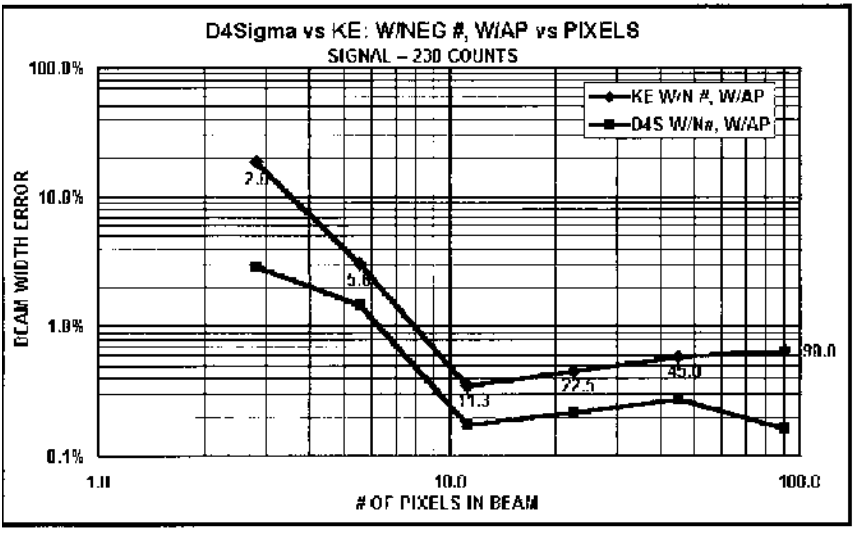


(b)

Figure 25 (a) Measured beam width error versus irradiance with and without negative baseline numbers; (b) measured beam width error versus number of pixels with and without negative baseline numbers.



(a)



(b)

Figure 26 (a) $D4\sigma$ versus KE beam width error versus irradiance (with aperture and negative numbers); (b) $D4\sigma$ versus KE beam width error versus number of pixels (with aperture and negative numbers).

3×3 pixels before the beam width error measurement rises to 3%. Without an aperture in the beam the beam width error is always in the 3–5% range, and at 3 pixels the error rises to over 60%.

Other conditions that are necessary to accurately measure the second moment beam width include accurate baseline control. This is done by having the software perform a multiple frame average of each individual pixel in the camera while the camera is not illuminated. This baseline is then subtracted from the signal when the laser is being measured. This baseline subtraction eliminates not only total offset of the baseline, but also any shading in the camera. (Shading is defined as the offset in the baseline not being uniform across the camera, but varying from one side to the other.)

In addition to accurate baseline control and $2\times$ software apertures mentioned in the previous paragraph, it is also very important to maintain the negative numbers derived from background subtraction as described previously. The following figures illustrate measurements made on an actual laser beam to determine the relative accuracy of making beam width measurements using both second moment and knife-edge under varying conditions. Figures 25a and 25b illustrate the ability of the second moment algorithm to accurately measure beam width with and without negative numbers in the baseline. Notice that in Fig. 25a where the beam is reduced in irradiance, the beam can be as low as 15 counts or 5% of saturation, with only 3% error. However, without negative numbers in the baseline, at 15 counts the beam width error is 100%. Figure 25b illustrates the ability to accurately measure beam width as a function of the number of pixels in the beam with and without negative numbers in the baseline. Having the negative numbers improves the accuracy by about a factor of 5 for larger beams. At 3×3 pixels the accuracy is 3% with negative numbers, and just over 7% without negative numbers.

Figures 26a and 26b illustrate the measurement accuracy of the second moment beam width method compared to the accuracy of the knife-edge algorithm. In Fig. 26a the measurements are compared versus irradiance of the beam. In Fig. 26b they are compared versus the number of pixels in the beam. In both cases an aperture and negative numbers are used. Note that both second moment and knife-edge have approximately the same measurement accuracy in these conditions.

F. Other Important Beam Profile Measurements

1. Beam Ellipticity

With camera-based beam profiling systems it is relatively simple for the software to measure the ellipticity of laser beams. The software typically

finds the major axis of a beam, and then sets the minor axis perpendicular to the major axis. Once the major axis is found, the angle that the major axis deviates, typically from the X-axis, is given, and the ratio of the major to minor axis widths is calculated. This is an extremely useful measurement in laser beam alignment. It is particularly useful in aligning lenses to laser diodes, which are highly elliptical. Typically a special lens is used with diodes to circularize the beam. The alignment of this lens to the diode is extremely critical. With mechanical scanning systems it is very cumbersome to find the major and minor axes. Whereas with a camera-based system the entire beam profile is obtained in every frame of the camera, so the ellipticity can be found instantaneously. This makes it extremely rapid to do beam and component alignment in real time.

Another important reason for knowing the ellipticity of the laser beam is in industrial applications. Typically if the beam becomes elliptical, a laser used for cutting irregular shapes will have a different cut width in one axis than in the other. By measuring the ellipticity and correcting it when it goes beyond acceptable limits, industrial users can eliminate creating scrap materials.

2. Gaussian Fit

In many cases the desired beam irradiance profile is a Gaussian beam with its irradiance at any point in the X–Y plane corresponding to Eq. 4. There are a number of ways to perform a fit of the real beam to the Gaussian equation. One of these is to minimize the deviation, which is defined in Eq. 5. This fit can either be along an X/Y-axis, a major/minor axis, or can be performed over the entire laser beam. Being performed over the entire beam is useful in that it means that any energy off axis contributes to determining how well the beam fits a perfect Gaussian. In addition to these equations, the actual data of beam profile irradiance can be exported to a spreadsheet and user's can perform the calculations according to their own method.

