The Infrared &
Electro-Optical
Systems Handbook

ELECTRO-OPTICS

VOLUME 8

Emerging Systems and Technologies
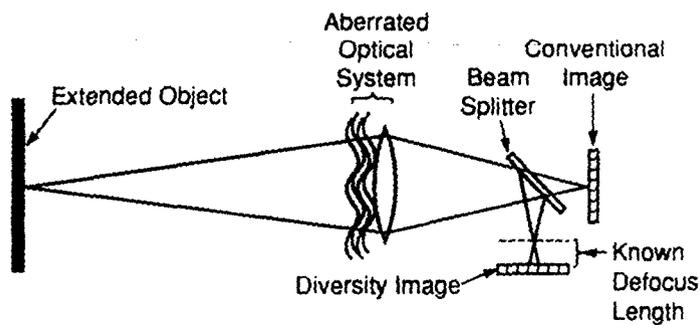
Stanley R. Robinson, *Editor*

Extended Object · Aberrated Optical System · Beam Splitter · Conventional Image · Diversity Image · Known Defocus Length

Emerging Systems and Technologies

VOLUME

8

The Infrared and Electro-Optical Systems Handbook

The Infrared and Electro-Optical Systems Handbook

Joseph S. Accetta, David L. Shumaker, *Executive Editors*

Emerging Systems and Technologies

Stanley R. Robinson, *Editor*
Environmental Research Institute of Michigan

1999060404 042

V O L U M E

8

The Infrared and Electro-Optical Systems Handbook

Joseph S. Accetta, David L. Shumaker, *Executive Editors*
Environmental Research Institute of Michigan

Preface

The Infrared and Electro-Optical Systems Handbook is a joint product of the Infrared Information Analysis Center (IRIA) and the International Society for Optical Engineering (SPIE). Sponsored by the Defense Technical Information Center (DTIC), this work is an outgrowth of its predecessor, *The Infrared Handbook,* published in 1978. The circulation of nearly 20,000 copies is adequate testimony to its wide acceptance in the electro-optics and infrared communities. *The Infrared Handbook* was itself preceded by *The Handbook of Military Infrared Technology.* Since its original inception, new topics and technologies have emerged for which little or no reference material exists. This work is intended to update and complement the current *Infrared Handbook* by revision, addition of new materials, and reformatting to increase its utility. Of necessity, some material from the current book was reproduced as is, having been adjudged as being current and adequate. The 45 chapters represent most subject areas of current activity in the military, aerospace, and civilian communities and contain material that has rarely appeared so extensively in the open literature.

Because the contents are in part derivatives of advanced military technology, it seemed reasonable to categorize those chapters dealing with systems in analogy to the specialty groups comprising the annual Infrared Information Symposia (IRIS), a Department of Defense (DoD) sponsored forum administered by the Infrared Information Analysis Center of the Environmental Research Institute of Michigan (ERIM); thus, the presence of chapters on active, passive, and countermeasure systems.

There appears to be no general agreement on what format constitutes a "handbook." The term has been applied to a number of reference works with markedly different presentation styles ranging from data compendiums to tutorials. In the process of organizing this book, we were obliged to embrace a style of our choosing that best seemed to satisfy the objectives of the book: to provide derivational material data, descriptions, equations, procedures, and examples that will enable an investigator with a basic engineering and science education, but not necessarily an extensive background in the specific technology, to solve the types of problems he or she will encounter in design and analysis of electro-optical systems. Usability was the prime consideration. In addition, we wanted each chapter to be largely self-contained to avoid time-consuming and tedious referrals to other chapters. Although best addressed by example, the essence of our handbook style embodies four essential ingredients: a brief but well-referenced tutorial, a practical formulary, pertinent data, and, finally, example problems illustrating the use of the formulary and data.

The final product represents varying degrees of success in achieving this structure, with some chapters being quite successful in meeting our objectives and others following a somewhat different organization. Suffice it to say that the practical exigencies of organizing and producing a compendium of this magnitude necessitated some compromises and latitude. Its ultimate success will be judged by the community that it serves. Although largely oriented toward system applications, a good measure of this book concentrates on topics endemic and fundamental to systems performance. It is organized into eight volumes:

Volume 1, edited by George Zissis of ERIM, treats sources of radiation, including both artificial and natural sources, the latter of which in most military applications is generally regarded as background radiation.

Volume 2, edited by Fred Smith of OptiMetrics, Inc., treats the propagation of radiation. It features significant amounts of new material and data on absorption, scattering, and turbulence, including nonlinear propagation relevant to high-energy laser systems and propagation through aerodynamically induced flow relevant to systems mounted on high-performance aircraft.

Volume 3, edited by William Rogatto of Santa Barbara Research Center, treats traditional system components and devices and includes recent material on focal plane array read-out electronics.

Volume 4, edited by Michael Dudzik of ERIM, treats system design, analysis, and testing, including adjunct technology and methods such as trackers, mechanical design considerations, and signature modeling.

Volume 5, edited by Stephen Campana of the Naval Air Warfare Center, treats contemporary infrared passive systems such as FLIRs, IRSTs, IR line scanners, and staring array configurations.

Volume 6, edited by Clifton Fox of the Night Vision and Electronic Sensors Directorate, treats active systems and includes mostly new material on laser radar, laser rangefinders, millimeter-wave systems, and fiber optic systems.

Volume 7, edited by David Pollock, consultant, treats a number of countermeasure topics rarely appearing in the open literature.

Volume 8, edited by Stanley Robinson of ERIM, treats emerging technologies such as unconventional imaging, synthetic arrays, sensor and data fusion, adaptive optics, and automatic target recognition.

Acknowledgments

It is extremely difficult to give credit to all the people and organizations that contributed to this project in diverse ways. A significant amount of material in this book was generated by the sheer dedication and professionalism of many esteemed members of the IR and EO community who unselfishly contributed extensive amounts of precious personal time to this effort and to whom the modest honorarium extended was scarcely an inducement. Their contributions speak elegantly of their skills.

Directly involved were some 85 authors and editors from numerous organizations, as well as scores of technical reviewers, copyeditors, graphic artists, and photographers whose skill contributed immeasurably to the final product.

We acknowledge the extensive material and moral support given to this project by various members of the managements of all the sponsoring and supporting organizations. In many cases, organizations donated staff time and internal resources to the preparation of this book. Specifically, we would like to acknowledge J. MacCallum of DoD, W. Brown and J. Walker of ERIM, and J. Yaver of SPIE, who had the foresight and confidence to invest significant resources in the preparation of this book. We also extend our appreciation to P. Klinefelter, B. McCabe, and F. Frank of DTIC for their administrative support during the course of this program.

Supporting ERIM staff included Ivan Clemons, Jenni Cook, Tim Kellman, Lisa Lyons, Judy Steeh, Barbara Wood, and the members of their respective organizations that contributed to this project.

We acknowledge Lorretta Palagi and the publications staff at SPIE for a professional approach to the truly monumental task of transforming the manuscripts into presentable copy and the patience required to interact effectively with the authors.

We would like to pay special tribute to Nancy Hall of the IRIA Center at ERIM who administrated this at times chaotic project with considerable interpersonal skill, marshaling the numerous manuscripts and coordinating the myriad details characteristic of a work of this magnitude.

We properly dedicate this book to the people who created it and trust it will stand as a monument to their skills, experience, and dedication. It is, in the final analysis, a product of the community it is intended to serve.

Joseph S. Accetta
David L. Shumaker
Ann Arbor, Michigan

January 1993

Notices and Disclaimer

Contents

Unconventional Imaging Systems

Carl C. Aleksoff
Environmental Research Institute of Michigan
Ann Arbor, Michigan

Christopher Dainty
Imperial College of Science, Technology and Medicine
London, England

James R. Fienup
Environmental Research Institute of Michigan
Ann Arbor, Michigan

Robert Q. Fugate
Starfire Optical Range
Phillips Laboratory
Kirtland Air Force Base, New Mexico

Jean-Marie Mariotti
Observatoire de Paris–Meudon
Meudon, France

Peter Nisenson
Harvard-Smithsonian Center for Astrophysics
Cambridge, Massachusetts

François Roddier
University of Hawaii at Manoa
Institute for Astronomy
Honolulu, Hawaii

Anthony M. Tai
Environmental Research Institute of Michigan
Ann Arbor, Michigan

CONTENTS

1.1   INTRODUCTION

Sections 1.2 through 1.7 cover the following types of unconventional imaging systems: stellar speckle interferometry, aperture-plane interferometry, passive interferometric range-angle imaging, imaging that uses nonimaged laser speckle patterns, optical range-Doppler imaging, and laser guide star adaptive optics for compensated imaging.

1.2   STELLAR SPECKLE INTERFEROMETRY
Written by Christopher Dainty and Peter Nisenson

The use of optical interferometry to determine the spatial structure of astronomical objects was first suggested by Fizeau[1] in 1868. Stellar interferometers measure, in modern terminology, the spatial coherence of light incident on the earth, and the object intensity (or some parameter such as its diameter) is calculated using the van Cittert-Zernike theorem.[2] Fizeau's suggestion led to the development of specialized long baseline interferometers; Michelson's stellar interferometer[3,4] directly applied Fizeau's method (amplitude interferometry), while the intensity interferometer[5] of Hanbury Brown and Twiss enabled the squared modulus of the spatial coherence function to be measured for thermal sources.

Until recently, single optical telescopes were used in a conventional (noninterferometric) way, their spatial resolution being limited to approximately 1″0 (1 arcsec) due to the presence of atmospheric turbulence or *seeing*. In 1970, Labeyrie[6] invented the technique of stellar speckle interferometry, in which diffraction-limited resolution is obtained from a large single telescope despite the seeing. The diffraction-limited angular resolution $\Delta\alpha$ of a telescope of diameter D operating at wavelength λ is conveniently expressed by the Rayleigh criterion,

$$\Delta\alpha = 1.22\frac{\lambda}{D} \ , \tag{1.1}$$

yielding approximately 0″025 at $\lambda = 400$ nm for a 4-m telescope. The first results by Labeyrie and collaborators were published[7] in 1972 and since then approximately 1150 papers on speckle interferometry have been published.[8]

Labeyrie's important contribution was to recognize that the speckles formed at the focus of a large telescope have an angular size determined by diffraction, i.e., their smallest dimension is given by Eq. (1.1). Diffraction-limited information about an astronomical object can therefore be extracted from short-exposure, narrow-band images by an appropriate method of data reduction.

The basic principles of stellar speckle interferometry are outlined in non-mathematical terms in Sec. 1.2.1, and this is followed by a detailed mathematical discussion of the technique in Sec. 1.2.2. In astronomy, the objects

under observation are often faint and only a limited observation time is available, so that the question of signal-to-noise ratio is very important; this is evaluated in Sec. 1.2.3. In Sec. 1.2.4 we discuss the problem of finding images (or maps) of astronomical objects using speckle data. This is an area of considerable activity both by theoreticians and observers. The equipment required to implement speckle interferometry is described in Sec. 1.2.5, which includes a discussion of the technique of one-dimensional infrared speckle interferometry, which has proved fruitful in recent years. Finally, we conclude with a brief summary of the astronomical results produced by speckle interferometry—these range from measurements of asteriods to quasars and supernovas.

1.2.1 Basic Principles

Figure 1.1 shows highly magnified images of an unresolvable (*point*) and a resolved star taken using a large telescope with an exposure time of approximately 10^{-2} s through a filter of 10-nm bandwidth. In the case of the point source (upper row), the image has a speckle-like structure and it is found that, as with conventional laser speckle patterns, the minimum speckle "size" is of the same order of magnitude as the Airy disk of the telescope. A long-exposure image is simply the sum of many short-exposure ones, each with a speckle structure that is different in detail and is therefore a smooth intensity distribution whose diameter is typically 1″0 in good seeing. Long-exposure images of the point source and resolved star of Fig. 1.1 would show little, if any, difference. The minimum speckle size, on the other hand, is approximately 0″025 for a 4-m telescope at a mean wavelength of 400 nm. By extracting correctly the information in short-exposure images, it is possible to observe detail as small as the limit imposed by diffraction and not be limited to the 1″0 resolution of conventional images.
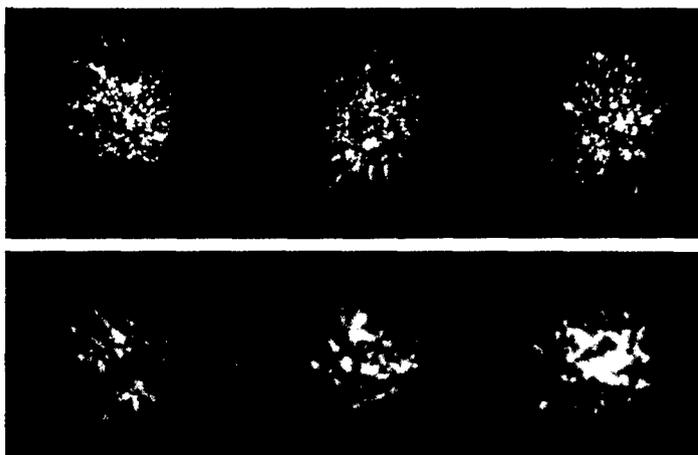
Fig. 1.1 Short-exposure photographs of an unresolved point source (*upper row*) and a resolved star, α-Orionis (*lower row*), taken on a 4-m class telescope. The exposure time and filter bandwidth are 10^{-2} s and 10 nm, respectively. (Courtesy of B. L. Morgan and R. J. Scaddan, Imperial College, London)
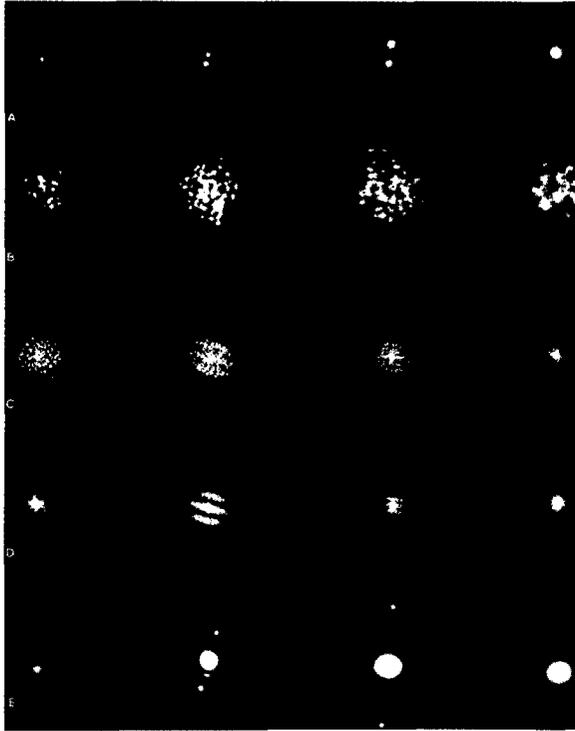
Fig. 1.2 Laboratory simulation showing the principles of stellar speckle interferometry: (A) objects, (B) typical short-exposure photographs, (C) diffraction patterns of row B, (D) sum of 20 diffraction patterns, and (E) diffraction pattern of row D. (Courtesy of A. Labeyrie, CERGA)

A laboratory simulation illustrating the basic method is shown in Fig. 1.2 for an unresolved star, binary stars of two separations, and a resolved star (shown as a uniformly illuminated disk). A large number of short-exposure records are taken, each through a different realization of the atmosphere, typical examples being shown in row B. For a binary star, each component produces an *identical* speckle pattern (assuming isoplanatism and neglecting photon noise), and a *double-speckle* effect may be visible in each short-exposure image in favorable circumstances. The optical diffraction pattern, or squared modulus of the Fourier transform, of a typical short-exposure record is shown in row C for each object. The signal-to-noise ratio (SNR) is low for a single record and may be improved by adding many such diffraction patterns (row D). The unresolved object has a diffraction halo of relatively large spatial extent, the binaries give fringes of a period inversely proportional to their separation, and the resolved object gives a diffraction halo whose diameter is inversely proportional to the diameter of the object. By taking a further Fourier transform of each ensemble-average diffraction pattern we obtain the average spatial (or angular) autocorrelation of the diffraction-limited images of each object (row E).

The term *speckle interferometry* was adopted by Gezari, Labeyrie, and Stachnik.[7] The interferometer is, in fact, the telescope—light from all parts of the pupil propagates to the image plane where it interferes to become a speckle pattern. In other forms of stellar interferometry, the light in the pupil is combined in a different way, for example, using a rotation-shearing interferometer. The beauty of the speckle technique is that the interferometer (i.e., the telescope) is already constructed to the required tolerances.

1.2.2 The Theory of Speckle Interferometry

1.2.2.1 Outline of Theory. For each short-exposure record, the usual quasi-monochromatic, isoplanatic imaging equation applies, provided that the angular extent of the object is not too large[a]:

$$I(\alpha,\beta) = \int\int_{-\infty}^{\infty} O(\alpha',\beta')P(\alpha - \alpha',\beta - \beta') \, d\alpha' \, d\beta'$$

or, in notation,

$$I(\alpha,\beta) = O(\alpha,\beta) \circledast P(\alpha,\beta) \ , \tag{1.2}$$

where $I(\alpha,\beta)$ is the instantaneous image intensity as a function of angle (α,β), $O(\alpha,\beta)$ is the object intensity, $P(\alpha,\beta)$ is the instantaneous point spread function of the atmosphere/telescope system normalized to unit volume, and $\circledast$ denotes the convolution integral.

As we demonstrated in Sec. 1.2.1, the analysis of these data may be carried out in two equivalent ways. In the angular, or spatial, domain, the ensemble-averaged angular autocorrelation function of the image is found; this is defined as

$$C_I(\alpha,\beta) \equiv \left\langle \int\int_{-\infty}^{\infty} I(\alpha',\beta')I(\alpha' + \alpha,\beta' + \beta) \, d\alpha' \, d\beta' \right\rangle$$

or, in notation,

$$C_I(\alpha,\beta) \equiv \langle I(\alpha,\beta) \ast I(\alpha,\beta) \rangle \ , \tag{1.3}$$

where $\ast$ denotes angular autocorrelation. Combining Eqs. (1.2) and (1.3) yields the following relationship between object and image autocorrelation functions:

$$C_I(\alpha,\beta) = C_o(\alpha,\beta) \circledast \langle P(\alpha,\beta) \ast P(\alpha,\beta) \rangle \ , \tag{1.4}$$

where $C_o(\alpha,\beta)$ is the angular autocorrelation function of the object intensity. Note that Eq. (1.4) for the object and image autocorrelation functions is similar

---

[a]Throughout this review, the object and image plane coordinates are taken to be angles (α,β), the coordinates in the Fourier transform plane being angular frequencies (arcsec^{-1}).

in form to Eq. (1.2) for object and image intensities, but with an impulse response equal to $\langle P(\alpha,\beta) \ast P(\alpha,\beta) \rangle$.

In the angular (or spatial) frequency domain, the average squared modulus of the Fourier transform of the image intensity is found. This is correctly referred to as the *average energy spectrum*[b]:

$$\Phi_I(u,v) \equiv \langle |i(u,v)|^2 \rangle \ , \tag{1.5}$$

where

$$i(u,v) \equiv \int\limits_{-\infty}^{\infty}\!\!\int I(\alpha,\beta) \, \exp[-2\pi i(u\alpha + v\beta)] \, d\alpha \, d\beta \ . \tag{1.6}$$

Combining Eqs. (1.2), (1.5), and (1.6) yields the following simple relationship between the energy spectrum of the image $\Phi_I(u,v)$ and that of the object $\Phi_O(u,v)$:

$$\Phi_I(u,v) = \Phi_O(u,v) \cdot \mathcal{T}(u,v) \ , \tag{1.7}$$

where

$$\mathcal{T}(u,v) \equiv \langle |T(u,v)|^2 \rangle \ ,$$

and $T(u,v)$, the instantaneous transfer function, is equal to the Fourier transform of the point spread function:

$$T(u,v) = \int\limits_{-\infty}^{\infty}\!\!\int P(\alpha,\beta) \, \exp[-2\pi i(u\alpha + v\beta)] \, d\alpha \, d\beta \ . \tag{1.8}$$

Because of the similarity between Eq. (1.7) and the Fourier-space isoplanatic imaging equation (in which image frequency components are equal to object frequency components multiplied by an optical transfer function[10]), the quantity $\mathcal{T}(u,v)$ is referred to as the transfer function for speckle interferometry or *speckle transfer function*. Equations (1.4) and (1.7) in the real (angular) and Fourier (angular frequency) domains, respectively, are completely equivalent; Eq. (1.7) is simply obtained by taking the Fourier transform of both sides of Eq. (1.4).

The conventional ("long-exposure") image intensity is found from Eq. (1.2) by ensemble averaging:

$$\langle I(\alpha,\beta) \rangle = O(\alpha,\beta) \circledast \langle P(\alpha,\beta) \rangle \ , \tag{1.9}$$

---

[b]The *energy spectrum* of a function equals the squared modulus of its Fourier transform. If the function is a realization of a square-integrable nonstationary random process, an *ensemble-averaged energy spectrum* can be defined as in Eq. (1.5). A realization of a stationary random process does not possess a Fourier transform, but a *power spectrum* can be defined in terms of a generalized Fourier transform.[9]

where $\langle P(\alpha,\beta)\rangle$ is the average point spread function of the atmosphere/telescope system. In Fourier space, Eq. (1.9) becomes

$$\langle i(u,v)\rangle = o(u,v)\langle T(u,v)\rangle \; , \qquad\qquad (1.10)$$

where $o(u,v)$ is the Fourier transform of the object intensity, and $\langle T(u,v)\rangle$ is the average, or long-exposure, transfer function.

Comparing the conventional long-exposure imaging of Eq.(1.10) to the speckle interferometry of Eq. (1.7), it is clear that the resolution of conventional imaging is governed by the form of the average transfer function $\langle T(u,v)\rangle$, whereas in speckle interferometry the relevant transfer function is $\mathscr{T}(u,v) \equiv \langle |T(u,v)|^2\rangle$. In the following sections we show that the latter function retains high angular-frequency information that is lost in conventional imaging. However, remember that $\mathscr{T}(u,v)$ is a transfer function for energy spectra, whereas $\langle T(u,v)\rangle$ is a transfer function for Fourier components. The loss of Fourier phase information in speckle interferometry is a severe limitation to its usefulness. Possible methods of retrieving the Fourier phase are discussed in Sec. 1.2.4.

1.2.2.2   The Long-Exposure Transfer Function.   To find the optical transfer function of a system, we must consider the imaging of a quasimonochromatic point source as given in Fig. 1.3. For an isoplanatic, incoherent imaging system, the optical transfer function $T(u,v)$ is equal to the normalized spatial autocorrelation of the pupil function $H(\xi,\eta)$:

$$T(u,v) = \frac{\displaystyle\int_{-\infty}^{\infty} H(\xi,\eta)H^*(\xi + \lambda u,\eta + \lambda v) \, d\xi \, d\eta}{\displaystyle\iint_{-\infty}^{\infty} |H(\xi,\eta)|^2 \, d\xi \, d\eta} \; , \qquad\qquad (1.11)$$

where (u,v) are angular frequency coordinates, (ξ,η) are distance coordinates in the pupil, and λ is the mean wavelength.[11] The pupil function $H(\xi,\eta)$ is the
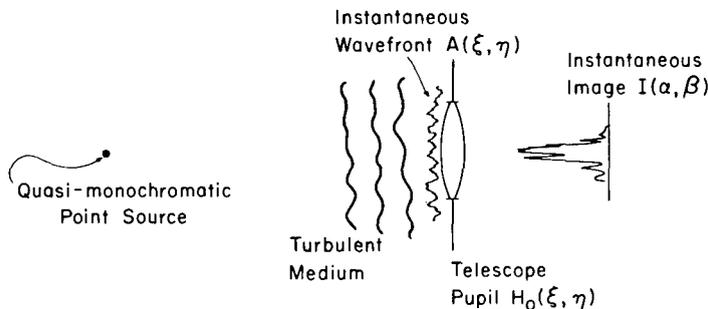
Fig. 1.3   The formation of an instantaneous image of a point source through the atmosphere.

complex amplitude in the exit pupil, relative to the reference sphere centered on the Gaussian focus, due to a point source and in the case of propagation through the turbulent atmosphere may be written as

$$H(\xi,\eta) = A(\xi,\eta)H_0(\xi,\eta) \ , \tag{1.12}$$

where $A(\xi,\eta)$ is the complex amplitude of light from a point source that has propagated through the atmosphere and $H_0(\xi,\eta)$ is the pupil function of the optical system alone.

Substitution of Eq. (1.12) into (1.11) gives

$$T(u,v) = \frac{\displaystyle\int\!\!\int_{-\infty}^{\infty} A(\xi,\eta)A^*(\xi+\lambda u,\ \eta+\lambda v)H_0(\xi,\eta)H_0^*(\xi+\lambda u,\ \eta+\lambda v)\ d\xi\ d\eta}{\displaystyle\int\!\!\int_{-\infty}^{\infty} |A(\xi,\eta)|^2|H_0(\xi,\eta)|^2\ d\xi\ d\eta} \ . \tag{1.13}$$

The long-exposure or average transfer function is found by averaging Eq. (1.13). The lower line is simply the intensity of light integrated over the telescope pupil and is effectively constant for a large telescope and/or weak scintillation. We also assume that $A(\xi,\eta)$ is a (wide-sense) stationary process [i.e., its mean and autocorrelation function in Eq. (1.13) are independent of the ξ,η coordinates], so that the expression for the long-exposure transfer function becomes[12]

$$\langle T(u,v)\rangle = T_s(u,v)T_0(u,v) \ , \tag{1.14}$$

where $T_s(u,v)$ is the atmospheric or *seeing* transfer function,

$$T_s(u,v) = \frac{\langle A(\xi,\eta)A^*(\xi\ +\ \lambda u,\ \eta\ +\ \lambda v)\rangle}{\langle |A(\xi,\eta)|^2\rangle} \ , \tag{1.15}$$

and $T_0(u,v)$ is the optical transfer function of the telescope alone,

$$T_0(u,v) = \frac{\displaystyle\int\!\!\int_{-\infty}^{\infty} H_0(\xi,\eta)H_0^*(\xi\ +\ \lambda u,\ \eta\ +\ \lambda v)\ d\xi\ d\eta}{\left\langle \displaystyle\int\!\!\int_{-\infty}^{\infty} |H_0(\xi,\eta)|^2\ d\xi\ d\eta\right\rangle} \ . \tag{1.16}$$

Thus the long-exposure transfer function is equal to the product of the transfer functions of the atmosphere and telescope.

A detailed discussion of the atmospheric transfer function and other relevant properties of turbulence may be found in Refs. 13 through 16, particularly in the comprehensive review by Roddier.[15] For a Kolmogorov spectrum of tur-

bulence, the average transfer function is rotationally symmetric and is given by

$$T_s(w) = \exp\left[-3.44\left(\frac{\lambda|w|}{r_0}\right)^{5/3}\right] , \qquad (1.17)$$

where $w = (u^2 + v^2)^{1/2}$ and the parameter r_0, first defined by Fried,[13] is equal to the diameter of the diffraction-limited telescope whose Airy disk has the same area as the seeing disk. The parameter r_0 plays an important role in both long-exposure imaging and speckle interferometry. It can be shown that[15]

$$r_0 \propto \lambda^{6/5}(\cos\gamma)^{3/5} , \qquad (1.18)$$

where γ is the zenith angle. Typical values of r_0 lie in the range of 5 to 20 cm at a good observing site in the visible range. Since an r_0 value of 10 cm at $\lambda = 500$ nm is equivalent to $r_0 = 3.6$ m at $\lambda = 10$ μm, it follows that a 4-m class telescope is severely seeing limited in the visible but essentially diffraction limited at 10 μm.

The angular "diameter" of the seeing disk, or seeing angle ω, is defined by

$$\omega \equiv \frac{\lambda}{r_0} , \qquad (1.19)$$

and is therefore proportional to $\lambda^{-1/5}$. At $\lambda = 500$ nm and $r_0 = 10$ cm, the seeing disk has a diameter of approximately 5×10^{-6} rad or 1″0.

Measurements of the long-exposure transfer function and the parameter r_0 have been reported by Dainty and Scaddan,[17] Roddier,[18] and Brown and Scaddan[19] and there is good agreement with Eq. (1.17).

1.2.2.3 The Speckle Transfer Function.

The transfer function of speckle interferometry, $\mathscr{T}(u,v) \equiv \langle|T(u,v)|^2\rangle$, relates the average energy spectrum of the image to that of the object. Using Eq. (1.13), we can write $|T(u,v)|^2$ as $|T(u,v)|^2 = \mathscr{A}(u,v)/\mathscr{B}$, where

$$\mathscr{A}(u,v) = \int\!\!\!\int\!\!\!\int\!\!\!\int_{-\infty}^{\infty} A(\xi_1,\eta_1)A^*(\xi_1 + \lambda u, \eta_1 + \lambda v)A^*(\xi_2,\eta_2)$$

$$\times A(\xi_2 + \lambda u, \eta_2 + \lambda v)H_0(\xi_1,\eta_1)H_0^*(\xi_1 + \lambda u, \eta_1 + \lambda v)$$

$$\times H_0^*(\xi_2,\eta_2)H_0(\xi_2 + \lambda u,\eta_2 + \lambda v)\, d\xi_1\, d\eta_1\, d\xi_2\, d\eta_2$$

and

$$\mathscr{B} = \left[\int\!\!\!\int_{-\infty}^{\infty} |A(\xi,\eta)|^2|H_0(\xi,\eta)|^2\, d\xi\, d\eta\right]^2 . \qquad (1.20)$$

As before, $A(\xi,\eta)$ is assumed to be a stationary random process with weak scintillation. For convenience we define $\langle |A|^2 \rangle \equiv 1$ and the pupil area $\mathscr{P}$,

$$\mathscr{P} \equiv \int\!\!\!\int_{-\infty}^{\infty} |H_0(\xi,\eta)|^2 \, d\xi \, d\eta \tag{1.21}$$

(this is the true pupil area for an unapodized, or clear, pupil).

With the substitution $\xi' = \xi_2 - \xi_1$ and $\eta' = \eta_2 - \eta_1$, Eq. (1.20) yields the following expression for the speckle transfer function:

$$\mathscr{T}(u,v) = \mathscr{P}^{-2} \int\!\!\!\int_{-\infty}^{\infty} \mathscr{M}(u,v;\xi',\eta') \mathscr{H}(u,v;\xi',\eta') \, d\xi' \, d\eta' \ ,$$

where $\mathscr{M}$ is a fourth-order moment,

$$\mathscr{M}(u,v;\xi',\eta') \equiv \langle A(\xi_1,\eta_1) A^*(\xi_1 + \lambda u, \eta_1 + \lambda v) A^*(\xi_1 + \xi', \eta_1 + \eta')$$

$$\times A(\xi_1 + \xi' + \lambda u, \eta_1 + \eta' + \lambda v) \rangle \tag{1.22}$$

and

$$\mathscr{H}(u,v;\xi',\eta') = \int\!\!\!\int_{-\infty}^{\infty} H_0(\xi_1,\eta_1) H_0^*(\xi_1 + \lambda u, \eta_1 + \lambda v)$$

$$\times H_0^*(\xi_1 + \xi', \eta_1 + \eta') H_0$$

$$\times (\xi_1 + \xi' + \lambda u, \eta_1 + \eta' + \lambda v) \, d\xi_1 \, d\eta_1 \ .$$

Clearly, the quantity $\mathscr{M}$ characterizes the atmospheric contribution and $\mathscr{H}$ the telescopic contribution to the speckle transfer function.

Further simplification of Eq. (1.22) requires that an assumption about the joint probability distribution of the process $A(\xi,\eta)$ be made. The most satisfactory distribution is the log normal, in which the log modulus and phase each have a Gaussian probability density. Korff[20] evaluated the speckle transfer function using this model and results are given below; however, neither this model nor the zero-scintillation versions[15] of it have a simple analytical solution and require extensive numerical calculations.

To illustrate in a qualitative way the form of the speckle transfer function, we assume that $A(\xi,\eta)$ is a complex Gaussian process.[21] This is a poor assumption in good seeing, although it improves as the seeing deteriorates; this assumption also violates the weak scintillation requirement for normalization. For a complex Gaussian process, the fourth-order moment of Eq. (1.22) reduces to a product of second-order moments:

$$\mathcal{M}(u,v;\xi',\eta') = \langle A(\xi_1,\eta_1)A^*(\xi_1 + \lambda u, \eta_1 + \lambda v)\rangle$$

$$\times \langle A^*(\xi_1 + \xi', \eta_1 + \eta')A(\xi_1 + \xi' + \lambda u, \eta_1 + \eta' + \lambda v)\rangle$$

$$+ \langle A(\xi_1,\eta_1)A^*(\xi_1 + \xi', \eta_1 + \eta')\rangle$$

$$\times \langle A^*(\xi_1 + \lambda u, \eta_1 + \lambda v)A(\xi_1 + \xi' + \lambda u, \eta_1 + \eta' + \lambda v)\rangle ,$$

which, when substituted into (1.22) yields

$$\mathcal{T}(u,v) = |T_s(u,v)|^2|T_0(u,v)|^2 + \mathcal{P}^{-2}\int\!\!\int_{-\infty}^{\infty} |T_s(\xi'/\lambda, \eta'/\lambda)|^2$$

$$\times \mathcal{H}(u,v;\xi',\eta')\, d\xi'\, d\eta' \ . \tag{1.23}$$

Now $|T_s(\xi'/\lambda, \eta'/\lambda)|^2$ is of width of order r_0/λ and $\mathcal{H}$ is essentially constant for such values of ξ', η', provided that $(u^2 + v^2)^{1/2} < (D - r_0)/\lambda$. The second term of Eq. (1.23) therefore reduces to

$$\mathcal{P}^{-2}\int\!\!\int_{-\infty}^{\infty} |T_s(\xi'/\lambda, \eta'/\lambda)|^2\, d\xi'\, d\eta' \times \mathcal{H}(u,v;0,0) \ , \tag{1.24}$$

except for $(u^2 + v^2)^{1/2} > (D - r_0)/\lambda$.

The first integral in Eq. (1.24) can be evaluated using Eq. (1.17) to give $0.109\pi r_0^2$. The quantity $\mathcal{H}(u,v; 0,0)$ is simply

$$\mathcal{H}(u,v; 0,0) = \int\!\!\int_{-\infty}^{\infty} |H_0(\xi_1,\eta_1)|^2|H_0(\xi_1 + \lambda u, \eta_1 + \lambda v)|^2\, d\xi_1\, d\eta_1 \ ,$$

which, when multiplied by $\mathcal{P}^{-1}$ is the *diffraction-limited* optical transfer function $T_D(u,v)$ for an unapodized, or clear, pupil; and, finally, the remaining $\mathcal{P}^{-1}$ equals $4/\pi D^2$.

Thus the expression for the speckle transfer function reduces to

$$\mathcal{T}(u,v) = |\langle T(u,v)\rangle|^2 + 0.435(r_0/D)^2 T_D(u,v) \ , \tag{1.25}$$

or, defining the number of speckles as

$$N_{\text{sp}} \approx 2.3\left(\frac{D}{r_0}\right)^2 \ ,$$

$$\mathcal{T}(u,v) = |\langle T(u,v)\rangle|^2 + \frac{1}{N_{\text{sp}}}T_D(u,v) \ . \tag{1.26}$$

In both equations we assume that $(u^2 + v^2)^{1/2} \leqq (D - r_0)/\lambda$.

The essential feature of the speckle transfer function, Eqs. (1.25) or (1.26), is that there is a term proportional to the diffraction-limited optical transfer
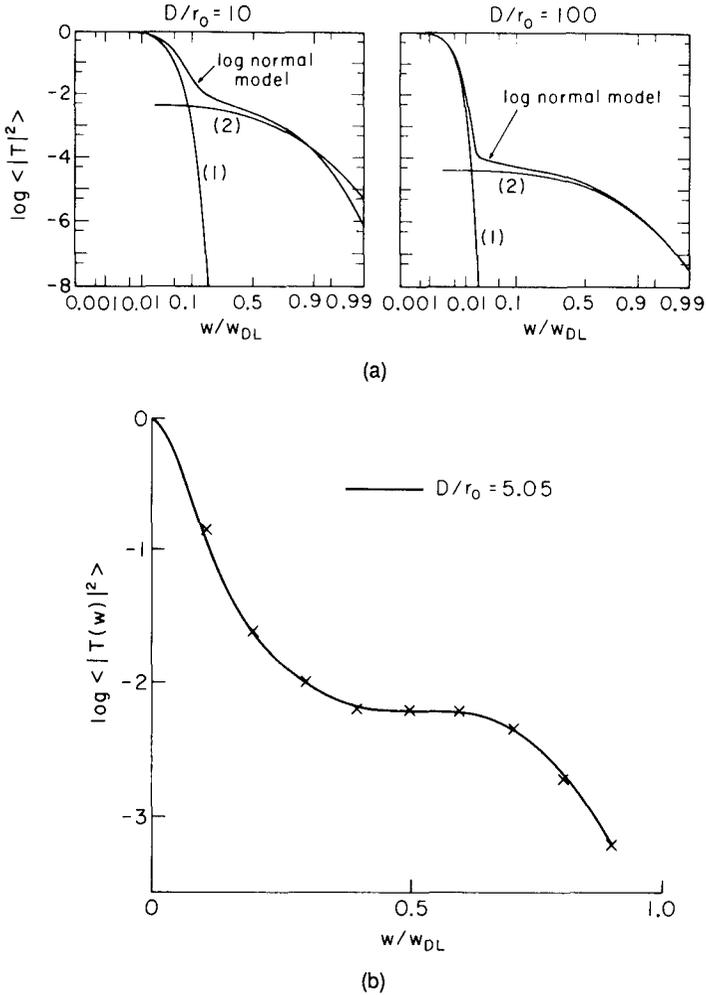
Fig. 1.4 (a) Comparison of the speckle transfer function predicted by the log-normal model with terms (1) and (2) of Eq. (1.26) for the complex Gaussian model, for $D/r_0 = 10$ and 100 (Ref. 22). (b) Comparison of the speckle transfer function predicted by the log-normal model with experimental results.[24]

function that extends almost up to the diffraction-limited cutoff D/λ; expressions (1.25) and (1.26) indicate that this result is *independent of telescope aberrations*,[21] although there is, in fact, a weak dependence on aberrations as discussed in Sec. 1.2.2.4. With $D \cong 4$ m and $r_0 \cong 0.1$ m, the number of speckles N_{sp} is approximately 3.7×10^3, indicating that the diffraction-limited information in the image may be carried with a low signal-to-noise ratio (SNR). However, the normalization of Eqs. (1.25) and (1.26) to unity at zero spatial frequency gives a misleading impression of the SNR, which is best evaluated by other methods (Sec. 1.2.3).

Because Eqs. (1.25) and (1.26) are based on the assumption that $A(\xi,\eta)$ is a complex Gaussian process, they give only the qualitative behavior of the transfer function. The speckle transfer function can be calculated using the log-normal model and these results are compared[22] to Eq. (1.26) in Fig. 1.4(a) for

$D/r_0 = 10$ and 100. The main differences lie in the region between the low- and high-frequency terms. In fact, at low spatial frequencies, the correct asymptotic form of the speckle transfer function is $|\langle T \rangle_{SE}|^2$, where $\langle T \rangle_{SE}$ is the so-called "short-exposure" average[13,20] (i.e., the average when each point image is recentered). Careful measurements by Aime et al.[23] and Chelli et al.[24] are in excellent agreement with the log-normal model, particularly if the effect of the central obstruction and the (small) effect of defocus are allowed for. Figure 1.4(b) shows the result of a measurement in the infrared.

1.2.2.4 Effect of Aberrations. Telescope aberrations have two potential effects on the speckle transfer function. If they are very severe, optical path differences greater than the coherence length of the light may be introduced, which would lead to a strong attenuation of the transfer function. Proper analysis of this effect requires a detailed consideration of temporally partially coherent imaging; this is not carried out here since the effects in normal circumstances are small, as the following analysis shows.

Consider the simplest aberration—defocus—of magnitude m waves at the edge of the pupil. The longitudinal and angular transverse ray aberrations Δz and $\Delta \alpha$, respectively, are given by

$$\Delta z = \frac{8m\lambda f^2}{D^2} ,$$

$$\Delta \alpha = \frac{8m\lambda}{D} .$$

$$(1.27)$$

Under most observing conditions, focus can be established to a tolerance $\Delta \alpha$ of less than $1\rlap{.}{''}0$, giving a maximum value of m of approximately 5λ for a 4-m telescope. The coherence length l_c of light of bandwidth $\Delta \lambda$ is given approximately by

$$l_c \cong \frac{\overline{\lambda}^2}{\Delta \lambda} ,$$

$$(1.28)$$

and with typical bandwidths ($\Delta \lambda = 20$ nm, $\overline{\lambda} = 500$ nm) it is clear that $l_c > m\lambda$. In practice, aberrations only introduce path differences greater than the coherence length if the bandwidth is large or the aberrations are severe.

Aberrations also affect the shape of the speckle transfer function in the quasimonochromatic case; their effect reduces as the ratio D/r_0 increases and disappears in the limit $D/r_0 \to \infty$. The effect of several aberrations was investigated by Dainty[25] using the complex Gaussian model for atmospheric turbulence. More precise calculations for defocus and astigmatism were made by Roddier et al.[26] using the log-normal model and were compared to the measurements of Karo and Schneiderman.[27] These results are shown in Fig. 1.5; it should be emphasized that the defocus in this case was made artificially large to illustrate the effect, with $m \cong 6.4\lambda$ corresponding to an angular transverse ray aberration (of extremal rays) of $\Delta \alpha \cong 3\rlap{.}{''}3$.