Gaussian equation:

$$J = J_0 e^{-2[(x-\bar{x}/w_x)^2+(y-\bar{y}/w_y)^2]} + A \tag{4}$$

where:

J = amplitude at the point (x, y)

J_0 = amplitude at the Gaussian center

$x = x$ location of pixel

$\bar{x} = x$ location of the Gaussian center

$w_x =$ horizontal radius at $1/e^2$ of energy

$y = y$ location of pixel

$\bar{y} = y$ location of the Gaussian center

$w_y =$ vertical radius at $1/e^2$ of energy

$A =$ offset

Minimization of the deviation can be performed by varying the parameters of Eq. 4 using the spreadsheet solve feature. A is an offset term which is set to zero, i.e. disregarded in beam analyzers, because as stated above, the background is carefully set to zero. The definition of the deviation is

$$\sigma = \sqrt{\frac{\sum (Z - s)^2}{n - 2}} \tag{5}$$

where:

$\sigma =$ standard deviation

$Z =$ pixel irradiance

$s =$ Gaussian surface irradiance

$n =$ number of pixels

Gaussian fit as a measure of the quality of a laser beam is becoming less important. It has been shown that a multimode beam with the right combination of modes can look Gaussian (26), and can very closely fit to a Gaussian curve. Nevertheless, the beam has many modes, and is far from true TEM_{00} mode. A multimode will not follow the propagation laws of a perfect Gaussian beam, and a user can be misled by the Gaussian fit. Instead, the parameter M^2 has become more popular as representing the reality of how close the beam is to a true TEM_{00} Gaussian. M^2 will be discussed in more detail below.

3. Top Hat Measurement

Many real beams are intended to be flat top. Some of the beam shaping chapters earlier in this book discuss how to obtain uniform flat tops from Gaussian and other input beams. A flat top beam is useful in many applications where the irradiance should be uniform over a given cross section.

Applications include medical processes such as wine spot removal and photorefractive keratotomy, wherein a uniform portion of the cornea of the eye is removed. Industrial applications in which a flat top is useful include cleaning of surfaces and marking.

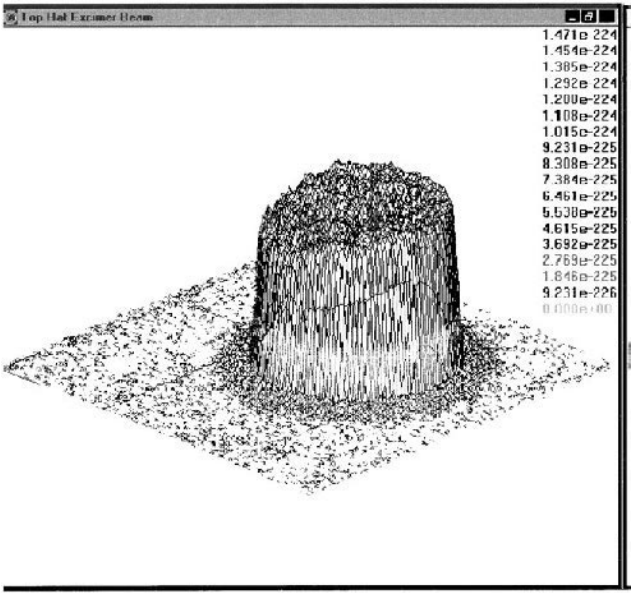
Camera-based systems enable easy and accurate measurements of flat top beams. The software is programmed to give a readout of the average irradiance, or the mean across the flat top, the standard deviation of the variations from the mean, and the standard deviation divided by the mean, which gives a percentage flatness. Also the minimum and maximum can be provided, which give additional information about the relative flatness of the beam. The flat top factor (30) is a way to give a quantitative and intuitive measure of how flat a top hat beam is. (The equations are given in Ref. 30.) A typically square beam would have a top hat factor of 1. A Gaussian beam has a top hat factor of 0.5. Therefore, most beams will fall somewhere between 0.5 and 1. In addition to measuring the flatness of the top hat, the software can also calculate the top hat area and the size or width of the top hat beam. Figure 27a shows a typical top hat beam, and Fig. 27b shows the typical calculations.

4. Divergence Measurement

Divergence is an important characteristic of laser beams. It gives the angle at which the beam is diverging from a collimated parallel beam. It is important because the lower the divergence, the longer the beam will remain at a given diameter. Typically when low divergence is necessary, a beam is often expanded to a large width, and then the divergence of this large width beam is smaller. Nevertheless, beam divergence by itself does not provide the true characteristics of a beam, since as just mentioned, simply expanding the beam to a larger waist can change it. This will be explored in more detail in the M^2 section to follow.

5. Statistical Measurement

Statistics on all measurements can provide information on long term stability of the laser beam. A typical example of statistical measurement is shown in Fig. 28. This figure illustrates a number of the basic measurements possible from the software, along with the statistics provided by sampling twenty calculations to determine the beam stability. Statistics can be performed in a large variety of ways. For example, software can be arranged so that only one measurement is made out of every few hundred frames, then statistics are calculated on thousands of such frames. This enables one to track the stability of a laser with respect to time, temperature fluctuations, or other characteristics of interest. Statistics typically provide the mean or



(a)

Frame 9

	Current	Units
—Quantitative—90/10 Knife Edge—		
Total	2..02.6e-.225	mw
% Above Clip	84.16	%
Peak	8.481e-225	mw/cm²
Min	-5.769e-227	mw/cm²
Peak Loc X	8.300e+03	um
Peak Loc Y	6.500e+03	um
Centroid X	7.035e+03	um
Centroid Y	6.765e+03	um
Width X	6.371e+03	um
Width Y	6.328e+03	um
Diameter	6.349e+03	um
—Top Hat—Data—		
Mean	7.761e-225	mw/cm²
StdDev	3.903e-226	mw/cm²
SD/Mean	5.03	%
Min	6.808e-225	mw/cm²
Max	8.481e-225	mw/cm²
Factor	0.823	
Effective Area	2.197e+07	um²
Effective Diam	5.289e+03	um
—Divergence—	—Focal Length—	
Divergence X	6.371e+00	mrad
Divergence Y	6.328e+00	mrad

(b)

Figure 27 (a) Typical top hat beam; (b) top hat calculations.