For aberrations other than defocus, intuitive reasoning based on the approximations necessary to obtain Eqs. (1.25) and (1.26) suggests that the effect
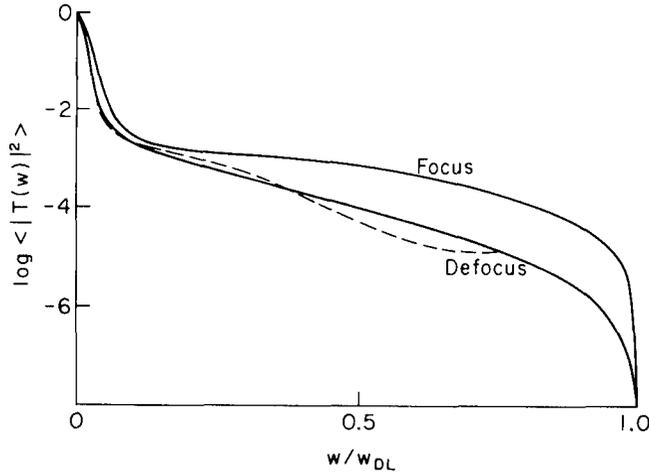
Fig. 1.5 Solid line—theoretical speckle transfer functions for $D/r_0 = 19.2$ in focus and defocused by $6.4l$. Broken line—curve observed by Karo and Schneiderman[27] under defocused conditions.

of aberrations is small if the seeing disk is larger than the point spread function due to telescope aberrations alone. Thus, a telescope of poor optical quality achieves *diffraction-limited* angular resolution if sufficiently severe atmospheric turbulence (real or artificially induced) is present.[21] Unfortunately, poor seeing (small r_0) also results in a low SNR (Sec. 1.2.3).

1.2.2.5 Effect of Exposure Time.

In practice, each image is the result of a finite exposure time Δt, which always has the effect of attenuating the speckle transfer function. Let the instantaneous point spread function at time t be denoted $P(\alpha,\beta,t)$ and the instantaneous transfer function be $T(u,v,t)$. The speckle transfer function for instantaneous exposures ($\Delta t \to 0$) is defined by

$$\mathcal{T}(u,v) \equiv \langle |T(u,v,t)|^2 \rangle \ , \tag{1.29}$$

whereas for an exposure time Δt it is equal to

$$\mathcal{T}_{\Delta t}(u,v) = \frac{1}{\Delta t^2} \int\int_0^{\Delta t} \langle T^*(u,v,t)T(u,v,t') \rangle \, dt \, dt' \ . \tag{1.30}$$

The term $\langle \cdot \rangle$ in Eq. (1.30) is called the *temporal cross-energy spectrum* and plays an important role in time-integration effects. Assuming temporal stationarity of the process $T(u,v,t)$, Eq. (1.30) may also be written

$$\mathcal{T}_{\Delta t}(u,v) = \frac{1}{\Delta t} \int_{-\Delta t}^{+\Delta t} \left(1 - \frac{|\tau|}{\Delta t} \right) \langle T^*(u,v,t)T(u,v,t+\tau) \rangle \, d\tau \ . \tag{1.31}$$

The finite exposure time speckle transfer function, $\mathcal{T}_{\Delta t}(u,v)$, is always less than (or equal to) the instantaneous transfer function $\mathcal{T}(u,v)$, as the following analysis shows.[28] The Schwartz inequality implies

$$|\langle T^*(u,v,t)T(u,v,t + \tau)\rangle| \leqq \langle|T(u,v,t)|^2\rangle \ ,$$

so that, using Eq. (1.31),

$$\mathcal{T}_{\Delta t}(u,v) \leqq \frac{1}{\Delta t} \int_{-\Delta t}^{+\Delta t} \left(1 - \frac{|\tau|}{\Delta t}\right) |\langle T^*(u,v,t)T(u,v,t + \tau)\rangle| \ d\tau$$

$$\leqq \frac{1}{\Delta t} \int_{-\Delta t}^{+\Delta t} \left(1 - \frac{|\tau|}{\Delta t}\right) \langle|T(u,v,t)|^2\rangle \ d\tau$$

$$= \langle|T(u,v,t)|^2\rangle \equiv \mathcal{T}(u,v) \ ,$$

that is,

$$\mathcal{T}_{\Delta t}(u,v) \leqq \mathcal{T}(u,v) \ . \tag{1.32}$$

This is a general result and is independent of the detailed nature of the turbulence. The magnitude of the attenuation of $\mathcal{T}(u,v)$ due to an exposure time Δt depends, from an experimental point of view, on the form of the temporal cross-energy spectrum $\langle T^*(u,v,t)T(u,v,t + \tau)\rangle$; only qualitative estimates of this function have been reported.[29]

The temporal cross-energy spectrum is equal to the Fourier transform of the spatially averaged space-time[c] intensity correlation function,

$$\langle T^*(u,v,t)T(u,v,t + \tau)\rangle$$

$$= \int\int_{-\infty}^{\infty} \left[\int\int_{-\infty}^{\infty} \langle P(\alpha,\beta,t)P(\alpha + \Delta\alpha, \beta + \Delta\beta, t + \tau)\rangle \ d\alpha \ d\beta\right]$$

$$\times \exp[-2\pi i(u\Delta\alpha + v\Delta\beta)] \ d\Delta\alpha \ d\Delta\beta \ , \tag{1.33}$$

where $\langle P(\alpha,\beta,t)P(\alpha + \Delta\alpha, \beta + \Delta\beta, t + \tau)\rangle$ is the space-time cross-correlation function of the instantaneous point spread function. A few measurements of the spatially integrated space-time cross-correlation function have been made.[30,31] They show that, in general, this function is not cross-spectrally pure, so that it cannot be written as the product of two separable functions of (u,v) and t,

$$\langle T^*(u,v,t)T(u,v,t + \tau)\rangle \neq \mathcal{T}(u,v)C(\tau) \ . \tag{1.34}$$

---

[c]As given in Eq. (1.33), this is an angle-time correlation function; the name *space-time* is more widely used, distances (x,y) in the image plane being related to angles (α,β) by $x = \alpha f$ and $y = \beta f$, where f is the focal length.

(This result is referred to in Sec. 1.2.3.3 on the optimum exposure time.)

When $\Delta\alpha = \Delta\beta = 0$, the space-time cross-correlation is simply equal to the temporal autocorrelation of the point spread function $\langle P(\alpha,\beta,t)P(\alpha, \beta, t + \tau)\rangle$. Several measurements of this function have been reported[30-33] and a sample of results taken at Mauna Kea, Hawaii, are shown in Fig. 1.6; the average correlation time of the image intensity was 15 ms for a 61-cm telescope. In site testing for new locations for stellar interferometry, it is important to measure both the spatial and temporal properties of seeing.

Although it is the cross-energy spectrum that most directly influences the effect of a finite exposure time Δt, from a more fundamental point of view, the important quantity is the fourth-order correlation function of the complex amplitude in the pupil:

$$\langle A(\xi,\eta,t)A^*(\xi + \xi_1, \eta + \eta_1, t)$$
$$\times A^*(\xi + \xi_2, \eta + \eta_2, t + \tau)A(\xi + \xi_3, \eta + \eta_3, t + \tau)\rangle \tag{1.35}$$

[compare with the expression for $\mathcal{M}$ in Eq. (1.22)]. For both complex Gaussian and log-normal complex amplitude, this fourth-order moment is determined by the behavior of the second-order moment. Roddier and coworkers[34,35] have calculated the effect of a finite exposure time on the speckle transfer function using the log-normal model and the assumption that the complex amplitude $A(\xi,\eta,t)$ moves rigidly across the telescope pupil (the Taylor approximation). For a velocity v along the ξ axis.

$$\langle A(\xi,\eta,t)A^*(\xi + \Delta\xi, \eta + \Delta\eta, t + \tau)\rangle = f(\Delta\xi - v\tau, \Delta\eta) \ . \tag{1.36}$$
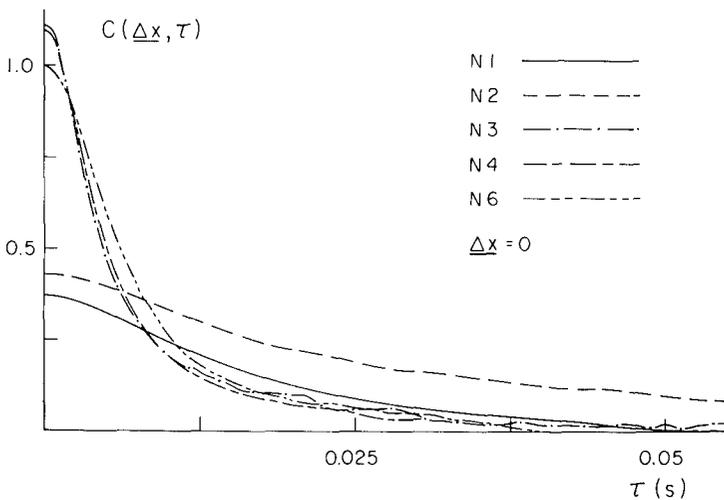

Fig. 1.6 Temporal image intensity autocorrelations over five nights at Mauna Kea, Hawaii, measured using a 61-cm telescope.[31]
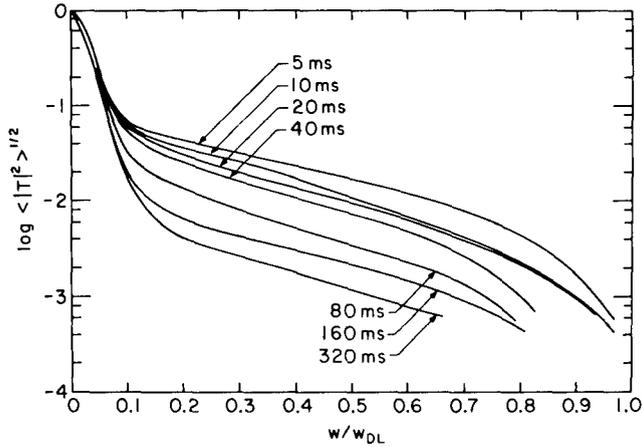
Fig. 1.7   The effect of finite exposure time on the speckle transfer function.[36]

For a telescope of diameter D, a velocity v of the turbulence implies a characteristic image time scale of D/v. The results of Ref. 34 show that the attenuation of the transfer function is not too severe provided that $\Delta t < D/v$.

Spatio-temporal measurements of $|A|^2$ imply that, in addition to the rigid translation described by Eq. (1.36), there is also a strong decorrelation due to *boiling* of $A(\xi,\eta,t)$. This can be explained by a multilayer model for the turbulence[15] with a velocity distribution Δv of the atmospheric layers; this leads to a characteristic time scale of $r_0/\Delta v$ and a uniform attenuation of the high-frequency part of the speckle transfer function.

Karo and Schneiderman[36] have measured the effect of a finite exposure time on the speckle transfer function; their results obtained on the 1.6-m telescope at Maui, Hawaii, are shown in Fig. 1.7. Unfortunately, the spatio-temporal atmospheric data required to compare these measurements with theory were not available. However, the uniform attenuation suggests that the wavefront boiling dominated over simple rigid translation and implies a time-scale consistent with $r_0/\Delta v \cong 20$ ms.

1.2.2.6   Effect of Finite Bandwidth.   A finite bandwidth $\Delta\lambda$, centered at $\bar{\lambda}$, has two effects, both of which attenuate the speckle transfer function. These effects are identical to those observed in polychromatic laboratory-generated Fraunhofer plane speckle patterns. The two effects are (1) a radial dispersion effect similar to that produced by a grating and (2) a loss of speckle contrast caused by atmospheric (or, possibly, telescope induced) optical path differences being comparable to the coherence length $l_c \equiv \bar{\lambda}^2/\Delta\lambda$ of the radiation.

In accordance with simple first-order grating theory, a spread of wavelengths $\Delta\lambda/\bar{\lambda}$ causes a spread in diffraction angles $\Delta\omega/\bar{\omega}$:

$$\frac{\Delta\omega}{\bar{\omega}} = \frac{\Delta\lambda}{\bar{\lambda}} \ .$$

Taking $\bar{\omega}$ to be the seeing angle λ/r_0 [Eq. (1.19)], and defining ω_0 to be the angular diameter of a speckle ($\equiv \lambda/D$), we find the fractional radial elongation of speckles, $\Delta\omega/\omega_0$, at the seeing angle to be given by

$$\frac{\Delta\omega}{\omega_0} = \frac{\Delta\lambda}{\bar{\lambda}} \frac{D}{r_0} \ .$$

To determine a criterion for the maximum permissible value of $\Delta\lambda/\bar{\lambda}$, we require that $\Delta\omega/\omega_0 < 1$, yielding

$$\left|\frac{\Delta\lambda}{\bar{\lambda}}\right|_1 < \frac{r_0}{D} \ . \tag{1.37}$$

To calculate a criterion for coherence length effects to be negligible, we require a formula for the root-mean-square optical path fluctuation $\sigma_z(\xi)$ between two points spaced ξ apart in the telescope pupil. The Kolmogorov theory[15] predicts

$$\sigma_z(\xi) \cong 0.42\bar{\lambda}\left(\frac{\xi}{r_0}\right)^{5/6} \ , \tag{1.38}$$

in which $\sigma_z(\xi)$ is in fact independent of wavelength since $r_0 \propto \lambda^{6/5}$. Thus over a telescope aperture of diameter D we may estimate σ_z by substituting $\xi = D$ in Eq. (1.38); requiring that the coherence length $l_c > \sigma_z$, we obtain

$$\left(\frac{\Delta\lambda}{\bar{\lambda}}\right)_2 < 2.4\left(\frac{r_0}{D}\right)^{5/6} \ . \tag{1.39}$$

Other, more stringent, criteria have been suggested.[14] For typical $r_0 = 0.1$ m and $D = 4$ m, criteria (1.37) and (1.39) yield

$$\left(\frac{\Delta\lambda}{\bar{\lambda}}\right)_1 < 0.025 \ , \qquad \left(\frac{\Delta\lambda}{\bar{\lambda}}\right)_2 < 0.111 \ ,$$

implying that the chromatic dispersion effect is dominant and that, for $\bar{\lambda} = 500$ nm, the bandwidth $\Delta\lambda$ should be less than 12.5 nm.

Measurements by Karo and Schneiderman[36] with $D/r_0 \cong 14$ show no discernible effect on the speckle transfer function for $\Delta\lambda/\bar{\lambda} < 0.06$; this is consistent with $(\Delta\lambda/\lambda)_1 < 0.07$ given by criterion (1.37). Even for $\Delta\lambda/\bar{\lambda} \cong 0.14$, the midfrequencies of $\mathcal{T}(u,v)$ were attenuated by only a factor of 2.

Since the chromatic dispersion effect is important, it may be worthwhile to design a relay optical system that removes the dispersion.[37] Various optical systems have been suggested for this,[38] but they suffer by having a very small effective field angle and no design has yet been successfully incorporated into a speckle camera system.

1.2.2.7 Isoplanicity.

If a linear system is nonisoplanatic (i.e., if its point spread function depends on both object and image coordinates), then the ele-

mentary convolution relationship of Eq. (1.2) is replaced by

$$I(\alpha,\beta) = \int\!\!\int_{-\infty}^{\infty} O(\alpha',\beta')P(\alpha - \alpha', \beta - \beta'; \alpha', \beta') \, d\alpha' \, d\beta' \ , \tag{1.40}$$

where $P(\Delta\alpha,\Delta\beta; \alpha',\beta')$ is the instantaneous point spread function for an object point at (α',β'). There is now no meaningful concept of an instantaneous transfer function or a speckle transfer function. However, defining $T(u,v; \alpha',\beta')$ to be the Fourier transform of $P(\alpha - \alpha', \beta - \beta'; \alpha', \beta')$ with respect to the variables (α,β), the average image energy spectrum $\Phi_I(u,v)$ reduces to

$$\Phi_I(u,v) \equiv \langle|i(u,v)|^2\rangle = \int\!\!\int_{-\infty}^{\infty} C_O(\alpha_1,\beta_1)\langle T(u,v; \alpha',\beta')$$

$$\times\ T^*(u,v; \alpha' - \alpha_1, \beta' - \beta_1)\rangle \exp[-2\pi i(u\alpha_1 + v\beta_1)] \, d\alpha_1 \, d\beta_1 \ , \tag{1.41}$$

where $\alpha_1 = \alpha - \alpha'$ and $\beta_1 = \beta - \beta'$.

If the function $T(u,v; \alpha',\beta')$ is independent of the object point (α',β'), i.e., the imaging is isoplanatic, then Eq. (1.41) simplifies to the usual result of Eq. (1.7):

$$\Phi_I(u,v) = \Phi_O(u,v)\langle|T(u,v)|^2\rangle \ .$$

However, according to Eq. (1.41), there is no longer a simple relationship between object and image properties, and the form of the cross spectrum,

$$\langle T(u,v; \alpha',\beta')T^*(u,v); \alpha' - \alpha_1, \beta' - \beta_1)\rangle \ ,$$

between speckle patterns produced by two point sources separated by angle (α_1,β_1) plays an important role.

Korff et al.,[39] Shapiro,[40] and Fried[22] have investigated this problem using the log-normal model for atmospheric turbulence. However, a more complete analysis can be carried out if the complex Gaussian model of the wavefront $A(\xi,\eta)$ is used, as shown by Roddier et al.[41] Using a multiple-layer model for the turbulence, they estimate the *atmospheric isoplanatic angle* $\delta\omega$ to be given by

$$\delta\omega \cong 0.36\frac{r_0}{\Delta h} \ , \tag{1.42}$$

where Δh is a measure of the altitude dispersion of the turbulent layers.[41] This simple relationship does not reveal the fact that high angular frequencies decorrelate more rapidly than lower ones as the angle of separation (α_1,β_1) increases, but gives a good estimate of the extent of the isoplanatic region. Based on measured profiles of the variation of turbulence with altitude (see Vernin in Ref. 34), predicted isoplanatic angles[41] were in the range 1″9 to 5″4 over six nights at Haute Provence Observatory, with an average of 3″1.

Several measurements of the isoplanatic angle or related quantities have been reported.[29,42–44] The values vary widely, with the most reliable quantitative estimates[43,44] being in the range of $1\overset{\prime\prime}{.}5$ to $5\overset{\prime\prime}{.}0$, i.e., the same order of magnitude as the theoretical predictions. Qualitative estimates, based on the successful implementation of speckle holography,[29,44] indicate some correlation of image intensity for stars as far apart as $22\overset{\prime\prime}{.}0$.

1.2.2.8 Amplitude Calibration for Speckle Interferometry.

To recover the energy spectrum of the object $\Phi_O(u,v)$, the average energy spectrum of the image $\Phi_I(u,v)$ is divided by the speckle transfer function

$$\Phi_O(u,v) = \Phi_I(u,v)/\mathcal{T}(u,v) \ . \tag{1.43}$$

A more robust approach to amplitude calibration for noisy data is to construct a Wiener (or optimal) filter[45,46] in the form

$$F(u,v) = \frac{T^*(u,v)}{|T(u,v)|^2 + K|N(u,v)|^2} \ , \tag{1.44}$$

where $N(u,v)$ is the energy spectrum of the noise in the data and K is a scaling constant that allows adjustment of the *smoothness* of the result. The object spectrum then is given by

$$\Phi_O(u,v) = \Phi_I(u,v)F(u,v) \ . \tag{1.45}$$

At high spatial frequencies, where the speckle transfer function is small and the SNR is generally low, the Wiener filter provides a form of damping in the division, limiting overshoot in the quotient.

In practice, the speckle transfer function is estimated by finding the average energy spectrum for a point source (or reference star). Unfortunately, as we have seen in previous sections, the speckle transfer function depends on a number of atmospheric parameters (such as r_0 and time scale) and these parameters themselves vary considerably over both short (seconds) and long (hours) periods of time. Under stable atmospheric conditions, application of Eq. (1.43) is straightforward, but under (more typical) unstable conditions, the use of Eq. (1.43) can lead to considerable error in the estimation of the object's spectrum. This is less critical for measurements of simple structural features of an object (e.g., the vector separation of a binary star) but crucial for photometric features (e.g., magnitude difference of a binary star).

Several approaches to this problem have been suggested. One is to make simultaneous measurements of r_0 and use the established theory to predict the form of the speckle transfer function.[15] The measurements of Aime et al.[23] and Chelli et al.[24] suggest that the instantaneous ($\Delta t \rightarrow 0$), narrow-band ($\Delta\lambda \rightarrow 0$) speckle transfer function can be predicted for an aberration-free telescope, but in practice focusing errors, aberrations, the finite exposure time, and other effects may influence it. Nevertheless, this appears to be a promising technique, particularly in the infrared where other approaches are less reliable.

The most commonly used amplitude calibration technique is to choose a reference star with as small an angular separation from the object as possible.

Use of a detector with a large, linear dynamic range greatly increases the number of possible reference stars in any area of the sky. Observations are performed by frequently switching from object to reference (on a time scale of one to a few minutes). Under most conditions, the atmosphere is sufficiently stable so that adjacent observations have very similar statistics and a stable speckle transfer function. This is particularly true for visible wavelength observations.

In the infrared, better amplitude calibration is required and a technique called *binning*[47] has been shown to be very effective in reducing calibration problems. Binning requires two passes through the data. The first pass sorts the frames by their estimated transfer function and "bins" frames (i.e., performs a speckle average) with nearly the same seeing, but separately for the object and reference data. Calibration is then performed by deconvolving subsets of object data by subsets of reference data and the results summed to obtain the final result. One limitation is that the SNR in each frame must be sufficient to allow reasonably accurate binning, so the process is limited to bright objects.

1.2.3   Signal-to-Noise Ratio

In the visible region of the spectrum, the SNR of a measurement and the limiting magnitude of speckle interferometry are ultimately determined by the fluctuations imposed by the atmospheric turbulence and the quantum nature of radiation. Although early film-based speckle cameras were limited by other types of noise, the improvement in detector technology over the past decade has made available detectors that are photon-noise limited.[48] Thus in this section we discuss only the fundamental noise sources relevant to visible light speckle interferometry (the infrared case is discussed in Sec. 1.2.5.2).

Let Q be some quantity that is to be estimated by speckle interferometry; Q may be (1) a point in the energy spectrum $\Phi_O(u,v)$ of the object, (2) a point in the autocorrelation function $C_O(\alpha,\beta)$ of the object, or (3) a parameter derived from the autocorrelation function or energy spectrum, such as the diameter of a star, binary separation, or magnitude difference.

We define the SNR of this measurement as

$$\text{SNR} \equiv \frac{\text{expected value of quantity}}{\text{standard deviation of estimate}} \; ,$$

or

$$\text{SNR} \equiv \frac{\langle Q \rangle}{[\text{var}(Q)]^{1/2}} \; , \qquad (1.46)$$

where $\text{var}(Q) \equiv \langle Q^2 \rangle - \langle Q \rangle^2$ is the variance of Q. In the analysis that follows, the SNRs relate to an estimate of Q based on a single frame of data. Normally, one would take M frames of data and, provided these are statistically independent, the overall SNR for the M frames $(\text{SNR})_M$, is simply given by

$$(\text{SNR})_M = \text{SNR} \times M^{1/2} \; . \qquad (1.47)$$

The SNR is the inverse of the relative error of measurement and in a given astronomical application we would normally be interested in the relative error on some parameter (such as diameter), as in (3). However, each problem has its own specific parameters of interest and to keep our results as general as possible we will consider the SNR of the energy spectrum or autocorrelation function.

Several investigations of the SNR of a measurement of the autocorrelation function have been made[25,49–53] and Ref. 54 outlines this approach. However, it has been shown[55] that the autocorrelation and energy spectrum approaches give exactly equivalent SNRs, although the detailed expressions show little apparent similarity. The decision whether to use the autocorrelation method or the energy spectrum method of data reduction should be based on operational considerations and not on SNR considerations. Thus in the following subsection we evaluate only the SNR of the energy spectrum of the object.

1.2.3.1   SNR at a Point in the Energy Spectrum.

The SNR at a point in the energy spectrum was first evaluated by Roddier[56] and subsequently in more detail by several authors[57–60] and reviewed in detail in Ref. 61. In this analysis we use one-dimensional notation for simplicity, and it is convenient to deal with energy spectra of the image and object that are normalized to unity at zero angular frequency, denoted by $\hat{\Phi}_I(u)$ and $\hat{\Phi}_O(u)$, respectively. These are related in the usual way,

$$\hat{\Phi}_I(u) = \hat{\Phi}_O(u)\mathcal{T}(u) ,\tag{1.48}$$

where the speckle transfer function $\mathcal{T}(u)$ in the frequency range of interest is given by Eq. (1.26):

$$\mathcal{T}(u) = \frac{1}{N_{\rm sp}}T_D(u) , \qquad \frac{r_0}{\lambda} < u < \frac{(D - r_0)}{\lambda} .$$

We model the j'th image, $D_j(\alpha)$, as an inhomogeneous or compound Poisson process, which has a rate proportional to the classical image intensity $I_j(\alpha)$, i.e.,

$$D_j(\alpha) = \sum_{k=1}^{N_j} \delta(\alpha - \alpha_{j\,k}) ,$$

where each delta function represents a detected photon event, $\alpha_{j\,k}$ is the location of the k'th event in the j'th frame and N_j is the number of detected photons in the j'th frame. In an observation, the squared modulus of the Fourier transform $|d_j(u)|^2$ is computed for each frame. It is straightforward to show that the average of this is given by[57]

$$\langle|d_j(u)|^2\rangle = \overline{N}^2\hat{\Phi}_I(u) + \overline{N} ,\tag{1.49}$$

where $\overline{N}$ is the average number of detected photons per frame. It follows that the energy spectrum of the photon data, $\langle|d_j(u)|^2\rangle$, is a biased estimate of $\hat{\Phi}_I(u)$

to the presence of the $\overline{N}$ term; in the realistic case in which the photon events are not delta functions but have a unit volume spread function $S(\alpha)$, the second term would be $\overline{N}|s(u)|^2$.

There are two estimators Q whose average yield an unbiased estimate of the image energy spectrum. One possibility is to subtract the average number $\overline{N}$ from each $|d_j(u)|^2$,

$$Q_1 = |d_j(u)|^2 - \overline{N} \; , \tag{1.50}$$

and the second possibility is to subtract the actual number N_j,

$$Q_2 = |d_j(u)|^2 - N_j \; . \tag{1.51}$$

In either case, the average values of Q are unbiased estimators,

$$\langle Q_1 \rangle = \langle Q_2 \rangle = \overline{N}^2 \hat{\Phi}_I(u) \; . \tag{1.52}$$

For the first estimator, the variance is equal to[58]

$$\mathrm{var}(Q_1) = \overline{N} + \overline{N}^2 + 2(2 + N)N^3\hat{\Phi}_I(u) + \overline{N}^2\hat{\Phi}_I(2u) + \overline{N}^4\hat{\Phi}_I^2(u) \; . \tag{1.53}$$

As in all problems of this type, the fluctuation at frequency u is influenced by the value of the energy spectrum at frequency $2u$. At exceedingly low light levels $\overline{N} \ll 1$ (probably of no practical interest!), the SNR per frame for estimate Q_1 is, from Eqs. (1.46), (1.52), and (1.53),

$$\mathrm{SNR} = \overline{N}^{3/2}\hat{\Phi}_I(u) \; , \qquad \overline{N} \ll 1 \; . \tag{1.54}$$

The use of definition (1.50) for the estimate Q_1 has the disadvantage that the noise associated with Q_1 contains a contribution arising from N_j, the actual number of photons per frame. These fluctuations are related to the brightness of the object and not to its structure. If one is interested in the morphology of the object, Q_2 is a better estimate; its variance is given by[60]

$$\mathrm{var}(Q_2) = \overline{N}^2 + \overline{N}^2\hat{\Phi}_I(2u) + 2\overline{N}^3\hat{\Phi}_I(u) + \overline{N}^4\hat{\Phi}_I^2(u) \; . \tag{1.55}$$

If we consider only frequencies $u > \frac{1}{2}D/\lambda$, the second term in Eq. (1.55) can be ignored, yielding a SNR per frame of

$$\mathrm{SNR} = \frac{\overline{N}\hat{\Phi}_I(u)}{1 + \overline{N}\hat{\Phi}_I(u)} \; , \tag{1.56}$$

which is the general expression for the SNR at any point ($u > \frac{1}{2}D/\lambda$) in the energy spectrum of the image. If the speckle transfer function is known exactly

(this is never true in practice), then Eq. (1.56) is also the SNR at a point in the energy spectrum of the object. Substituting Eqs. (1.26) and (1.48) into Eq. (1.56) and defining the average number of detected photons per speckle $\bar{n}$ as

$$\bar{n} \equiv \frac{\overline{N}}{N_{\mathrm{sp}}} = \frac{\overline{N}}{2.3}\left(\frac{r_0}{D}\right)^2 , \tag{1.57}$$

we find that the SNR per frame becomes

$$\mathrm{SNR} = \frac{\bar{n} T_D(u)\hat{\Phi}_O(u)}{1 + \bar{n} T_D(u)\hat{\Phi}_O(u)} . \tag{1.58}$$

Two limiting cases are of interest: (1) very bright objects and (2) very faint ones. For very bright objects, such that

$$\bar{n} T_D(u)\hat{\Phi}_O(u) \gg 1 ,$$

then

$$\mathrm{SNR} \cong 1 . \tag{1.59}$$

Note that the SNR per frame cannot exceed unity in speckle interferometry and this is one of the disadvantages of the speckle technique, compared to pupil plane interferometry, for bright objects.

For very faint objects, such that

$$\bar{n} T_D(u)\hat{\Phi}_O(u) \ll 1 ,$$

then

$$\mathrm{SNR} \cong \overline{N}\hat{\Phi}_I(u)$$

$$\cong \bar{n} T_D(u)\hat{\Phi}_O(u) , \tag{1.60}$$

where, as before, $\overline{N}$ is the average number of detected photons per frame and $\bar{n}$ is the average number per speckle. This particularly simple formula for the SNR per frame at a point ($u > \frac{1}{2}D/\lambda$) in the energy spectrum of the object is valid in practice for all fainter objects.

An example of the variation of SNR per frame as a function of $\overline{N}$ is shown in Fig. 1.8 for $D/r_0 = 10, 20$, and 40. For faint objects, the SNR is proportional to r_0^2, so that there is a strong dependence of SNR on the seeing. On the other hand, since the average number of photons per speckle [Eq. (1.57)] is independent of telescope diameter, the SNR at a point in the energy spectrum is also independent of telescope diameter for faint objects. Of course, a larger telescope yields more independent points in the energy spectrum.
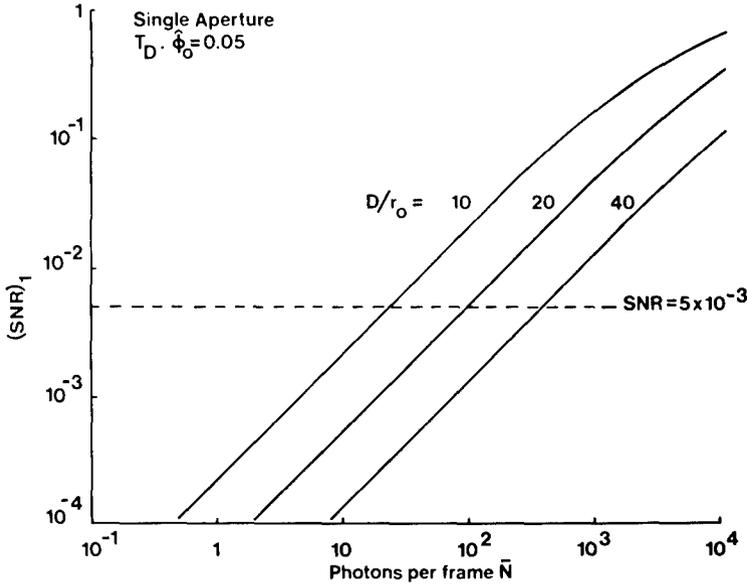
Fig. 1.8 The variation of SNR per frame with the average number of detected photons per frame $\bar{N}$ for $D/r_0 = 10$, 20, and 40, for $T_D d\hat{\Phi}_O = 0.5$ (Ref. 61).

1.2.3.2 Optimum Exposure Time. In the low-light-level case, the SNR at a point in the energy spectrum for M statistically independent frames is, from Eqs. (1.47) and (1.60),

$$(\text{SNR})_M \cong \sqrt{M}\bar{N}\hat{\Phi}_I(u) \ . \tag{1.61}$$

It appears at first sight that a larger exposure time (i.e., increasing $\bar{N}$) leads to a higher SNR; however, this is true only up to an optimum exposure time, after which the decrease in M and $\hat{\Phi}_I(u)$ dominates. The optimum exposure time has been evaluated by Walker[52] and O'Donnell and Dainty.[62]

Let the exposure time be denoted by Δt, the experiment time by T_e, and the photon rate by $\mu \equiv \bar{N}/\Delta t$; then, assuming that neighboring exposures are always statistically independent[d] Eq. (1.61) can be rewritten as

$$(\text{SNR})_M \cong \mu(T_e \Delta t)^{1/2}\hat{\Phi}_{I,\Delta t}(u) \ , \tag{1.62}$$

where $\hat{\Phi}_{I,\Delta t}(u)$ is the measured image energy spectrum for an exposure time Δt. The temporal behavior of the image intensity was discussed in Sec. 1.2.2.5,

---

[d]Clearly, neighboring exposures cannot be statistically independent unless $\Delta t \gg$ correlation time of the image intensity; when there are only a small number of detected photons per frame, however, there is an approximate statistical independence for neighboring frames.

where we showed that the measured image energy spectrum may always be written [see Eq. (1.31)]:

$$\hat{\Phi}_{I,\Delta t}(u) = \frac{1}{\Delta t}\int_{-\Delta t}^{\Delta t}\left(1 - \frac{|\tau|}{\Delta t}\right)\langle i^*(u,t)i(u,t + \tau)\rangle \, d\tau \ . \tag{1.63}$$

Both theory and experiment show that, in general, the cross spectrum $\langle i^*(u,t)i(u,t + \tau)\rangle$ is *not* separable. On the other hand, measurements[31] indicate that the approximation [see Eq. (1.34)]

$$\langle i^*(u,t)i(u,t + \tau)\rangle \cong \hat{\Phi}_I(u)C(\tau) \tag{1.64}$$

may not be unreasonable under typical observing conditions. In Eq. (1.64), $\Phi_I(u)$ is the normalized instantaneous energy spectrum and $C(\tau)$ is the normalized temporal autocorrelation function of the stellar image (some measurements are shown in Fig. 1.6). Substituting Eqs. (1.63) and (1.64) into (1.62) we obtain

$$(\text{SNR})_M \cong \hat{\Phi}_I(u)2\mu\left(\frac{T_e}{\Delta t}\right)^{1/2}\int_0^{\Delta t}\left(1 - \frac{|\tau|}{\Delta t}\right)C(\tau) \, d\tau \ . \tag{1.65}$$

In Fig. 1.9, the SNR is plotted as a function of exposure time for two models of the temporal correlation function $C(\tau)$, Gaussian and negative exponential, each having a $1/e$ correlation time of τ_c; the Gaussian model appears to give a better fit to the experimental data of Fig. 1.6. It can be seen that the overall
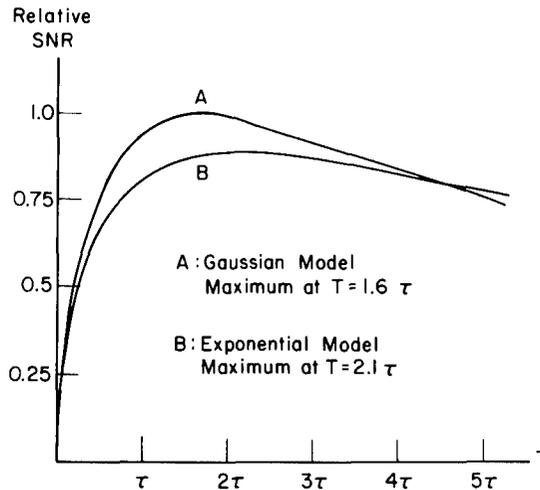
Fig. 1.9   Relative SNR at a point in the power spectrum as a function of the length of the individual short exposures for two models of the time correlation of the image intensity. The overall time of observation is assumed to be constant and it is also assumed that the average number of detected photons per speckle is very much less than one.[62]

SNR is highest for exposure times Δt equal to $1.6\tau_c$ for the Gaussian model and $2.1\tau_c$ for the exponential one. This is somewhat larger than might be expected and certainly much larger than desirable at high light levels where photon noise is negligible. Since the SNR decreases rather slowly for exposure times longer than $\cong 2\tau_c$, we can also conclude that, if there is some doubt as to the value of τ_c, longer rather than shorter exposures should be used.

1.2.3.3   Limiting Magnitude. Labeyrie concluded his original paper[6] on speckle interferometry with the comment that "the technique appears to be limited to objects brighter than $m_v = 7$." It was quickly recognized by Labeyrie and others that, in fact, the faintest objects that can be resolved by this technique are a factor of 10^5 fainter, of apparent visual magnitude $m_v \cong 20$.

Any estimate of the limiting or just-observable magnitude depends on the criterion adopted for "just-observable" as well as on the usual parameters such as detector quantum efficiency, bandwidth, exposure time, and so on. Three examples are given here: an estimate of the complete object energy spectrum, the detection of a binary star, and the measurement of the diameter of a star. In each case we define a factor F to be the product of the exposure time Δt in seconds, [s], the optical bandwidth $\Delta\lambda$ in nanometers, and the quantum efficiency q of the detector,

$$F \equiv \Delta t \Delta \lambda q \ . \tag{1.66}$$

We also use the fact that a source of apparent visual magnitude m_v gives rise to an average number of detected photons per square meter per frame, $\overline{N}_A$, of[63]

$$\overline{N}_A = F 10^{(8 - 0.4 m_v)} \ . \tag{1.67}$$

Estimation of the Object Energy Spectrum.   At low light levels, combination of Eqs. (1.47), (1.60), and (1.67) gives a SNR of

$$(\text{SNR})_M = M^{1/2} \frac{\pi D^2}{4} F 10^{(8 - 0.4 m_v)} \left[0.435 \left(\frac{r_0}{D} \right)^2 \right] \hat{\Phi}_O(u) T_D(u) \ ,$$

which can be rearranged to give[61]

$$m_v = 18.8 + 2.5 \log F - 2.5 \log (\text{SNR})_M + 1.25 \log M$$

$$+ 2.5 \log[\hat{\Phi}_O(u) T_D(u)] + 5 \log r_0 \ . \tag{1.68}$$

For $r_0 = 0.1$ m, $M = 10^5$, $\Delta t = 0.02$ s, $\Delta\lambda = 25$ nm, $q = 0.1$, $\hat{\Phi}_O(u) T_D(u) = 0.2$, and a limiting $(\text{SNR})_M = 5$, Eq. (1.68) predicts a limiting apparent visual magnitude of approximately $m_v = 13.3$, corresponding to approximately 300 detected photons per frame in a 4-m telescope. Note that the limiting magnitude defined in this way is independent of telescope diameter and depends quite strongly on the seeing parameter r_0; in fact, the dependence on r_0 is stronger than Eq. (1.68) indicates since the bandwidth and exposure time both change with r_0 (Ref. 14). The value $m_v = 13.3$ is a conservative estimate of the limiting magnitude for many purposes, since it is based on the criterion that the SNRs have the value 5 at every point in the energy spectrum.
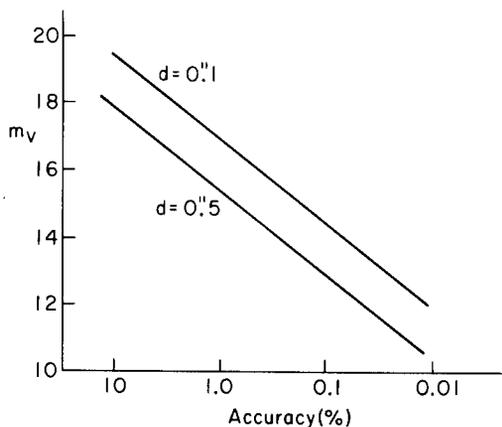
Fig. 1.10 Limiting magnitude m_v as a function of the desired fractional accuracy for a typical set of observing parameters on a 4-m class telescope (observing period: 2000 s).[65]

Detection of Binary Stars.   Using a formula for the SNR based on the autocorrelation approach,[51,61] in which the estimated quantity is the height of the binary star autocorrelation peak above its local background, we can derive the following limiting magnitude for a binary whose components are equally bright:

$$m_v = 17.3 + 2.5 \log F - 2.5 \log(\mathrm{SNR})_M + 1.25 \log M$$

$$+ 2.5 \log D + 2.5 \log r_0 . \tag{1.69}$$

Substituting the same parameters as earlier leads to a limiting magnitude of $m_v = 17.6$, corresponding to approximately five detected photons per frame on average. By increasing the number of independent frames to 10^6 and slightly increasing the exposure time and bandwidth, binaries as faint as $m_v = 20$ should be observable.

The limiting magnitude predicted by Eq. (1.69) has been effectively achieved by Hege et al.[64] in their measurement of the 16.2 magnitude component of the triple quasar PG 1115 + 08 using approximately 20,000 independent frames.