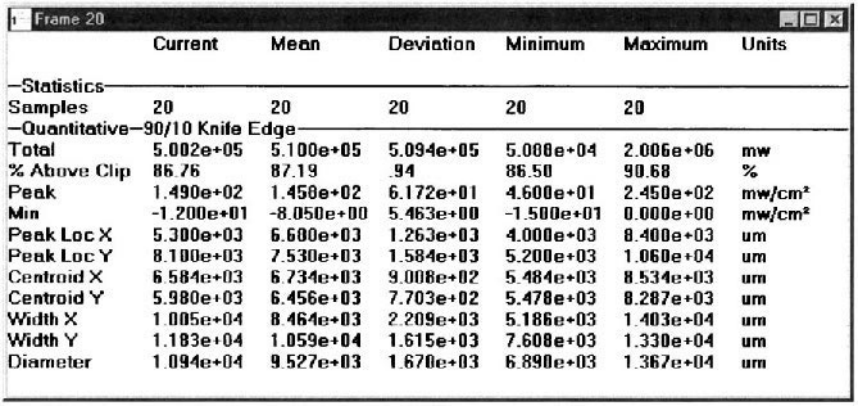


Figure 28 Statistical measurement of the basic laser beam parameters.

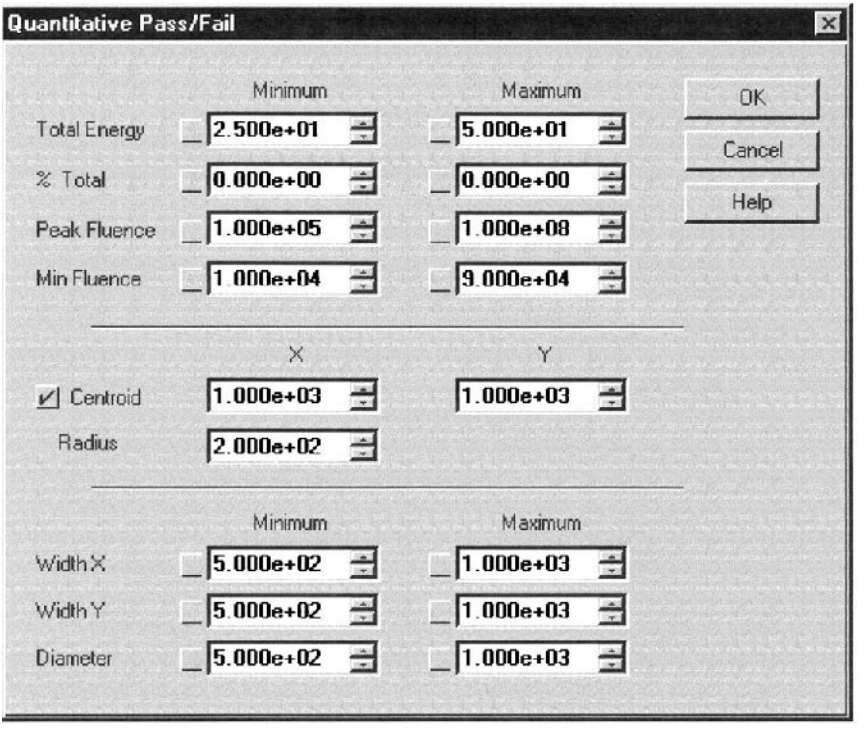


Figure 29 Pass/fail dialog box.

average measurement of a parameter, the standard deviation, and the minimum and maximum to which that characteristic has drifted.

6. Pass/Fail Measurements

Figure 29 shows a typical dialog for a pass/fail measurement. Essentially all of the quantitative measurements being made on the laser beam can have pass/fail limits set in one of these dialog boxes. Thus, for example, if the centroid location is critical in a manufacturing or other environment, a maximum radius from a given position can be set. The software can then be programmed to provide an alarm if the parameter of interest drifts outside the limits. This feature can be used in many environments, including industrial, instrument design, laser stability and design, and others.