Estimation of Object Diameter.   Walker[49] has made a comprehensive study of the accuracy with which the diameter of an object can be estimated by speckle interferometry, assuming a known limb darkening profile of the star. His results are summarized in Fig. 1.10 for a collection of observing parameters that are similar (but not identical) to the previous two cases. For 1% statistical error in a diameter whose value is 0.″5, the limiting magnitude is approximately $m_v = 16$. Of course, other deterministic effects such as those due to atmospheric calibration are not included in this or previous cases.

1.2.3.4   Space-Time Speckle Interferometry.

In the analysis of the optimum exposure time in Sec. 1.2.3.2, we found that exposures as long as twice the temporal correlation time of the image could be optimum from the point of view of the SNR. Such long exposure times result in attenuation of the high angular frequency components in the measured energy spectrum, and those

remain uncorrected. Another drawback of the straightforward speckle method is that no use is made of the fact that photons detected at the end of one exposure are associated with essentially the same classical intensity as those detected at the beginning of the next exposure; thus there is a potential loss of information.

Space-time speckle interferometry[62,66] is an extension of speckle interferometry that includes correlations in the time domain as well as in the spatial or angular domain. In one such scheme, the temporal cross-energy spectrum $\langle i^*(u,v,t)i(u,v,t + \tau)\rangle$ is estimated and used to find an estimate of $\hat{\Phi}_I(u,v)$ that is not biased by the effects of a finite exposure time. However, the SNR of this technique does not appear to be any higher than that associated with the optimum exposure time method.[62] It does not appear to be worthwhile implementing space-time speckle interferometry unless other benefits can be found (such as obtaining object maps[66]).

1.2.4   Reconstruction of the Object Intensity

The fundamental equation of speckle interferometry relates the average energy spectrum of the image $\Phi_I(u,v)$ to that of the object $\Phi_O(u,v)$,

$$\Phi_I(u,v) = \Phi_O(u,v)\mathcal{T}(u,v) \ ,$$

where $\mathcal{T}(u,v)$ is the speckle transfer function. Under favorable conditions this equation can be inverted to yield an estimate of the object energy spectrum

$$\Phi_O(u,v) \equiv |o(u,v)|^2$$

$$= \left| \int\int_{-\infty}^{\infty} O(\alpha,\beta) \exp[-2\pi i(\alpha u + \beta v)] \, d\alpha \, d\beta \right|^2 \ , \tag{1.70}$$

where $O(\alpha,\beta)$ is the angular distribution of object intensity and $o(u,v)$ is its Fourier transform. It should be noted that, by the van Cittert-Zernike theorem,[2] $o(u,v)$ is a *spatial coherence function* (strictly, the mutual intensity) and $|o(u,v)|$ is often called a *visibility function*.

It is impossible, in general, to calculate a unique object intensity $O(\alpha,\beta)$ from a knowledge of only its energy spectrum $\Phi_O(u,v)$; this simple fact cannot be stressed too strongly. In some special cases, unique reconstruction of $O(\alpha,\beta)$ is possible; in a second set of special cases, unique reconstructions can be formed almost always; and in a third set of special cases, additional information is available that enables a unique solution to be found.

The object energy spectrum $\Phi_O(u,v)$ contains no obvious information about the phase of the Fourier transform of $O(\alpha,\beta)$ and for this reason the problem of reconstructing the object intensity from $\Phi_O(u,v)$ is referred to as the *phase problem*. Phase problems arise in many branches of physics—scattering, x-ray diffraction, coherence theory, and microscopy—and a detailed review is beyond the scope of this chapter (see Refs. 67 and 68). Our review is limited strictly to the phase problem as it occurs in the measurement of angular coherence functions by stellar speckle interferometry; short reviews of this may be found in Refs. 54 and 69 through 72 and a comprehensive review was given by

Bates.[73] It is interesting to note that some of the earliest work on the phase problem by Lord Rayleigh[74] and, in the modern era, by Wolf[75] was also concerned with coherence theory.

In Sec. 1.2.4.1 we discuss the basic reason for the ambiguity of the phase problem. The next two sections deal with attempts at object reconstruction from the energy spectrum only, while in Secs. 1.2.4.4 through 1.2.4.8 we describe other methods that incorporate information in addition to the energy spectrum. The subject is summarized in Sec. 1.2.4.9. The review is limited to the speckle method of stellar interferometry; in this regard it should be noted that there is increasing evidence[76–79] that other methods of stellar interferometry are probably more appropriate for object reconstruction.

1.2.4.1   Ambiguity of the Phase Problem.

An essentially theoretical restriction in the phase problem, which is always satisfied in practice, is that the object intensity $O(\alpha,\beta)$ has a finite angular extent with support (2a, 2b); thus $o(u,v)$ is the *finite* Fourier transform,

$$o(u,v) = \int_{-a}^{a} \int_{-b}^{b} O(\alpha,\beta) \exp[-2\pi i(u\alpha + v\beta)] \, d\alpha \, d\beta \; . \qquad (1.71)$$

It can be shown that the analytical continuation of $o(u,v)$ to the complex plane, $o(z_1,z_2)$ where z_1 and z_2 are complex variables is an entire function of exponential type. Such functions are completely specified by their (complex) zeros. The zeros provide a unifying concept for the study of all phase retrieval methods; their importance in interferometry was discussed by Bates[80] and in a more general context by Ross and colleagues.[81–83] Although the zeros are the unifying concept, they are not necessarily of practical value in computer-based algorithms due to the complexity of determining their locations.

Before discussing the reason for the ambiguity of the phase problem, we should note that certain phase ambiguities do not affect the form of the object intensity and are ignored in the following analysis. Defining the phase of $o(u,v)$ as phase $\{o(u,v)\}$, we are not concerned with the following variants:

$$\text{phase } \{o(u,v)\} + \phi, \text{ where } \phi \text{ is a constant}, \qquad (1.72a)$$

phase $\{o(u,v)\} + 2\pi(u\alpha_1 + v\beta_1)$,

  where (α_1,β_1) is a constant vector, $\qquad (1.72b)$

$- \text{ phase } \{o(u,v)\}$. $\qquad (1.72c)$

The addition of a constant phase [Eq. (1.72a)] does not alter the object intensity $O(\alpha,\beta)$; the second variant [Eq. (1.72b)] leads to a shifted object $O(\alpha + \alpha_1, \beta + \beta_1)$; the third case [Eq. (1.72c)] gives $O(-\alpha, -\beta)$, which is a 180-deg rotated version of the object. In the following discussion, these trivial ambiguities are ignored.

Our approach to describing the phase problem is, following Bruck and Sodin,[84] to represent the object by a finite number of samples, equally spaced (for simplicity) by Δ on a grid of $(N + 1)$ by $(M + 1)$ points. Defining new complex variables w_1 and w_2

$$w_1 = \exp(-2\pi i z_1 \Delta), \qquad w_2 = \exp(-2\pi i z_2 \Delta) \ , \tag{1.73}$$

the Fourier transform $o(w_1, w_2)$ can be written as a finite polynomial in w_1 and w_2

$$o(w_1, w_2) = w_1^{-a/\Delta} w_2^{-b/\Delta} \sum_{n=0}^{N} \sum_{m=0}^{M} O(n\Delta - a, \, m\Delta - b) w_1^n w_2^m \ .$$

The terms $w_1^{-a/\Delta}$ and $w_2^{-b/\Delta}$ merely define the (α, β) origin; ignoring these, and simplifying the notation we write

$$o(w_1, w_2) = \sum_{n=0}^{N} \sum_{m=0}^{M} O_{nm} w_1^n w_2^m \ . \tag{1.74a}$$

The most important feature of Eq.(1.74a) is that the (discrete) Fourier transform of the object intensity can be written as a finite polynomial in the complex variables w_1 and w_2, the coefficients of the polynomial being the sampled values of the object intensity. In this approach to the phase problem, the mathematics of polynomials is important. Note, however, that this approach is less general than required by the original problem, which was for continuous, not discrete, object functions.

Consider now the one-dimensional case:

$$o(w_1) = \sum_{n=0}^{N} O_n w_1^n \ . \tag{1.74b}$$

A one-dimensional polynomial can always be factorized, or reduced, into prime factors.

$$o(w_1) = C \prod_{j=1}^{N} (w_1 - w_{1,j}) \ , \tag{1.75}$$

where C is a constant and $w_{1,j}$ are the roots or zeros. The N zeros and the constant C completely determine the Fourier transform $o(w_1)$ and hence the object O_n. If the object is real, as in the present case, the zeros lie on the unit circle or in complex conjugate pairs around the unit circle and only $N/2$ zero locations are required to specify the object; positivity requires that no zeros lie on the positive real w_1 axis. Figures 1.11(a) and (b) illustrate these results.

In a similar manner, we can represent the energy spectrum $\Phi_O(u)$ as an analytical function $o(z_1)o^*(z_2^*)$. The conjugate function to $o(w_1)$ is just $o(1/w_1)$ so that $\Phi_O(w_1)$ can be written as a polynomial of degree $2N$:

$$\Phi_O(w_1) = C^2 \prod_{j=1}^{N} (w_1 - w_{1,j})(w_1 - 1/w_{1,j}^*) \ . \tag{1.76}$$

That is, the complex zeros of $\Phi_O(w_1)$ consist of the original N zeros of the object transform plus their inverses. This is illustrated in Fig. 1.11(c).
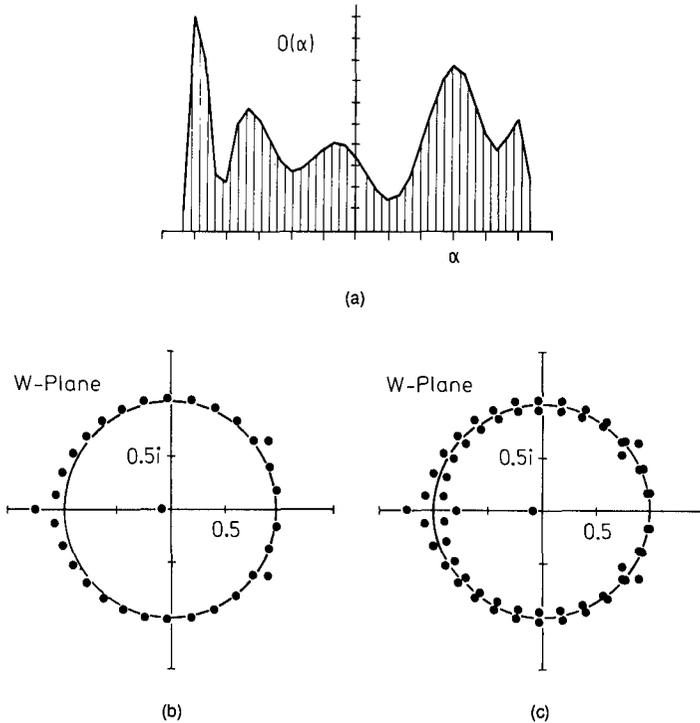
Fig. 1.11 (a) A real positive object, (b) zeros of its Fourier transform, and (c) zeros of its energy spectrum. (Courtesy of B. J. Brames)

Thus, the essence of the phase problem is that, without some basis for choosing between the correct zero and its inverse, we could construct $2^{N/2}$ equally valid sets of N zeros each representing a real, possibly positive, object. In the one-dimensional case, there is no unique solution to the phase problem, in either a theoretical or practical sense; additional information is required to find the object intensity.

Consider now the two-dimensional case, where the Fourier transform of the object intensity can be written as a polynomial in two complex variables,

$$o(w_1, w_2) = \sum_{n=0}^{N} \sum_{m=0}^{M} O_{nm} w_1^n w_2^m \ .$$

Napier and Bates[85] were the first to find that a unique solution to the phase problem was more likely to occur in this case. In one dimension, ambiguity resulted from the factorizability of the polynomial of Eq. (1.74b); in two dimensions, as shown by Bruck and Sodin,[84] ambiguity may also exist if the two variable polynomial of Eq. (1.74a) is factorizable (or reducible) and the degree of ambiguity is determined by the number of non-self-conjugate irreducible factors. However, there is a very small probability that any two-dimensional polynomial is reducible; in fact, reducible polynomials in two dimensions are

a set of measure zero.[86] Thus one is tempted to assume that the two-dimensional phase problem has a unique solution "almost always."

The uniqueness of the two-dimensional phase problem is the subject of much current research. The results of applying the algorithms to be described in Sec. 1.2.4.3 strongly suggest that effectively unique solutions may exist for certain objects, although of course it is always possible to produce counter-examples.[87,88] Fiddy et al.[89] and Fienup[90] have used Eisenstein's irreducibility theorem to define one particular class of objects for which a unique solution is guaranteed.

There are three basic approaches to solving the phase problem in stellar speckle interferometry. In the first, it is assumed that something about the object is known. For example, for a symmetric object intensity

$$O(\alpha,\beta) = O(-\alpha,-\beta) \ , \tag{1.77}$$

the Fourier transform $o(u,v)$ is purely real and continuity arguments enable it to be found from $|o(u,v)|$; a rotationally symmetric object is included in this category. Speckle holography, discussed in the next subsection, also assumes that the object has a known property. In the second approach, one assumes that the two-dimensional phase problem is almost unique and seeks an algorithm to recover the object intensity from the modulus information alone. In the third approach, additional information is extracted from the speckle images in a number of different ways (see Secs. 1.2.4.4 through 1.2.4.7).

1.2.4.2   Speckle Holography.   The technique of speckle holography, in its original and most elementary form,[91,92] relies on the presence of a reference object, preferably a point source. Let the object field be written as the sum of a point centered at the origin and the object under investigation $O_1(\alpha,\beta)$ centered at (α_1,β_1),

$$O(\alpha,\beta) = \delta(\alpha)\delta(\beta) + O_1(\alpha - \alpha_1,\beta - \beta_1) \ . \tag{1.78}$$

The spatial autocorrelation of Eq. (1.78) consists of four terms

$$C_O(\alpha,\beta) = \int\int_{-\infty}^{\infty} \delta(\alpha')\delta(\beta')\delta(\alpha' + \alpha)\delta(\beta' + \beta) \, d\alpha' \, d\beta'$$

$$+ \int\int_{\infty}^{\infty} O_1(\alpha', \beta')O_1(\alpha' + \alpha, \beta' + \beta) \, d\alpha' \, d\beta'$$

$$+ O_1(\alpha - \alpha_1, \beta - \beta_1) + O_1(-\alpha - \alpha_1, -\beta - \beta_1) \ . \tag{1.79}$$

The first two terms are located in the region of the origin, the third is the object centered at (α_1,β_1), and the fourth term is a 180-deg rotation of the object centered at $(-\alpha_1,-\beta_1)$. Provided that $\alpha_1 > 3a/2$ and $\beta_1 > 3b/2$, where
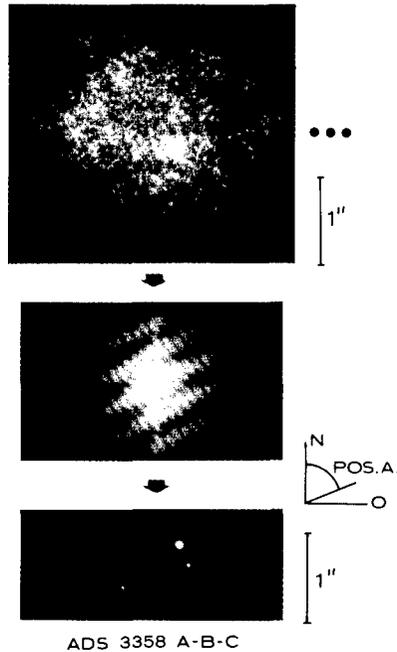
Fig. 1.12 Speckle holography of ADS 3358 (Ref. 94).

the object extent is (a,b), the third and fourth terms are separated in angle from the first two and a reconstruction of the object is obtained (with the 180-deg rotation ambiguity). Weigelt[93-95] has demonstrated that this is a useful astronomical technique and Fig. 1.12 shows an example of the reconstruction of a triple star using speckle holography. The extent of the atmospheric isoplanatic angle is clearly important in speckle holography (Sec. 1.2.2.7).

If the reference point is not separated by the *holographic distance* then, in general, the object intensity cannot be reconstructed unambiguously unless further information is available. For example, Liu and Lohmann[96] suggested using the long-exposure image as a mask, and Baldwin and Warner[97] used the knowledge that one star is brighter than the others to unravel the object (star clusters) from the autocorrelation function. Indeed, if the object consists of a discrete set of points and no vector separation between points occurs more than once (i.e., nonredundant spacings), then a unique solution to the problem exists.[98,99] In another special case described by Bruck and Sodin,[83] a one-dimensional object can be reconstructed uniquely provided that the reference point is not in line with the object in the two-dimensional plane. The irreducibility criterion described by Fiddy et al.[89] also involves the use of reference points less than the usual holographic distance. In practice, speckle holography is rarely applicable to astronomical problems, since very few interesting objects have bright sources conveniently located at a small enough separation to be within the atmospheric isoplanatic angle.

Weigelt[100,101] has suggested a technique called *speckle masking* that is related to holography in the sense that the speckle short-exposure images are preprocessed to yield an approximation to the instantaneous point spread function. Details are given in Sec. 1.2.4.8.

1.2.4.3 Modulus-Only Algorithms.

In this approach to object reconstruction in stellar speckle interferometry, it is implicitly assumed that the two-dimensional problem *does* have a unique solution. Three algorithms that attempt to recover this solution are described here. Any result produced by these algorithms is therefore subject to two uncertainties: (1) Did a unique solution to the phase problem exist, even in principle? (2) If it did exist, did the algorithm converge to this solution? Strictly speaking, uniqueness of the solution to the two-dimensional phase problem is not guaranteed and none of the algorithms described here has been shown to always converge to the unique solution when one is known *a priori* to exist. On the other hand, the overwhelming proportion of experimental evidence suggests that, for simple objects, some of these methods are successful in reconstructing object maps.

Iterative Algorithm. Fienup has suggested a number of iterative algorithms[102–106] for computing the object intensity from a knowledge of only the modulus of its Fourier transform and an estimate of the support of the object. Two possible schemes are shown in Fig. 1.13. The first scheme, called the *error reduction method* because the mean square error between iterations always decreases,[106] is a generalized form of the Gerchberg-Saxton algorithm.[107] Starting with an estimate of the object intensity at the k'th iteration $\hat{O}_k(\alpha,\beta)$, the transform $\hat{o}_k(u,v)$ is calculated. The modulus of this transform is replaced by the given modulus, forming a new estimate $\hat{o}'_k(u,v)$ that satisfies the constraints of the problem in the Fourier transform domain. This is inverse-transformed to give a new estimate of the object $\hat{O}'_k(\alpha,\beta)$, which is set to zero in the region where the object is known to be zero and set equal to zero where negative object values exist, thus forming a new estimate $\hat{O}_{k+1}(\alpha,\beta)$, which is the starting point for the next cycle. In practice, the error reduction algorithm converges very slowly and it is generally most useful when applied with one of the input/output algorithms.

The second scheme is shown in Fig. 1.13(b) and is called the *input/output algorithm*. The only difference between this and the error reduction scheme lies in how the next starting input $\hat{O}_{k+1}(\alpha,\beta)$ is derived from the previous output estimate $\hat{O}'_k(\alpha,\beta)$ and input $\hat{O}_k(\alpha,\beta)$. To a first-order approximation, a small change in the input gives a small change in the output proportional to that in the input (plus nonlinear terms); thus, by changing the input it should be possible to drive the output in the desired direction. The most satisfactory version of this scheme, called the *hybrid input/output algorithm* is

$$\hat{O}_{k+1}(\alpha_1\beta) = \begin{cases} \hat{O}'_k(\alpha,\beta) & \text{when object constraints satisfied} \\ \hat{O}_k(\alpha,\beta) - \gamma\hat{O}'_k(\alpha,\beta) & \text{when not satisfied,} \end{cases}$$

$$(1.80)$$

where γ is a parameter, typically of the order of unity.
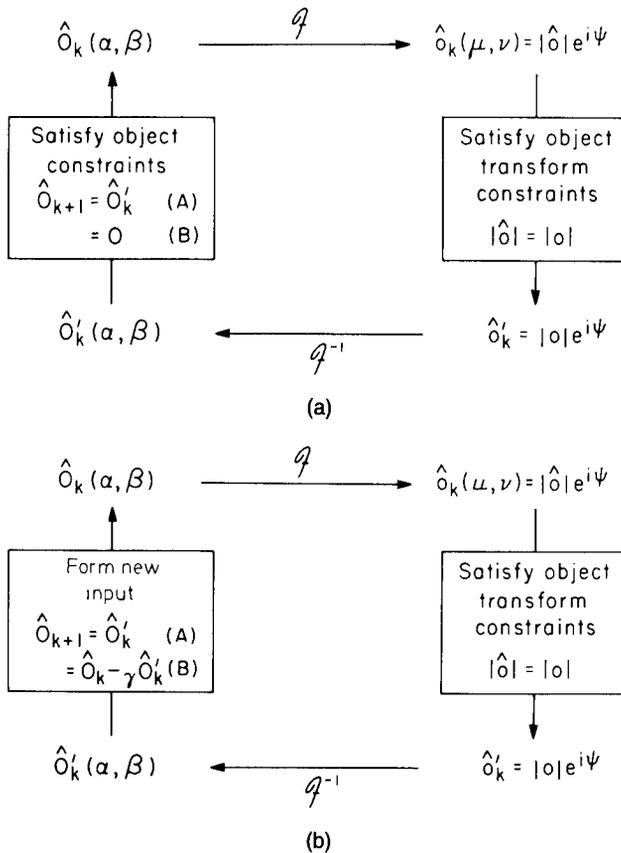
Fig. 1.13 The iterative algorithms for solving the phase problem: (a) error reduction in object space and (b) hybrid input/output. (A) is where object constraints are satisfied and (B) is where they are not satisfied.[106]

A discussion of the relative merits of different iterative algorithms is given in Ref. 106. At the present time, these algorithms are still rather ad hoc and their success appears to depend to some extent on the skill of the programmer. Figure 1.14 shows some results obtained by Fienup. These algorithms tend to recover the object intensity successfully for simple, but nonsymmetric, objects; the shape of the support of the object also appears to affect the success of the iterative method. It should be stressed that this (and other) algorithms can fail to converge to the correct solution for complicated objects.