G. M^2 Measurements

M^2, or the factor $k = 1/M^2$ in Europe, has become increasingly important in recent years in describing the quality of a laser beam (31–43). In many applications, especially those in which a Gaussian beam is the desired profile, M^2 is the most important characteristic describing the quality of the beam. Figure 30 illustrates the essential features of the concept of M^2 as

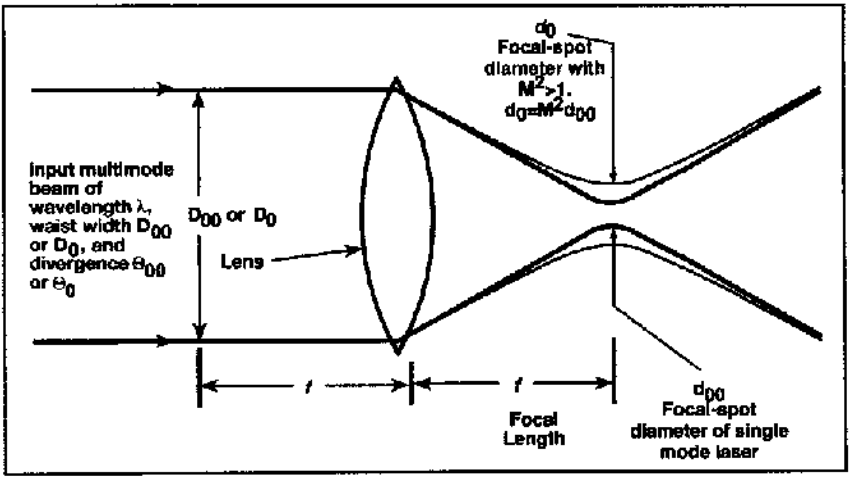


Figure 30 Curve showing M^2 (characteristics and equations relating M^2 to the beam focused spot size.)

defined by Eqs. 6a and 6b. As shown in Fig. 30, if a given input beam of width D_{in} is focused by a lens, the focused spot size and divergence can be readily predicted. If the input beam is a pure TEM_{00}, the spot size equals a minimum defined by Eq. 6a and d_{00} in Fig. 30. However, if the input beam D_{in} is composed of modes other than pure TEM_{00}, the beam will focus to a larger spot size, namely M^2 times larger than the minimum, as shown mathematically in Eq. 6b, and d_0 in Fig. 30. In addition to defining the minimum spot size, M^2 also predicts the divergence of the beam after the focused spot. Specifically, the real beam will diverge M^2 times faster than an equivalent TEM_{00} beam of the same width. Figure 30 illustrates what happens to a beam after going through a focusing lens, but the same principles apply if no lens is involved. That is, the beam will diverge more rapidly by a factor of M^2 than if it were true TEM_{00}. The ISO definition for the quality of a laser beam uses M^2 as the fundamental quality parameter.

$$d_{00} = 4\lambda f/\pi D_{in} \tag{6a}$$

$$d_0 = M^4 4\lambda f/\pi D_{in} \tag{6b}$$

where

λ = the wavelength

f = the focal length of the lens

D_{in} = the width of the input beam.

In measuring and depicting M^2, it is essential that the correct beam width be defined. The ISO standard, and beam propagation theory indicate that the second moment is the most relevant beam width measurement in defining M^2. Only the second moment measurement follows the beam propagation laws so that the future beam size will be predicted by Eqs. (6a) and (6b). Beam width measured by other methods may or may not give the expected width in different parts of the beam path.

M^2 is not a simple measurement to make. It cannot be found by measuring the beam at any single point. Instead a multiple set of measurements must be made as shown in Fig. 31 wherein an artificial waist is generated by passing the laser beam through a lens with known focal length. One essential measurement is to measure the beam width exactly at the focal length of the lens. This gives one measurement of the divergence of the beam. Other measurements are made near the focal length of a lens to find the width of the beam and the position at the smallest point. In addition, measurements are made beyond the Rayleigh range of the beam waist to confirm the divergence measurement. With these multiple measurements

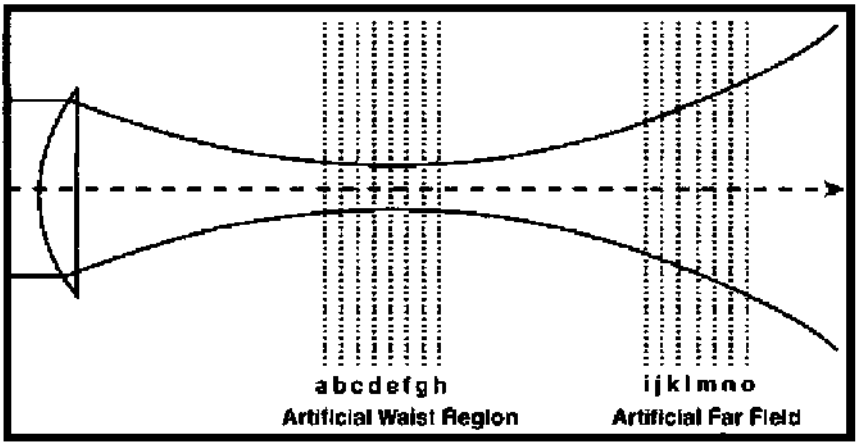


Figure 31 Multiple measurements made to measure M^2.

one can then calculate the divergence and minimum spot size, and then going backwards through Eq. (6b), one can find the M^2 of the input beam.

The measurement shown in Fig. 31 can be made in a number of ways. In one commercial instrument, shown in Fig. 12, a detector is placed behind a rotating drum with knife edges, and then the lens is moved in the beam to effectively enable the measurement of the multiple spots without having to move the detector. This instrument works extremely well as long as the motion of the lens is in a relatively collimated part of the laser beam. However, if the beam is either diverging or converging in the region where the lens is moving, the resultant M^2 measurement can be misleading.

The ISO method (28) for measuring M^2 is to have the lens in a fixed position, and then make multiple detector position measurements as shown in Fig. 31. This can be done by placing a lens on a rail and then moving the camera along the rail through the waist and through the far field region. There are commercial instruments that perform this measurement automatically without having to manually position the camera along the rail. One of these is shown in Fig. 32, wherein the lens and the camera are fixed, but folding mirrors are mounted on a translation table, and moved back and forth to provide the changing path length of the beam. A typical readout of an M^2 measurement is shown in Fig. 33. In this case a collimated laser diode was measured, which gave a much greater divergence in the X-axis than in the Y-axis. The steep V curve displayed is the X-axis of the beam coming to a focus following the lens. The more gradual curve is the focus of the less divergent Y-axis. Notice that while for most of the range the X-axis has a wider beam width, at focus the X axis focuses smaller than the Y-axis. Also,

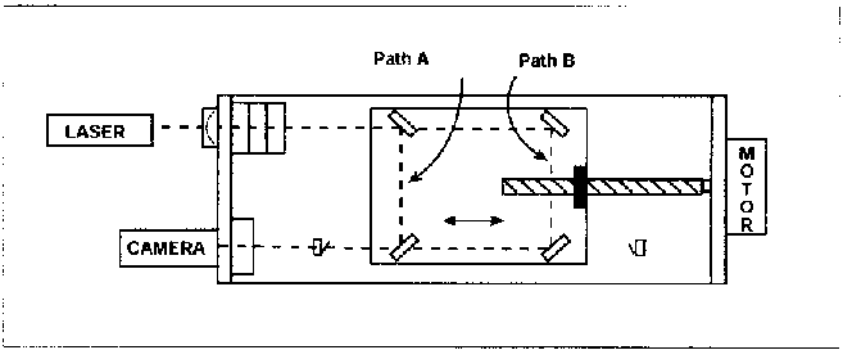


Figure 32 Instrument with fixed position lens for measuring M^2.

the X-axis M^2 was 1.46, whereas the Y-axis M^2 was only 1.10. The M^2 reported in the numbers section is calculated from the measurements of the beam width at the focal length, the minimum width, and the divergence in the far field according to the equations in the ISO standard.

One of the difficulties of accurately measuring M^2 is accurately measuring the beam width. This is one of the reasons that so much effort has been made to define the second moment beam width, and create algorithms

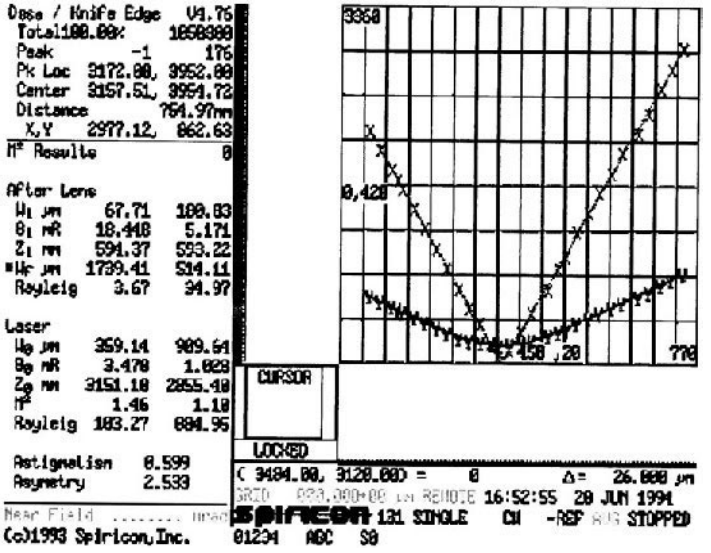


Figure 33 M^2 measurement display and calculation readout.

to accurately make this measurement. Another difficulty in measuring this beam width is that the irradiance at the beam focus is much greater than it is far from the Rayleigh length. This necessitates that the measurement instrument operates over a wide signal dynamic range. Multiple neutral density filters are typically used to enable this measurement. An alternative exists with cameras or detectors that have extremely wide dynamic range, typically 12 bits, so that sufficient signal-to-noise ratio is obtained when the irradiance is low, and still not saturate the detector near the focused waist.

There are some cases when M^2 is not a significant measure of the quality of a laser beam. For example, flat top beams for surface processing typically have a very large M^2, and M^2 is not at all relevant to the quality of the beam. Nevertheless, for many applications in nonlinear optics, industrial laser processing, and many others, the smallest possible beam with the M^2 closest to 1 is the ideal. Some flat top beam mappers are designed for an input Gaussian beam and then the M^2 of the input beam should be very close to 1, and the beam widths should closely match the design width. This is treated in more detail in Chapter 3.

H. Signal Processing

Careful control of the camera baseline and proper treatment of both positive and negative going noise enable signal processing that would not otherwise be possible. Figure 34 shows a HeNe laser beam at near saturation of a CCD camera. This beam was then blocked, and signal summing of 256 frames was performed to determine the noise distribution under summing conditions. This noise is shown in three dimension in Fig. 35a. The darker components of noise at the bottom of the distribution are the negative-going components. With accurate baseline control and treatment of negative noise components, Fig. 35b shows that the distribution of the noise is roughly Gaussian, and is centered at zero. This is what would be hoped for from summing many frames of random noise.