Phase-Closure Algorithms. Bates and coworkers[108–110] have suggested an algorithm that, in its original form, may be useful as a starting point for the Fienup algorithm,[111–113] and in a future improved form may be valuable on its own. Consider an array of N by M values of the Fourier transform of an object (for a real object of size N by N, $M = N/2 + 1$). The aim of this algorithm is to calculate the phases of each point θ_{ij}, allowing any one point (usually
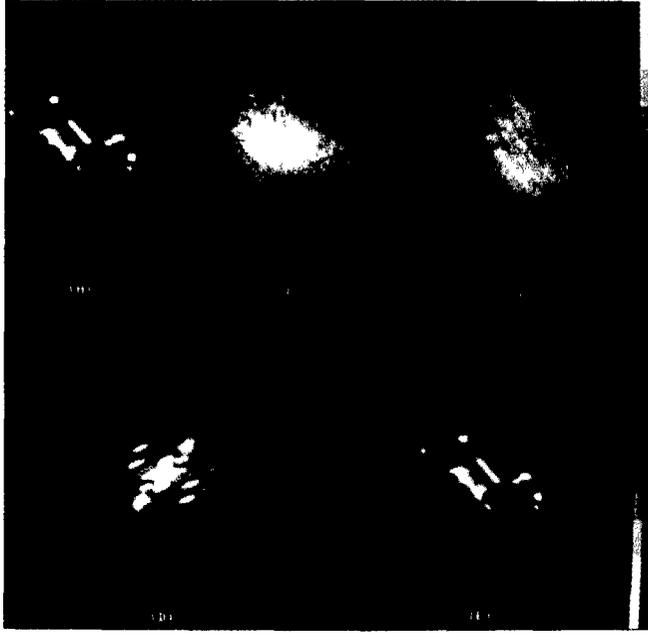
Fig. 1.14 (A) Original object; (B), (C) examples of simulated degraded images; (D) Fourier modulus estimate computed from degraded images; and (E) image reconstructed using iterative algorithm.[106]

the origin) to be set to zero. Bates and coworkers suggested the following two-step procedure:

1. Estimate the magnitude of the $(N - 1)$ by M phase differences along the u axis, $|\theta_{i+1,j} - \theta_{i,j}|$ and the N by $(M - 1)$ v phase differences $|\theta_{i,j+1} - \theta_{i,j}|$.
2. Compute the N by M phases from the magnitudes of these $(2NM - N - M) \cong 2NM$ phase differences.

Let us assume, for the moment, that step 1 is possible and see how phase closure might be used to determine the phases. Consider the rectangle comprising the first four points (0,0), (1,0), (1,1), and (0,1) and assume that the magnitudes of the four phase differences are known:

$$
\begin{aligned}
|\theta_{1,0} - \theta_{0,0}| &= \psi_1 \ , \\
|\theta_{1,1} - \theta_{1,0}| &= \psi_2 \ , \\
|\theta_{1,1} - \theta_{0,1}| &= \psi_3 \ , \\
|\theta_{0,1} - \theta_{0,0}| &= \psi_4 \ .
\end{aligned}
\tag{1.81}
$$

Clearly, we can set

$$
\theta_{0,0} = 0
\tag{1.82a}
$$

and

$$\theta_{1,0} = +\psi_1 \ . \tag{1.82b}$$

(If in fact $\theta_{1,0} = -\psi_1$, the object reconstruction will be rotated by 180 deg.) Proceeding around the rectangle anticlockwise,

$$\theta_{1,1} = \psi_1 \pm \psi_2 \tag{1.82c}$$

and this leads to four possible values of $\theta_{0,1}$,

$$\theta_{0,1} = \psi_1 \pm \psi_2 \pm \psi_3 \ . \tag{1.82d}$$

On the other hand, going directly from (0,0) to (0,1) yields

$$\theta_{0,1} = \pm \psi_4 \ . \tag{1.82e}$$

Bates[108] argued that only one of the four solutions [Eq. (1.82d)] will equal one of the two solutions [Eq. (1.82e)], thus determining the phases at each of the four points. If this is the case, then this procedure could be repeated for all points in the Fourier plane and the object intensity could be found by inverse Fourier transformation. Since the number of phase difference magnitudes is roughly twice the number of phases, it may be possible to use the methods mentioned in Sec. 1.2.4.7 for improving the phase estimates.

   Even if step 2 works, it is still necessary to find the magnitudes of phase differences from step 1. These can be estimated by oversampling the modulus in a scheme in which the Shannon interpolation formula is replaced by two-point interpolation[108]; this provides only a crude estimate of the phase differences (for example, a large proportion have to be set equal to 0 or π) and requires improvement for reliable object restoration by itself. Combined with a modified Fienup algorithm that incorporates a preprocessing step to remove the strong central lobe in the Fourier plane, this technique has been shown[112,113] to produce excellent reconstructions of simple objects.

Maximum Entropy Algorithm.   In general terms, the maximum entropy method reconstructs the smoothest object intensity distribution consistent with the available data. It was first suggested for use with phaseless data by Gull and Daniell[114]; as with the other algorithms, there is, of course, no way that the maximum entropy algorithm can resolve any inherent ambiguities.[115] If there are ambiguities, this method restores the smoothest object map.

1.2.4.4   Use of Exponential Filters.

The ambiguity of the phase problem arises because the $2N$ zeros of the object energy spectrum consist of the N zeros associated with the Fourier transform of the object, plus their inverses. Given only the $2N$ zeros of the energy spectrum, it is impossible, in general, to select the correct zeros from each zero pair. By making a second measurement of the energy spectrum of a modified object intensity distribution [the original $O(\alpha,\beta)$ multiplied by $\exp(-2\pi\alpha a)$, where a is a constant], it is possible to recover unambiguously the correct N zeros and hence the object intensity itself. This was first suggested by Walker[116] and Wood et al.[117]
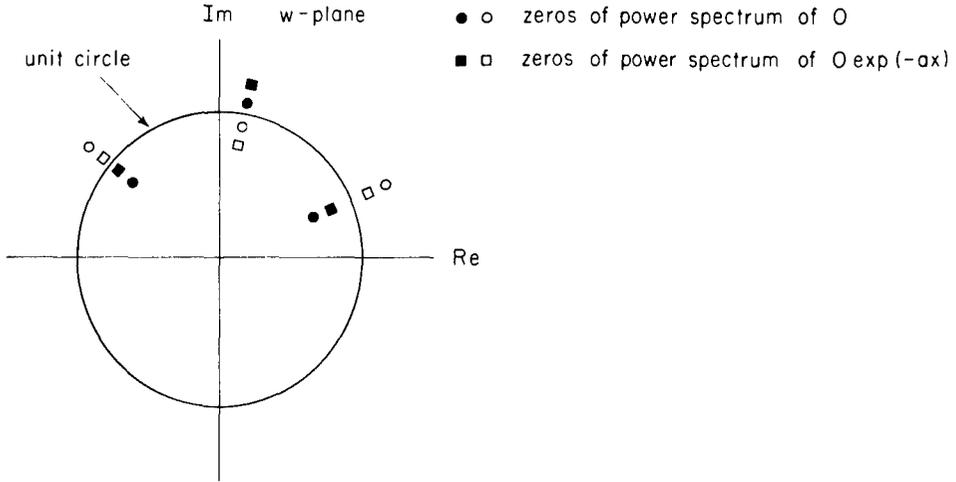
Fig. 1.15   The effect of exponential filtering on zero locations.

The basic principle of the method is shown in Fig. 1.15, where, for illustration, there are only three sets of zeros. The zeros corresponding to the original object are shown as solid circles ● and their inverses as ○; given only the object energy spectrum it is impossible to determine which is the "correct" one. When the object is multiplied by $\exp(-2\pi\alpha a)$, $a > 0$, the zeros in terms of the z variable move from z_j to $z_j - ia$, and in terms of the w variable, $w = \exp(-2\pi iz\Delta)$, from w_j to $w_j \exp(2\pi a\Delta)$. That is, the correct zeros all move radially outward by a constant factor, as shown in Fig. 1.15 (●→ ■). The energy spectrum of the modified object contains both these zeros (■) and their inverses (□); given both pairs of zeros (●, ○, ■, and □) the correct zero (●) can always be located. Although our description has been in terms of one dimension, the uniqueness of the solution also applies to the two-dimensional case.

In astronomy, it is, of course, impossible to place an exponential filter over the object! Walker[118] showed that this is not necessary and that the exponential filter may be placed in the image plane. Denote the instantaneous image intensity by $I(\alpha,\beta)$ and the exponential filter transmittance by $G(\alpha,\beta)$. The energy spectra of the image intensity and the modified image intensity $[I(\alpha,\beta)\cdot G(\alpha,\beta)]$ of Eq. (1.5) are

$$\Phi_I(u,v) = |o(u,v)|^2 \langle |T(u,v)|^2 \rangle$$

and

$$\Phi'_I(u,v) = \langle |i(u,v)\otimes g(u,v)|^2 \rangle$$

$$= \langle |\{o(u,v)T(u,v)\}\otimes g(u,v)|^2 \rangle \ , \tag{1.83}$$

where $\otimes$ denotes the convolution and the other symbols are defined in Sec. 1.2.2.1. Provided that
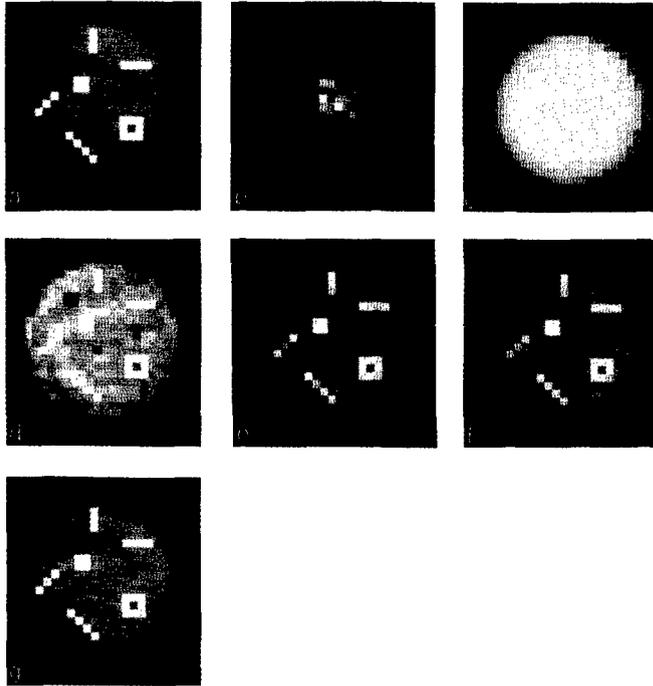
Fig. 1.16 (a) Test object; (b) a short-exposure point spread function; (c) a short-exposure image; (d)-(g) reconstructions using, respectively, 10, 25, 50, and 100 simulated point and object exposures.[119]

$$G(\alpha_1 + \alpha_2, \beta_1 + \beta_2) \equiv G(\alpha_1,\beta_1)G(\alpha_2,\beta_2) \ , \tag{1.84}$$

which is satisfied by the real exponential function, the convolution of Eq. (1.83) simplifies to yield

$$\Phi_I'(u,v) = |o'(u,v)|^2 \langle |T'(u,v)|^2 \rangle \ , \tag{1.85}$$

where

$$o'(u,v) = o(u,v) \circledast g(u,v) \ ,$$

$$T'(u,v) = T(u,v) \circledast g(u,v) \ .$$

Assuming that the forms of the two transfer functions $\langle |T(u,v)|^2 \rangle$ and $\langle |T'(u,v)|^2 \rangle$ can be found (using a reference star, for example), we can find the energy spectra of the object and of the modified object, which are sufficient data for a unique solution to the phase problem.

Having shown that a unique object reconstruction can be found from $|o(u,v)|^2$ and $|o'(u,v)|^2$, there remains the problem of finding a practical two-dimensional algorithm that converges to this unique solution. Walker[118,119] has used an extended version of the Fienup algorithm that includes both sets of Fourier constraints. Figure 1.16 shows an example of reconstructions obtained by Walker

in a computer simulation that used this algorithm. Note that this proposed method of object reconstruction uses only a single set of data for the object and for the reference, because the exponential filter can be applied numerically on the raw data.

1.2.4.5 Shift and Add.

The short-exposure speckle images shown in Fig. 1.1 are of an unresolvable star in the upper row and α-Orionis, or Betelgeuse, which is a red giant star in the lower row. In simplistic terms, each "speckle" in both sets of images may be regarded as an "image"; for the upper row, it is an image of a point source and for the lower row it is an image of α-Orionis. Such reasoning led Harvey and coworkers[120,121] to obtain the first diffraction-limited map of a star other than our own sun.

In the original method, a few bright speckles are selected from each exposure and superimposed with the aid of a digital microdensitometer and computer. Figure 1.17 shows the result of this process for a point object (a) and α-Orionis in the continuum (b) and TiO absorption band (c); clearly the giant star is

(a) (b)

0.05 "

(c) (d)
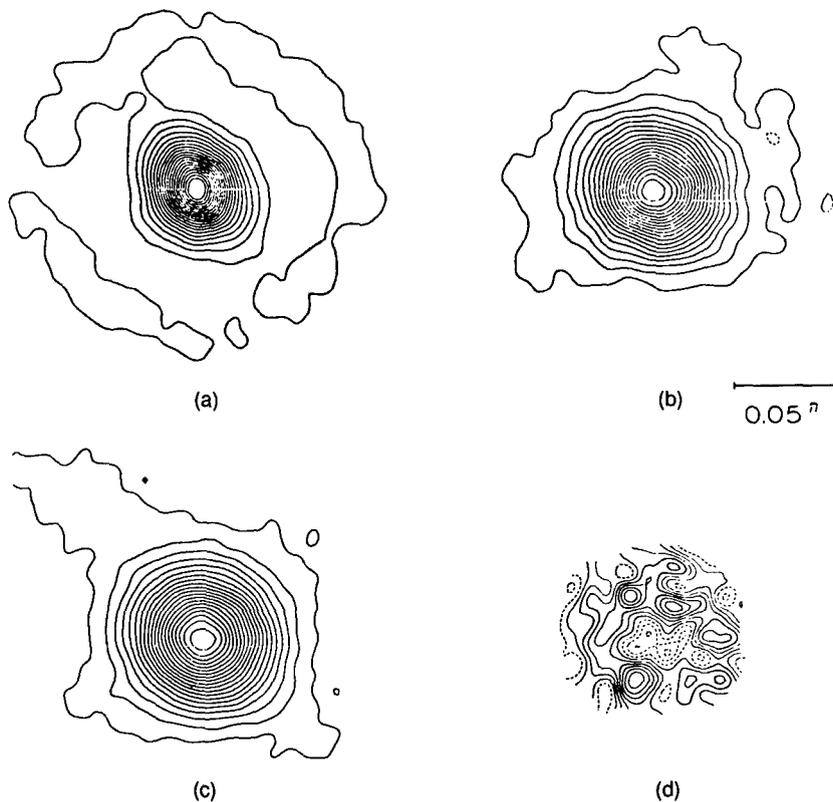
Fig. 1.17 Diffraction-limited images computed from short-exposure photographs by Lynds et al.[120]: (a) unresolved star (γ-Ori), (b) α-Ori or Betelgeuse, in the continuum, (c) α-Ori in the TiO band, and (d) the difference between (b) − (c). The contour levels are 5% of the peak intensity (a) − (c); in (d) the interval is 2%, with the broken curve indicating that the continuum is brighter.

resolved and the difference $|(b - c)|$ indicates possible temperature variation over the surface of the star. McDonnell and Bates[122] have applied super-resolution techniques to produce an enhanced image of Betelgeuse from these data.

This approach to forming object maps has been extended by Bates and Cady[123,124] in a technique they call *shift and add*. Let (α_j, β_j) denote the coordinates of the center of the brightest speckle in the j'th image. Each image is shifted such that (α_j, β_j) is at the origin and then added to all other similarly shifted images, giving the result

$$R(\alpha, \beta) = \frac{1}{N} \sum_{i=1}^{N} I_j(\alpha - \alpha_j, \beta - \beta_j) \ . \tag{1.86}$$

This process is carried out for both the object under study and a reference star; the image of the object is deconvolved using that of the reference and an algorithm[125] such as CLEAN. A theoretical study[126] has confirmed that diffraction-limited information is preserved in the shift-and-add method. Excellent reviews of shift-and-add techniques and their variations are found in Bates and McDonnell[127] and Hege.[128]

1.2.4.6 Phase Averaging. In the technique of speckle interferometry, the Fourier transforms of the instantaneous image intensity and the object intensity are related by

$$i(u,v) = o(u,v)T(u,v) \ , \tag{1.87}$$

where $T(u,v)$ is the instantaneous transfer function. The quantities $\langle |i(u,v)|^2 \rangle$ and $\langle |T(u,v)|^2 \rangle$ are measured and an estimate of the object energy spectrum $|o(u,v)|^2$ is obtained. Taking the logarithm of Eq. (1.87) we obtain

$$\text{phase}\{i(u,v)\} = \text{phase}\{o(u,v)\} + \text{phase}\{T(u,v)\} \tag{1.88}$$

and, taking the average

$$\langle \text{phase}\{i(u,v)\} \rangle = \text{phase}\{o(u,v)\} + \langle \text{phase}\{T(u,v)\} \rangle \ , \tag{1.89}$$

where, in all cases, the phase is the value in the interval $-\infty$ to ∞. Thus, provided that $\langle \text{phase}\{T(u,v)\} \rangle$ is known (or zero), the phase of the object transform can be obtained from the average phase of the image transforms; this method was first suggested by McGlamery.[129]

Using arguments based on the central limit theorem, it is not difficult to show that, for $D \gg r_0$ and angular frequencies

$$r_0/\lambda < (u,v) < (D - r_0)/\lambda \ ,$$

the quantity $T(u,v)$ is a circular complex Gaussian random process. It follows that

$$\langle \text{phase}\{T(u,v)\} \rangle = 0 \ ,$$

and that the phase $\{T(u,v)\}$ folded into the primary interval $-\pi$ to π is statistically uniformly distributed.

The crucial step in implementing the phase-averaging method is therefore the determination of the *unwrapped* phase (i.e., that in the interval $-\infty$ to ∞) from the phase in the primary interval $-\pi$ to π. In principle, this may be done by assuming continuity of the phase and following it out from the origin where it can be assumed to be zero. This procedure is subject to error when the modulus $|i(u,v)|$ is small. O'Donnell[28] has shown that the rms absolute error σ in the unwrapped phase is given approximately by

$$\sigma \simeq \frac{1}{(2\overline{N})^{1/2}|\hat{i}(u,v)|} , \tag{1.90}$$

where $\overline{N}$ is the average number of detected photons per frame and $\hat{i}(u,v)$ is the Fourier transform of the instantaneous image intensity normalized to unity at the origin. Clearly, a small value of $|\hat{i}(u,v)|$ leads to a large error. For example, for a point object $[\hat{\Phi}_O(u,v) = 1]$ at an intermediate frequency $[T_D(u,v) \simeq 0.5]$ and a large telescope $(D/r_0 \simeq 40)$, an *average* value of $|\hat{i}(u,v)|$ is of the order of 10^{-2}, implying $\overline{N} > 8 \times 10^4$ detected photons per frame for a phase error of less than 0.25 rad.

Despite this analysis, computer simulations of the phase-averaging method have shown some promise,[130,131] particularly for providing a starting point to the Fienup algorithm. Other algorithms for phase unwrapping have been suggested by Tribolet[132] and Swan[133]; in the latter, the average phase is calculated without explicit unwrapping. Finally, Mertz[66] has suggested following the phases of the angular frequency components in time in order to find their average value; the error has not yet been evaluated for this approach.