The laser beam of Fig. 34 was then passed through an ND2 filter, which attenuated it by a factor of roughly 100. At this point the laser beam was completely buried in the random noise for each single frame. Again 256 frames of signal were summed, and the signal rose out of the noise as shown in Fig. 36a. In this case the signal sums as the number of frames, whereas the noise sums roughly as the square root of the number of frames. Thus the signal-to-noise ratio is improved by approximately the square root of the number of frames summed. Note that this is possible only when negative noise components are used. Otherwise if negative components are clipped at zero, the noise will sum to a positive DC offset. Figure 36b shows the beam profile of 36a when adjacent pixels in a 4×4

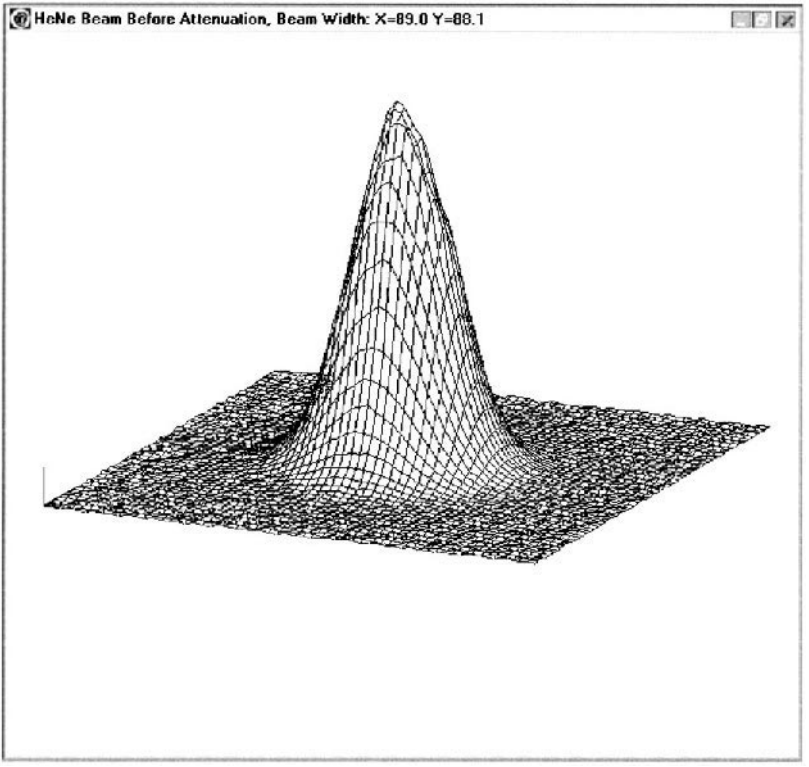
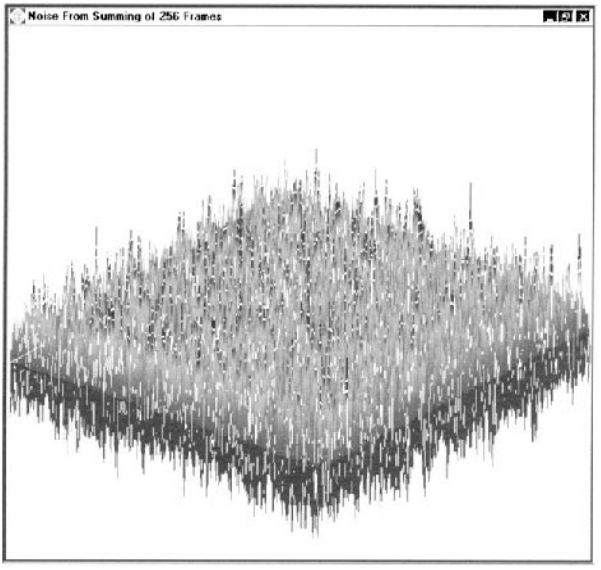


Figure 34 HeNe laser beam used in signal processing experiment.

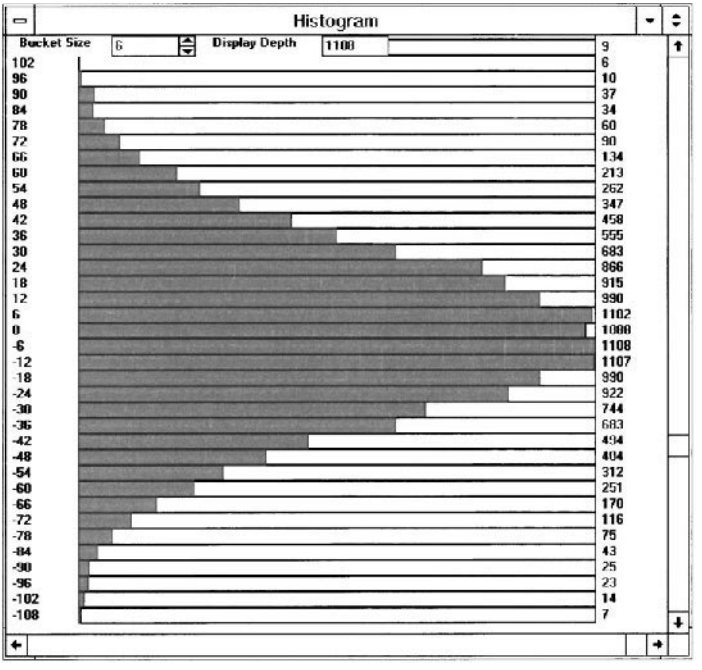
matrix are summed together. Notice a tremendous noise cancellation leaving a much cleaner view of the beam profile. This results from the summing of adjacent positive and negative noise components. Figure 36c shows a similar way of providing a clearer beam profile picture by using convolution to average out the noise in the background. In all three cases of Fig. 36 the beam width measurement, from the measurement of the beam in Fig. 34, was in error by only about 5–7%. This is quite phenomenal for a beam that started out buried in the noise.

I. Wavefront Phase

A more advanced measurement on laser beam profile is the wavefront phase of the laser beam. The profile simply measures the irradiance, but does not predict what the irradiance will be at any point further along the propaga-

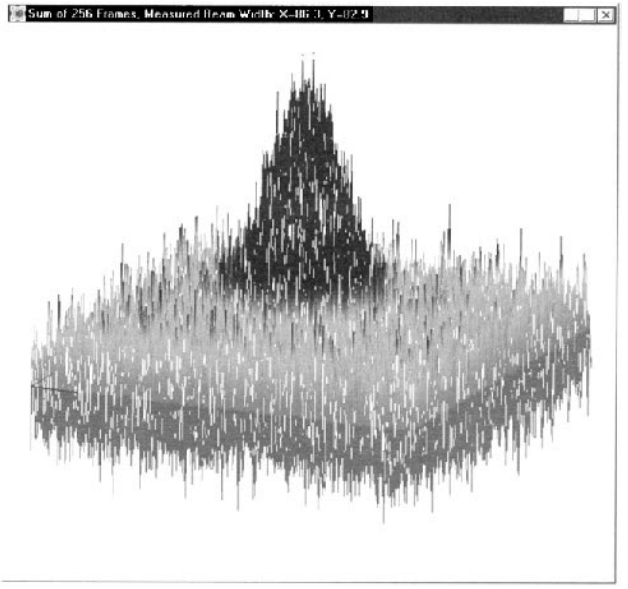


(a)

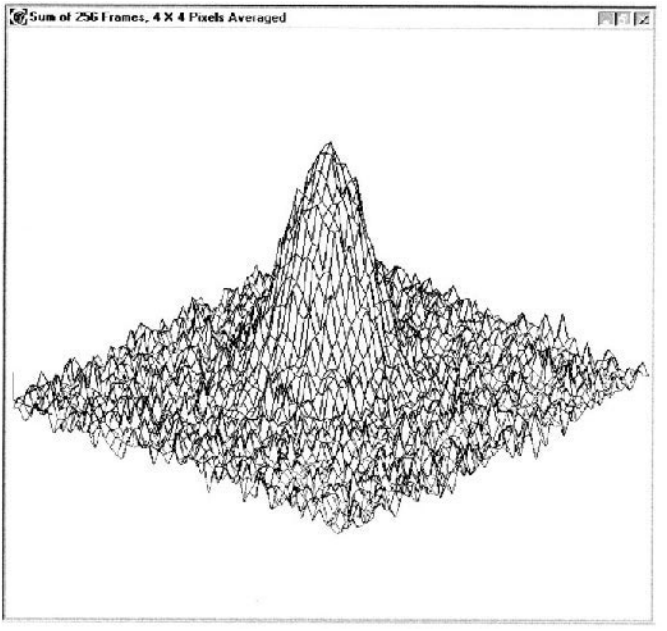


(b)

Figure 35 (a) CCD camera noise after sum of 256 frames; (b) distribution of noise shown in Fig. 35a.

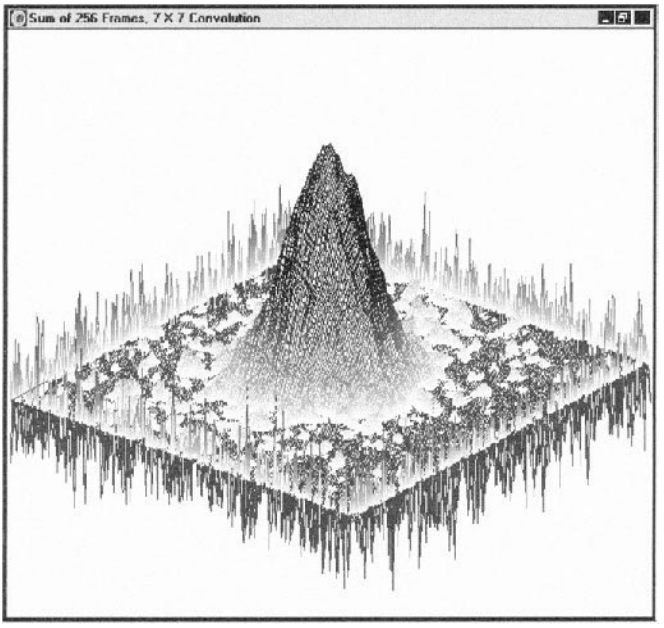


(a)



(b)

Figure 36 (a) Beam of Fig. 34 after attenuation of about 100 and summed for 256 frames; (b) beam of Fig. 36a with summing of pixels in a 4×4 matrix; (c) beam of Fig. 36a with convolution over a 7×7 array.



(c)

Figure 36 (*Continued*)

tion path. A measurement of M^2 tells how much more rapidly a beam will diverge, but does not give any information about the manner in which this divergence will occur. A measurement of wavefront phase gives the details of the beam distortion that are reported as a simple number in M^2. However, wavefront phase is a more complicated measurement to make, as well as to make use of. It is likely that as users become more sophisticated, wavefront phase will become an increasingly important measurement related to the beam profile. Currently there are two methods of measuring the wavefront phase in which commercial instruments are available. One is to use an interferometer, and the second is to use a Hartman array of individual detectors. For some beam shaping problems, knowledge of the wavefront phase is important. (This is discussed in Chapter 3 with respect to collimation and input beam requirements.)

VII. SUMMARY

Beam shaping generally requires beam profile measurement. This is required on the input beam to make sure that it has the proper characteristics. It

is also required on the output beam to make sure the beam shaping mechanism is operating properly. Mechanical scanning instruments can provide single axis profiles, which are sufficient in many cases.

Electronic measurements of laser beams using CCD and other solid-state cameras yield very detailed information on both the input beam and the output beam. Using such beam profilers, scientists and users in beam shaping and many fields of lasers are able to greatly enhance the operation of their instruments. Giving an accurate view of the beam profile, and making precise measurements of beam parameters, such as beam width and other characteristics, provides the ability to properly condition the input beam and measure the shaped output beam.