1.2.4.7   Knox-Thompson Method.

In this method, first suggested by Knox and Thompson,[134,135] the cross-energy spectrum of the image intensity is computed; following the notation of Sec. 1.2.2.1,

$$\langle i(u',v')i^*(u'',v'')\rangle = o(u',v')o^*(u'',v'')\langle T(u',v')T^*(u'',v'')\rangle . \tag{1.91}$$

Taking logarithms of each side and equating imaginary parts, we find that

$$\text{phase}\{\langle i(u,v)i^*(u + \Delta u, v + \Delta v)\rangle\}$$

$$= \text{phase}\{o(u,v)\} - \text{phase}\{o(u + \Delta u, v + \Delta v)\}$$

$$+ \text{phase}\{\langle T(u,v)T^*(u + \Delta u, v + \Delta v)\rangle\} , \tag{1.92}$$

where we have made the substitutions

$$\Delta u = u'' - u' \quad \text{and} \quad \Delta v = v'' - v' \quad \text{in Eq. (1.91).}$$

Thus, provided that
1. $\langle T(u,v)T^*(u + \Delta u, v + \Delta v)\rangle \neq 0$ and
2. phase $\{\langle T(u,v)T^*(u + \Delta u, v + \Delta v)\rangle\}$ is either known or zero,

it is possible to find *phase differences* in the object spectrum. This information is then used to find the phase of the object spectrum and hence the object intensity (if the energy spectrum is known). In the following we show that item 1 is satisfied when $(u,v) < r_0/\lambda$ and that phase $\{\langle T(u,v)T^*(u + \Delta u, v + \Delta v)\rangle\}$ is approximately zero; we then discuss how the phase difference information can be used to restore the actual phases.

To evaluate the quantity $\langle T(u',v')T^*(u'',v'')\rangle$, we use a similar approach to that given in Sec. 1.2.2.3 to evaluate the approximate speckle transfer function. In particular, we assume that the complex amplitude of the wave in the telescope pupil from a point source is a circular complex Gaussian process. Instead of Eq. (1.23) we now have the following expression for $\langle T(u',v')T^*(u'',v'')\rangle$:

$$\langle T(u',v')T^*(u'',v'')\rangle = T_0(u',v')T_0^*(u'',v'')T_s(u',v')T_s^*(u'',v'')$$

$$+ \mathcal{P}^{-2}\iiiint\limits_{-\infty}^{\infty} T_s\left(\frac{\Delta\xi}{\lambda}, \frac{\Delta\eta}{\lambda}\right)T_s^*\left(\frac{\Delta\xi}{\lambda} + \Delta u, \frac{\Delta\eta}{\lambda} + \Delta v\right)$$

$$\times\ H_0(\xi_1,\eta_1)H_0^*(\xi_1 + \lambda u', \eta_1 + \lambda v')H_0^*(\xi_2,\eta_2)$$

$$\times\ H_0(\xi_2 + \lambda u'', \eta_2 + \lambda v'')\ d\xi_1\ d\eta_1\ d\xi_2\ d\eta_2 \ , \qquad (1.93)$$

where

$$\Delta\xi = \xi_1 - \xi_2 \ , \qquad \Delta\eta = \eta_1 - \eta_2 \ ,$$

$$\Delta u = u'' - u' \ , \qquad \Delta v = v'' - v' \ ,$$

and the other symbols were defined in Sec. 1.2.2.3. Assuming that $H_0(\xi,\eta)$ is constant where $T_s(\xi/\lambda, \eta/\lambda)$ is effectively nonzero, the second term reduces to [see Eq. (1.24)]

$$\mathcal{P}^{-2}\iint\limits_{-\infty}^{\infty} T_s\left(\frac{\Delta\xi}{\lambda}, \frac{\Delta\eta}{\lambda}\right)T_s^*\left(\frac{\Delta\xi}{\lambda} + \Delta u, \frac{\Delta\eta}{\lambda} + \Delta v\right) d\Delta\xi\ d\Delta\eta$$

$$\cdot \iint\limits_{-\infty}^{\infty} |H_0(\xi,\eta)|^2 \left|H_0\left[\xi + \lambda\left(\frac{u'' + u'}{2}\right), \eta + \lambda\left(\frac{v'' + v'}{2}\right)\right]\right|^2 d\xi\ d\eta \ . \quad (1.94)$$

Bearing in mind that the seeing transfer function $T_s(u,v)$ has a width $\simeq r_0/\lambda$, then it is clear from Eq. (1.94) that $\langle T(u,v)T^*(u + \Delta u, v + \Delta v)\rangle$ can only be nonzero if $|\Delta u|$ and $|\Delta v| < r_0/\lambda$ [otherwise the first integral in Eq. (1.94) is zero].

If we make the further approximation that the seeing transfer function has a Gaussian shape [it is more accurately described by Eq. (1.17)], then it is straightforward to show that

$$\langle T(u,v)T^*(u + \Delta u, v + \Delta v)\rangle \simeq \langle |T(u,v)|^2\rangle|T_s(\Delta u/2, \Delta v/2)|^2 \ . \qquad (1.95)$$

That is, the Knox-Thompson transfer function is simply the product of the speckle transfer function at (u,v) and the squared modulus of the seeing transfer function at $(\Delta u/2, \Delta v/2)$. It follows from Eq. (1.95) that

$$\text{phase}\langle T(u,v)T^*(u + \Delta u, v + \Delta v)\rangle \simeq 0 \ .$$

Since Eq. (1.95) results from an oversimplified atmospheric model, it cannot be relied on quantitatively, but it does provide the correct qualitative condition on $|\Delta u|$ and $|\Delta v|$. Using the log-normal model, Fried[40] suggests that the optimum value of $|\Delta u|$ and $|\Delta v|$ is approximately $0.2r_0/\lambda$.

If we consider the Fourier transform of the object to be sampled on a grid of N by M points, there are approximately $2NM$ phase differences for a single choice of $(\Delta u, \Delta v)$; several schemes have been suggested for efficiently computing the required NM phases.[136–143] This problem is similar to that of calculating phases from shearing interferograms. It may be helpful[138] to use more than one value of $(\Delta u, \Delta v)$.

In a variation of the Knox-Thompson technique, Aitken and Desaulniers[144] suggest computing average ratios $\langle i(u,v)/i(u + \Delta u, v + \Delta v)\rangle$, a possible advantage being that a separate reference calibration may not be required. Sherman[145] has extended the technique to nonisoplanatic imaging. Brames and Dainty[146] have given an interpretation of the method in terms of the complex zero picture of Sec. 1.2.4.1; this picture may be useful for studying the role of noise in the technique. The effects of photon noise on speckle image reconstruction with the Knox-Thompson algorithm have recently been investigated.[147] Photon noise introduces a frequency-dependent bias, which must be corrected for successful reconstruction. In the photon-limited case (low light levels), Nisenson and Papaliolios[147] gave the lower bound on the number of frames M required for "good" image reconstruction of a point-like object as

$$M \geq 125 \left(\frac{N_{\text{sp}}}{\overline{N}} \right)^2 \ ,$$

where N_{sp} is the average number of speckles per frame and $\overline{N}(< N_{\text{sp}})$ is the average number of detected photons per frame.

1.2.4.8   Triple Correlation

Weigelt[100,101] has suggested another method for recovering the phase from a speckle integration. This method was originally called *speckle masking* from an early version that used photographic masks to extract the image from the autocorrelation. It is now more commonly referred to as the *triple correlation* (or bispectrum, in the frequency domain) method. The triple correlation is defined as

$$I_{\text{TC}}(\alpha,\beta,\alpha_0,\beta_0) = \langle I(\alpha,\beta)I(\alpha - \alpha_0, \beta - \beta_0)*I(\alpha,\beta)\rangle \ . \tag{1.96}$$

Fourier transforming this quantity yields the energy spectrum (or bispectrum) in the form

$$i_{\text{TC}}(u',v',u'',v'') = \langle i(u',v')i(u' + u'', v' + v'')i(u'',v'') \rangle \ . \tag{1.97}$$

Following the approach in Sec. 1.2.2.1, we can derive the bispectrum as the product of the object bispectrum and the telescope-atmospheric bispectrum:

$$i_{\text{TC}}(u',v',u'',v'') = o(u',v')o(u' + u'', v' + v'')o(u'',v'')$$

$$\times \ \langle T(u',v')T(u' + u'', v' + v'')T(u'',v'') \rangle \ . \tag{1.98}$$

The problem of correcting for the speckle transfer is identical to the amplitude for speckle interferometry (Sec. 1.2.2.8). The general approach is to calculate the bispectrum from data taken with similar atmospheric statistics on an unresolved reference star. We then calculate the object bispectrum as the ratio of the image bispectrum and the data bispectrum:

$$O_{\text{TC}}(u',v',u'',v'') = \frac{\langle i_{\text{TC}}(u',v',u'',v'') \rangle}{\langle R_{\text{TC}}(u',v',u'',v'') \rangle} \ , \tag{1.99}$$

where R_{TC} is estimated from the reference source data.

As with the Knox-Thompson algorithm, the phases of the object energy spectrum are extracted from the corrected bispectrum by a set of recursion relations. See, for example, the review by Weigelt.[148] This recursive algorithm has the limitation that phase errors accumulate from low to high frequencies. Glindemann et al.[149] have suggested using a least-squares method for recovering the phase that does not have this cumulative error problem.

A number of analyses[150–152] have been carried out calculating and comparing the expected SNR for Knox-Thompson and triple correlation. The results differ somewhat from one analysis to another. Ayers et al.[150] point out that both the Knox-Thompson and triple correlation methods may be considered as four-dimensional problems, in which phases are correlated over all possible baselines. In this case, Chelli[152] has found the triple correlation to be slightly superior at high light levels, and the Knox-Thompson to have slightly better SNR at low light levels. Ayers et al.[150] point out that the Knox-Thompson may be degraded by frame-to-frame centroiding errors, reducing the low- light-level advantage to approximate equality.

1.2.4.9 Summary. In this section we have reviewed a number of possible techniques for solving the phase problem (that is, reconstructing the object intensity) in stellar speckle interferometry. The methods fall into two categories; those that require only the modulus of the object Fourier transform (covered in Secs. 1.2.4.2 and 1.2.4.3) and those that utilize other information present in the original speckle exposures (Secs. 1.2.4.4 through 1.2.4.8). It seems obvious that methods in the latter category are preferable for this particular phase problem, because they make use of additional information present in the available data.

Great progress has been made in the last decade on the use of speckle imaging techniques with a large body of scientific results having been published over the last several years. Progress in algorithms, detectors, and fast

computers has made full image reconstruction a standard tool for high angular resolution astronomy.

1.2.5 Implementation

1.2.5.1 Data Collection and Processing.
Speckle camera systems have been constructed by a number of groups.[141,153–157] As an example, we describe a first-generation system used at Kitt Peak National Observatory for many years, a diagram of which is shown in Fig. 1.18.

Referring to Fig. 1.18, light from the telescope passes through an electromechanical shutter (1) at the front of the speckle camera system and reaches the Cassegrain or Ritchey-Chrétien focus at (2). At the 4-m Mayall telescope, the image scale at this focus is approximately 6.5 arcsec/mm so that a lens (3) is required to magnify the image 10 or 20 times giving final image scales of approximately 0.65 and 0.32 arcsec/mm, respectively. At 500 nm, the diffraction-limited angular frequency of a 4-m telescope is approximately 40 arcsec^{-1}, and the sampling theorem therefore requires image plane sampling at $\Delta\alpha \leqq 0\rlap{.}''012$, or 0.04 mm or less in the 20× magnified image plane; this value also determines the resolution (or MTF) of the image detection system.

It is necessary to correct for atmospheric dispersion except when observing close to the zenith. The magnitude of atmospheric dispersion depends on a number of factors,[158] but it is approximately given by

$$\Delta z \cong 0.3 \tan z \quad [\text{arcsec/100 nm}] \;,$$

in the middle of the visible spectrum. Either a grating system[153,155] or a pair of Risley prisms[154,156] can be used to correct this; Fig. 1.18 shows the use of a prism pair. The optimum choice of glasses for the prisms are LaK24 and KF9—these match the dispersion of air over the broadest wavelength range.[159] A narrow-band interference filter (5) selects the mean wavelength and bandpass (see Sec. 1.2.2.6 for a discussion of the permissible bandpass).

The most critical element in a speckle camera system is the image detector. Figure 1.18 shows an image intensifier/photographic film combination, which
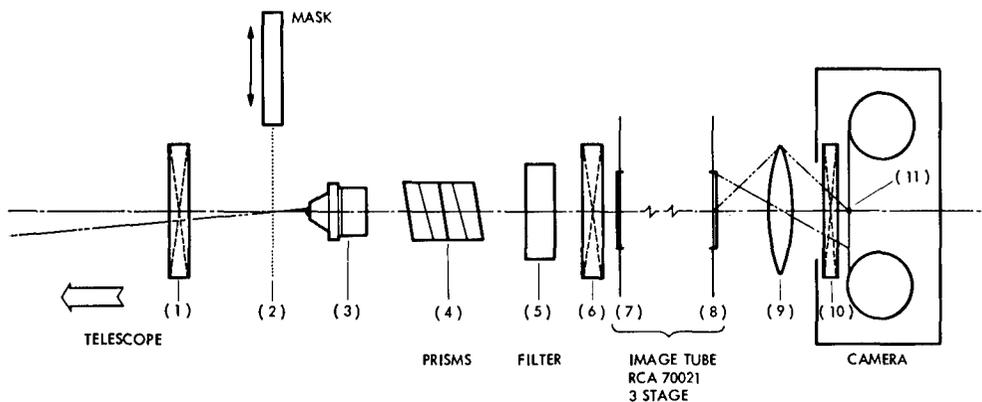
Fig. 1.18 Schematic cross-section view of a speckle camera.[156]

has the advantage of simplicity. A variety of image intensifiers may be suitable—magnetically or electrostatically focused cascade systems, or microchannel plate devices; a variable (high) gain and low background are two practical requirements for the intensifier. Recently constructed speckle cameras and those under construction all use some form of electronic readout; this has the potential advantages of overcoming the noise and nonlinearity of photographic film and of allowing the possibility of real-time analysis of the data.

The type of electronic image detector required depends to a certain extent on the type of astronomical speckle observations that are planned and the intended method of data analysis. Before describing possible detectors it is therefore appropriate to discuss methods of data reduction. In the first-generation speckle cameras, the photographic images were analyzed in a coherent optical processor; this extremely simple analog device gives as output the energy spectrum of the complex amplitude transmittance of the film, the average energy spectrum being found by summation of the energy spectra of M frames ($M < 1000$ in practice). This technique could also be used for other "real-time" photographic-type detectors,[137] but these analog systems tend to suffer from nonlinearities and noise. Digital processing appears to offer more flexibility and is the only way of implementing some of the object reconstruction algorithms described in Sec. 1.2.4.

For conventional speckle interferometry, two approaches are used to calculate the object information; one is via the average energy spectrum, as in Eq. (1.5), and the other is via the average spatial autocorrelation function as in Eq. (1.3). Allowing for moderate oversampling, large telescope speckle data require a format of at least 256 × 256 pixels, and a desirable frame rate is approximately 50 s^{-1}. Devices that compute Fourier transforms of this size at this rate are becoming available, but their cost may not be justified in this application. Consequently, the average energy spectrum method of analysis is currently done after the observations have been made and stored on a suitable medium such as videotape.

On the other hand, the autocorrelation method of analysis lends itself to real-time computation. Vokac[160] has described a prototype on-line digital autocorrelator for 16-level (4-bit) 64 × 64 pixel images taken at a rate of 2 s^{-1}, and predicted that full-scale throughput would be possible with current technology. Blazit[161] and the London group[162] have constructed 1-bit vector autocorrelators that process images containing a few photons (< 200) at 25 s^{-1}. Vector autocorrelators work on the principle that the autocorrelation function of an image consisting entirely of ones and zeros (presence or absence of a photon) is equal to the histogram of vector differences between all possible pairs of photons. This algorithm can either be hardwired in a special-purpose device or programmed into a fast commercial or customized microcomputer. An example of the resolution of a binary star obtained with such a device (in real time) is shown in Fig. 1.19. Another approach suggested by Cole[163] uses optical circuit elements to allow higher photon rates.

Depending on the type of data analysis to be used, there are several possible electronic detector systems. One of the most straightforward is to use an intensified television camera or intensifier plus television camera combination. The format of the data allows easy storage on videotape, but digital computer analysis, via a video-digitizer system, may be tedious. Another possibility is

Fig. 1.19 Output of a vector autocorrelator display in real time when observing a binary star. (Courtesy of B. L. Morgan and H. Vine, Imperial College, London)

to replace the television camera with a charge-coupled device (CCD).[164] The advantage of both of these approaches is that either analog (intensity) or digital (photon counting) data may be processed. For low light levels, photon counting devices in which the position and time of photoelectron events are recorded may be preferable,[165–167] particularly since the recorded data are already in a suitable format for vector autocorrelation processing.

Two-dimensional photon counting detectors such as the MAMA detector,[168] the RANICON detector,[169] and the PAPA detector[167] have several important advantages for speckle, particularly for image reconstruction. They record lists of photon positions and time of arrival, so the data may be regrouped into frames of arbitrary length, allowing optimization of exposure times for each data set. This is not possible with framing cameras such as CCDs and intensified video cameras. They also give exact counts of the number of photons and provide centroided photon positions, so that the photon noise biases that accumulate (and dominate for low light levels) can be accurately subtracted.[170]

1.2.5.2 One-Dimensional Infrared Speckle Interferometry.

Until recently, efficient two-dimensional array detectors in the near infrared (2 to 5 μm) have not been widely available and therefore infrared speckle interferometry had to be practiced using only a single detector element. This feature, some other special problems that are encountered, and its demonstrated astronomical success make it worthwhile to devote a section of this chapter to infrared speckle interferometry.

At first glance, infrared speckle interferometry would seem less fruitful than that in the visible range, particularly in view of the restriction to a single detector element. Table 1.1 summarizes the resolution according to the diffraction limit for a 4-m telescope (column 2) and the seeing limit (column 4) for the wavelengths of 0.5, 2.2(K), 3.45(L), and 4.8(M) μm. From column 2, it can be seen that the diffraction-limited angular resolution (Rayleigh criterion) is approximately 0″.03 at 0.5 μm, but only 0″.30 at 4.8 μm, whereas a

Table 1.1   Resolution According to the Diffraction Limit for a 4-m Telescope

(1) Wavelength [μm]	(2) $\Delta\alpha\,(D = 4$ m) [arcsec]	(3) r_0 [m]	(4) ω [arcsec]
0.5	0.03	0.1	1.00
2.2 (K)	0.14	0.6	0.74
3.45 (L)	0.22	1.0	0.68
4.8 (M)	0.30	1.5	0.64

1″0 seeing-limited image at 0.5 μm is slightly smaller, 0″64, at 4.8 μm. Thus, taking the ratio of columns 2 and 4, we see that there is typically a 33× increase in angular resolution possible by doing speckle interferometry at 0.5 μm, whereas the improvements at 2.2, 3.45, and 4.8 μm are only 5, 3, and 2, respectively (this does assume "good" seeing). The reason infrared speckle has been so valuable is that, despite the relatively poorer angular resolution, there are many more potentially resolvable (i.e., large) bright objects in the near infrared than in the visible.

The technique of one-dimensional infrared speckle interferometry is described in Refs. 171 through 175, particularly the comprehensive paper by Sibille et al.[172] In the method developed by the French group, the image is scanned over a long, narrow slit and the light collected by a single indium antimonide (InSb) detector cooled to liquid nitrogen or helium temperature. The bandwidth restrictions are much less severe in the infrared than in the visible [see Eq. (1.37) and Table 1.1], the maximum $\Delta\lambda/\bar\lambda$ being on the order of 0.13 at 2.2 μm and 0.37 at 4.8 μm. The scanning speed of the image across the slit has to be sufficient to "freeze" the speckle, rates of 50 to 100 arcsec s^{-1} being typical; the effect of scanning rate is described by Aime et al.[176]

If the scan is assumed to be along the α-axis (corresponding to the u axis in the angular frequency plane), the temporal average energy spectrum $\langle|i(f)|^2\rangle$ of the image intensity $I(t) \equiv I(\alpha/v)$, where v is the scan rate, is given by

$$\langle|i(f)|^2\rangle = |o(u,0)|^2\langle|T(u,0)|^2\rangle T_{\text{slit}}^2(u) \ , \tag{1.100}$$

where the temporal frequency f is related to the angular frequency u by $f = uv$ and where the slit transfer function for a slit of width α_{slit} is

$$T_{\text{slit}}(u) = \frac{\sin(\pi u\alpha_{\text{slit}})}{\pi u\alpha_{\text{slit}}} \ . \tag{1.101}$$

The one-dimensional temporal energy spectrum can easily be computed on-line using a commercial microcomputer. By observing a reference star, the speckle transfer function can be found, so that a section through the modulus of the object energy spectrum $|o(u,0)|$ can be found. The complete modulus could, in principle, be found by rotating the scan direction, although because of practical problems connected with atmospheric instability only north-south and east-west scans are usually made.