REFERENCES

1. J Darchuk. Beam profilers beat laser-tuning process. Laser Focus World 205–212, May 1991.
2. G Forrest. Measure for measure (letters). Laser Focus World 55, September 1994.
3. L Langhans. Measure for measure (letters). Laser Focus World 55, September 1994.
4. CB Roundy. A beam profiler that stands alone. Lasers And Optronics 81, June 1990.
5. CB Roundy. The importance of beam profile. Physics World 65–66, July 1990.
6. CB Roundy. Instrumentation for laser beam profile measurement. Industrial Laser Review 5–9, March 1994.
7. CB Roundy. So, who needs beam diagnostics? Lasers & Optronics 19–22, April 1994.
8. CB Roundy. Seeing is believing with visual laser-beam diagnostics. Laser Focus World 117–119, July 1994.
9. CB Roundy. Measure for measure (letters). Laser Focus World 55, September 1994.
10. CB Roundy. Practical applications of laser beam profiling. Lasers & Optronics 21, April 1994.
11. CB Roundy. Electronic beam diagnostics evaluate laser performance. Laser Focus World 119–125, May 1996.
12. CB Roundy. PC-based laser analyzers: New uses require improved devices. Photonics Spectra 97–98, January 1997.
13. MW Sasnett. Propagation of multimode laser beams: the M^2 factor. In: The Physics and Technology of Laser Resonators. DR Hall, PE Jackson, eds. New York: Adam Hilger, 1989, 132–142.
14. AF Siegman. Lasers. Mill Valley, CA: University Science Books, 1986, Chapter 7, p. 697.

15. YV Carts. Excimer-laser work spurs UV beam-profiler development. Laser Focus World 21:24–30, August 1989.
16. CB Roundy. Laser-assisted radialeratotomy. Photonics Spectra 122, October 1994.
17. CB Roundy. Applying beam profiling to industrial lasers. Lasers & Optronics, supplement to Metalworking Digest 5, August 1996.
18. MW Sasnett. Beam geometry data helps maintain and improve laser processes. (Part 1). Industrial Laser Review, 9–13, August 1993.
19. MW Sasnett. Beam geometry data helps maintain and improve laser processes (Part 2). Industrial Laser Review 15–16, May 1994.
20. MW Sasnett. Characterization of laser beam propagation. Coherent Mode-Master Technical Notes, 1990.
21. CB Roundy. Pyroelectric arrays make beam imaging easy. Lasers And Applications 55–60, January 1982.
22. CB Roundy. Digital imaging produces fast and accurate beam diagnostics. Laser Focus World 117, October 1993.
23. CB Roundy, GE Slobodzian, K Jensen, D Ririe. Digital signal processing of CCD camera signals for laser beam diagnostics applications. Electro Optics 11, November 1993.
24. CB Roundy. 12-bit accuracy with an 8-bit digitizer. NASA Tech Briefs 55H, December 1996.
25. RD Jones, TR Scott. Error propagation in laser beam spatial parameters. Optical and Quantum Electronics 26, 25–34, 1994.
26. AE Siegman, MW Sasnett, TF Johnston, Jr. Choice of clip level for beam width measurements using knife-edge techniques. IEEE Journal of Quantum Electronics 27 (4):1098–1104, 1991.
27. M Fleischer. Laser beam width, divergence, and propagation factor: Status and experience with the draft standard. SPIE pp. 2–11, 1991.
28. International Organization for Standardization. Test methods for laser beam parameters: Beam widths, divergence angle and beam propagation factor. Document ISO/11146, 1993.
29. M Sasnett, T Johnston, T Siegman, J Fleischer, D Wright, L Austin, D Whitehouse. Toward an ISO beam geometry standard. Laser Focus World 53, September 1994.
30. G Klauminzer, C Abele. Excimer lasers need specifications for beam uniformity. Laser Focus World, pp. 153–158, May 1991.
31. PA Belanger. Beam propagation and the ABCD ray matrices. Optics Letters 16:196–198. 1991.
32. R Borghi, M Santarsiero. Modal decomposition of partially coherent flat-topped beams produced by multimode lasers. Optics Letters. 23:313–315, 1998.
33. PB Chapple. Beam waist and M^2 measurement using a finite slit. Optical Engineering 33:2461–2466, 1994.
34. RM Herman, TA Wiggins. Rayleigh range and the M^2 factor for Bessel–Gauss beams. Applied Optics 37:3398–3400, 1998.

35. TF Johnston, Jr. M-squared concept characterizes beam quality. Laser Focus World 173–183, 1990.
36. TF Johnston, Jr. Beam propagation (M^2) measurement made as easy as it gets; the four-cuts method. Applied Optics 37:4840-4850, 1998.
37. GN Lawrence. Proposed international standard for laser-beam quality falls short. Laser Focus World 109–114, 1994.
38. M Sasnett, TF Johnston, Jr. Beam characterization and measurement of propagation attributes. SPIE pp. 21–32, 1991.
39. AE Siegman. New developments in laser resonators. SPIE 122:2–14. 1990.
40. AE Siegman. Conference on Laser Resonators. SPIE/OE LASE '90, Los Angeles, CA: January, 1990.
41. AE Siegman. Conference on Lasers and Electro-Optics. CLEO/IQEC Anaheim, May 1990.
42. AE Siegman. Output beam propagation and beam quality from a multimode stable-cavity laser. IEEE Journal of Quantum Electricity 29: 1212–1217, 1993.
43. W Woodward. A new standard for beam quality analysis. Photonics Spectra 139–142, 1990.