One of the greatest problems encountered in implementing infrared speckle interferometry is the instability of atmospheric turbulence. Because the seeing limited angular frequency portion of the speckle transfer function is a significant part of the whole transfer function, it is not possible to use Worden's scheme for self-calibrating the method. Accordingly, a typical observing sequence is object → sky → reference → sky → object, taking perhaps 100 to 1000 scans of each and repeating the sequence until consistent results are obtained. The "sky" measurement is required in the infrared due to emission from both the sky and the telescope, and an estimate of the energy spectrum of the object is obtained from

$$|o(u,0)|^2 \cong \frac{\langle |i_{\mathrm{obj}}(f)|^2 \rangle - \langle |i_{\mathrm{sky}}(f)|^2 \rangle}{\langle |i_{\mathrm{ref}}(f)|^2 \rangle - \langle |i_{\mathrm{sky}}(f)|^2 \rangle} \ . \qquad (1.102)$$

The SNR of the slit scan method is derived by Sibille et al.[172] In addition to the atmospheric fluctuation and photon noise of the signal that are the only fundamental contributions in the visible, there is now also the photon noise of the "sky" background and noise inherent in the detector, such as Johnson noise. Limiting magnitudes, based on the value of the object intensity that yields an energy spectrum equal to that of the noise sources for a single 100-ms scan, were predicted to be of the order of 5 to 6 for the $K, L,$ and M wavelengths, although practical experience indicates limiting magnitudes of approximately 7 (K) to 2 (M). Selby et al.[171] used a grating rather than a slit, thus measuring only a single-frequency component at a time; they claim fainter limiting magnitudes but these have not yet been achieved.

In a new development of one-dimensional speckle interferometry (visible or infrared) Aime et al.[177] suggested the use of a telescope with a one-dimensional aperture (e.g., 10×800 cm^2). This gives a contrast gain over a circular aperture and, associated with a spectroscope, allows investigation of a spectral-angular plane with no loss in light.

More recently, infrared arrays have become available[178] for speckle applications with current array sizes up to 256×256 pixels. These arrays have dramatically expanded the range of astronomical science that can be accomplished by means of speckle techniques in the infrared.

1.2.6 Astronomical Results

Observational speckle interferometry is now two decades old, and approximately 300 papers primarily concerned with astronomical results have been published. Despite the enthusiasm of a growing group of astronomers, it is only realistic to point out that the technique is not widely used by the astronomical community at large. Some possible reasons for this are (1) relatively few objects, particularly in the visible, are resolvable by 4-m class telescopes, whose diffraction limit at 400 nm is 0″.02; (2) calibration problems make it difficult to obtain photometric energy spectra of sufficient accuracy for the particular astronomical problem; and (3) only the most expensive equipment yields faint limiting magnitudes and enables the vast amounts of data to be reduced.

This section summarizes astronomical results in four parts: solar system objects, binary stars, single stars, and infrared objects. We include discussion of some of the many astronomical results that have been obtained with speckle interferometry and related techniques. In recent years, the improvements in detectors and algorithms and in the processing power of computers has led to an explosion of activity and a dramatic increase in the number and importance of scientific results. A complete bibliography of scientific papers in this field, through 1990, has been generated by Gezari et al.[179] and is available as a NASA publication.

In addition to the four types of results discussed, some more unusual objects have also been observed by means of speckle interferometry. For example, Hege et al.[64] resolved one of the components of the "triple" quasar PG 115 + 08 as a binary, one of the faintest objects studied by the speckle method ($m_v \simeq 16.2$). The Seyfert galaxy NGC 1068 has been observed in the visible[162] and at 2.2 μm,[180] both results revealing a nuclear core containing most of the luminosity. The study of Ebstein et al.[181] observed both NGC 1068 and NGC 4151, the two brightest Seyferts. They showed that, when observed in a narrow emission line (OIII), the very bright nucleus exhibited similar (but not identical) structure to maps generated at radio wavelengths with the Very Large Array telescope. Structure in the nucleus was also observed for NGC 1068 in the infrared.[182]

However, due to their intrinsic faintness, extragalactic astronomy is very difficult with speckle techniques, despite the improvements in detector technology.

1.2.6.1 Solar System Objects.

The angular diameters of the asteroids Pallas and Vesta were measured by Worden et al.,[183–185] the results for Pallas indicating some elongation of the object. The asteroids Juno and Amphitrite,[186] Eros,[187] Herculina,[188] and Davida[189] have also been resolved with speckle. Newer observations of Vesta include Knox-Thompson reconstructions that show surface features, and a sequence of images that show rotation.[190] The diameter of the planetary satellites Rhea and Iapetus[185] and Titan[191] have also been measured. Diameters have also been measured for the larger satellites of Uranus and Neptune.[192]

Observations of the planet Pluto and its moon Charon are near to the limiting magnitude of speckle interferometry, their magnitudes being approximately 15.3 and 16.9, respectively. Arnold et al.[193] estimate Pluto's diameter to be 3000 ± 400 to 3600 ± 400 km depending on whether limb darkening is incorporated in the model. This is slightly smaller than that measured by Bonneau and Foy,[194] 4000 ± 400 km with no limb darkening, who also estimate the diameter of Charon to be 2000 ± 200 km and propose a revised orbit for the moon. Both results imply a mean density of Pluto (and Charon) $\simeq 0.5 \text{ g cm}^{-3}$. The Pluto-Charon system has continued to be observed and studied, including image reconstruction using speckle masking over seven different nights.[195] Infrared imaging[196] of Jupiter's satellite Io shows hot spots on the surface that probably correspond to volcanic activity. This appears to be a promising approach to long-term monitoring of Io's rapidly evolving surface.

The Solar granulation has also been measured by speckle interferometry,[197–205] the main technical problem here is the absence of any reference source for estimation of the speckle transfer function. Image reconstruction

techniques (using the Knox-Thompson algorithm) have been applied to solar features.[199] Keller and von der Luhe[206] have combined speckle interferometry with polarimetry to allow the study of small-scale magnetic features on the sun. Keller[207] has employed the same techniques to possibly resolve magnetic flux tubes, which are thought to be the principal conduits of energy from the solar convection zone into the solar corona.

1.2.6.2 Binary Stars.

Speckle interferometry has been very successful when used to determine the angular separation and position angle of binary stars. McAlister and his colleagues[208–213] have reported more than 1000 measurements of resolved binaries, as well as a number of detailed studies of individual systems.[214–216]

Several hundred observations have also been reported by three other groups.[7,37,217–224] Several reasons have contributed to the success of speckle interferometry in this area; the measurements are among the simplest speckle observations to make, can be made rapidly on brighter stars (McAlister[225] reported 125 to 175 observations per clear night), and yield an accuracy far exceeding visual observations. McAlister[225] mentioned typical errors of 0.6% on the separation and ± 2 deg on the position angle, although other groups gave more conservative error estimates.[219]

The aim of making binary star measurements is usually to estimate the masses of each component. For a double-lined spectroscopic binary (i.e., one for which the radial velocities of both components are known) a minimum of two measurements of the angular separation and position angle yields both the masses of each component and the absolute distance (parallax). One example measured by McAlister[226] is 12 Persei; the masses are 1.25 ± 0.20 and 1.08 ± 0.17 times the mass of the sun and parallax is 0."046 ± 0."002, which combined with the known apparent magnitudes gives absolute visual magnitudes of 3.8 ± 0.1 and 4.1 ± 0.1, respectively.

Binaries that are both double-lined spectroscopic and resolvable by speckle interferometry are rather rare. If the binary is single-lined, then speckle observations cannot unambiguously give the individual masses and distances. However, if masses appropriate to the spectral type are assumed, a distance can be found. McAlister[227] and Morgan et al.[219] have applied this to binaries in the Hyades cluster, a distance marker in the universe, to confirm that its mass-luminosity relationship is normal and that its distance is approximately 10% greater than the original proper motion studies indicated. A full orbit for the Hyades binary Finsen 342 has been measured[228] and 28 combined visual/speckle orbits of close binary systems have been published.[229]

Beckers[230,231] has suggested a modification of the speckle technique called *differential speckle interferometry* that may enable submilliarcsecond separation of binary stars to be measured on a 4-m class telescope in the visible. The technique uses the Doppler shift and observation at two closely spaced wavelengths to modulate the position of speckles in the short-exposure photographs; since the speckle procedure measures shifts to an accuracy of a fraction of the speckle size, resolution of binaries whose separation is much less than the diffraction limit may be possible. This technique has been successfully employed to perform submillisecond imaging of rapidly rotating stars.[232] Speckle imaging in the infrared has proven to be extremely important in the study of

duplicity in pre-main sequence stars. Since Dyck et al.[233] discovered a protostar companion to T Tauri, there have been extensive efforts by many to look for companions to young stars in order to determine their influence of the early evolution of such systems.[234-239]

1.2.6.3 Stellar Physics.

One of the first stellar disks to be resolved by speckle interferometry was the supergiant α-Orionis (Betelgeuse),[7] which has subsequently been observed on many occasions.[120,121,221,240-242] The results include measurements of the mean diameter in $H\alpha$ and the limb-darkening coefficient[243-245]; two-dimensional image reconstruction[246,247]; detection of companions that may play a role in the star's episodic outbursts of mass loss[248]; and a study of the dust envelope close to the star.[249,250]

Several Mira variable stars—O Ceti (Mira), R. Leo, and χ Cygni[240,251,252]—have been observed by speckle interferometry. O Ceti, the closest and brightest Mira, has been studied with speckle techniques in attempts to better understand the mechanisms for the regular pulsation of these highly variable stars. Labeyrie et al.[251] have found that the apparent diameter of these stars changes by a factor of 2 or more when observed in and out of TiO absorption bands. Karovska et al.[253] have measured asymmetries in the shape of O Ceti's envelope, which change with the phase of the star's pulsation period.

Eta Carina, one of the most massive stars in the galaxy, shows jet-like features, a bipolar nebula, and extended emission features from recent violent eruptions when examined with speckle imaging.[254-257]

R136a, a bright feature in the 30 Doradus region of the Large Magellenic Cloud, was studied by Weigelt and Baier[258] by means of speckle masking and found to be a cluster of stars rather than an inexplicably massive bright single star.

SN1987a, the brightest and closest supernova in nearly 400 years, became the most intensely studied astronomical object in history. Speckle observations directly measured the size and the velocity of the expanding debris from early epochs[259-262]; detected asymmetries in the shape of the debris[263]; and detected a bright source close to the supernova, which has still not been completely understood.[264]

Speckle interferometry is now playing an important role in stellar research, particularly when the derived information is combined with observational results from other techniques, including photometry, spectroscopy, or polarimetry.

Finally we note that the first results of long-baseline two-telescope speckle interferometry have resolved the individual components of the binary star Capella, yielding values of 5 ± 1 and $4 \pm 2 \times 10^{-3}$ arcsec.[265]

1.2.6.4 Infrared Stars.

Although infrared speckle interferometry has been until recently restricted to a single detector across which the image is scanned, many interesting measurements have been made. This work is rapidly expanding now that array infrared detectors have become available.

The diameters of several protostar candidates have been measured, particularly WS-IRS 5, MonR 2-IRS 3, S 140-IRS 1 and the BN object.[173,175,266-268] McCarthy's measurements[268] of the triple nature of MonR 2-IRS 3 are a good example of the results possible with careful data analysis. Speckle interferometry in the infrared has been extensively employed in the search for "Brown Dwarfs," stars that have insufficient mass to "turn on" nuclear fusion but which

"glow" in the infrared.[236,269,270] While there have not been any certain detections, these objects are thought to be important candidates for the "missing mass," which is extrapolated from gravitational dynamics, but not detected directly as visible luminosity. The bright carbon star IRC + 10216 has been observed both in the continuum and in the CO lines.[271-272] Several Mira variables[175,271] and the dust shells around Wolf-Rayet stars[273-275] and the supergiant α-Orionis[173] have also been observed. The star T Tauri, after which the class of T Tauri variable stars is named, has been shown to have a protostar companion.[276] Object restoration via the Knox-Thompson algorithm and image enhancement techniques have been applied to infrared observations of the extended object η Carinae.[277]

1.3   APERTURE-PLANE INTERFEROMETRY

Written by Jean-Marie Mariotti and François Roddier

1.3.1   Theoretical Background

1.3.1.1   Aperture-Plane Versus Image-Plane Interferometry.   Aperture-plane or pupil-plane interferometry was first independently proposed by Breckinridge[278a] and KenKnight[279] in 1972, as a means to reconstruct turbulence degraded images. Since then, the technique has been developed to overcome some limitations of image-plane speckle interferometric techniques. One major shortcoming of image-plane speckle images is the random attenuation of the object spectral components introduced by the turbulent atmosphere. As a consequence, speckle interferometry requires a large number of independent images to be taken, even in the case of bright sources, and it requires careful calibration of the resulting modulation transfer function (MTF) on an unresolved star under exactly the same seeing conditions, which is difficult to achieve.

The reason for this random attenuation of the speckle transfer function is to be found in the so-called redundancy of the beam recombination.[280] One can think of the instantaneous speckle image as a superposition of interference fringe patterns. Each fringe pattern corresponds to a spatial frequency f and is produced by the interference of beams issued from aperture point pairs with the same separation and orientation. In the image plane, many such pairs of interfering beams contribute to the same spatial frequency (hence, the name *redundancy*). Because different pairs are affected differently by turbulence, the associated fringe patterns are randomly shifted by different amounts and do not add in phase. This incoherent addition of complex vectors (phasors) of identical amplitude but random phases produces a random attenuation of the spectral energy transfer function (also called speckle transfer function). Beyond the seeing frequency cutoff, the average attenuation is proportional to the inverse of the number of coherent areas in the aperture, or the number of speckles in the image plane.

By putting a mask with two small holes on a telescope aperture, one isolates a single pair of interfering beams producing a nonredundant beam recombination. This is a Michelson-Fizeau interferometer. Apart from small effects

due to stellar scintillation, atmospheric turbulence does not affect the amplitude of the fringes produced by a Michelson-Fizeau interferometer. It only introduces phase delays, which randomly shift the fringe pattern. Hence, as long as the exposure time is short enough to freeze the random fringe motion, the true fringe visibility is recorded. If the object is bright enough, the visibility can be estimated from a single exposure. The drawback of this approach is that only one object Fourier component is measured at a time. The technique can be extended by using masks with an array of holes in a suitable configuration forming nonredundant pairs. The technique was first demonstrated by Goodman and coworkers[281,282] and has only recently been applied to astronomical observations.[283-287] Unfortunately, the SNR for the reconstructed image decreases with the number of holes,[288] and most of the light is thrown away by the mask, which severely limits the potential of the technique. Clearly, the ideal would be to use an array of Michelson-Fizeau interferometers with as many different baselines as possible working all in parallel. This is precisely what aperture-plane interferometry provides.[289-292]

Compared to speckle interferometry, aperture-plane interferometry has the following advantages:

1. better SNR on bright sources
2. less sensitivity to calibration errors due to fluctuations of seeing
3. insensitivity of the object Fourier amplitudes to telescope aberrations (even if larger than atmospheric-induced aberrations)
4. local detector errors have only local effects in the Fourier plane and are less harmful
5. makes a better use of the detector dynamic range to detect faint extended structures such as stellar envelopes.

The only fundamental drawback of the technique is that it is essentially limited to sources smaller than the seeing disk. For larger sources, the SNR decreases as the source size increases and the advantage over image-plane techniques is lost.

1.3.1.2 Wavefront Shearing Interferometers.

Aperture-plane interferometry provides a way to recombine many light rays in a nonredundant fashion. A beamsplitter is used to produce two images of the telescope entrance pupil through a wavefront shearing interferometer. Interference fringes are observed within the common part of the two pupil images, hence the name *aperture-plane interferometry*. Figure 1.20 shows examples of various types of wavefront shears that can be used.

In a lateral shear interferometer, one pupil image is shifted with respect to the other. All the baselines over which interference occurs measure the same object Fourier component [Fig. 1.20(a)]. Many sequential measurements are required to cover the object Fourier spectrum as with a Michelson-Fizeau interferometer. Because it avoids the use of a pupil mask it is more efficient although light is still lost because the two pupil images only partially overlap. A regular Mach-Zehnder interferometer can be conveniently used as a lateral-shear interferometer.[293] Ribak uses two Dove prisms and introduces an acoustic modulator.[294-298] For an efficient use, it requires as many detectors as coherent areas on the telescope pupil. Instead of modulating very wide fringes,
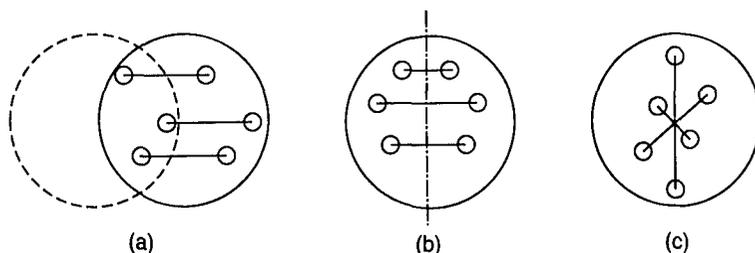
Fig. 1.20 Examples of various types of shears used in shearing interferometry: (a) lateral shear interferometer, (b) folding interferometer, and (c) 180-deg rotational shear interferometer. Interfering points are connected with a straight line.

one can introduce a tilt between the two interfering beams and record narrower fringes with a camera. However, the fringe spacing is wavelength dependent. This has the drawback of putting a more stringent constraint on the optical bandwidth than does speckle interferometry. To avoid this drawback achromatic shearing interferometers have been developed.[299-301] They have mainly been used for visual measurements of double stars.

In a wavefront folding interferometer, one pupil image is superimposed on its mirror image [Fig. 1.20(b)]. Several baselines still measure the same spatial frequency. All the baselines measure spatial frequencies in the same direction. As a result the interferometer produces a map of a one-dimensional Fourier transform.[302] Sequential measurements are still required to map the object two-dimensional Fourier transform. Currie uses a Köster prism to fold the wavefront.[303-305] Good astronomical results were obtained with a pair of detectors. Full efficiency would require a detector array. Another convenient configuration consists of replacing a flat mirror with a roof prism in a Michelson interferometer.[306-308]

In a rotational shear interferometer, one pupil image is rotated with respect to the other [Fig. 1.20(c)]. If the rotation axis coincides with the center of the pupil, the two images again exactly overlap and no light is lost. Each baseline measures a different spatial frequency, mapping the whole frequency plane up to a cutoff frequency that depends on the rotation angle. A 180-deg rotational shear produces a two-dimensional map of the object Fourier transform up to the cutoff frequency f_c of the telescope.[302] A rotation angle β produces a two-dimensional map up to the spatial frequency $f = f_c \sin(\beta/2)$. This zoom effect was first mentioned by Lowenthal, Serres, and Froehly[309] and later independently by Bryngdahl and Lohmann.[310] It is useful to match the Fourier plane coverage to the extent of the object Fourier transform. At the expense of a lower resolution, one can improve the SNR on extended objects by putting all the available photons in the useful part of the Fourier transform signals. In direct imaging this is done by choosing the appropriate detector pixel size. In interferometric imaging, a variable rotational shear is a unique means to provide the same flexibility.

Various optical configurations producing a rotational shear are described by Armitage and Lohmann.[311] This type of shear is particularly suitable for coherence measurements[312,313] and such interferometers were widely developed in the early 1970s for the study of optical propagation through turbulence.

Rotation shearing interferometers specially built for interferometric imaging in the visible are described by Breckinridge,[278b,314–316] by Itoh and Ohtsuka,[317] and by C. Roddier et al.[318–320] In all cases a small tilt angle is introduced between the two interfering beams and fringes are recorded with a camera. As discussed earlier, this is done at the expense of a reduced bandwidth. This constraint can again be alleviated by introducing appropriate spectral dispersion.[321] In practice, one can simply modify the setting of a dispersion compensator or simply use atmospheric dispersion with a proper choice of the fringe spacing and orientation. An interferometer built for infrared applications is described by Mariotti.[322–325] In this case, the tilt angle has been reduced to zero, which widens the fringe spacing to the point where they disappear. Fringes are said to be fluffed out. The local fringe contrast is measured by modulating the optical path difference between the two interferometer arms.

An important problem associated with rotation shearing interferometers is polarization matching. When fed with unpolarized light, the states of polarization of the two interfering output beams do not perfectly match, producing fringes with a reduced contrast. This problem is generally solved by putting a properly oriented polarizer at the interferometer input. For low-light-level astronomical applications this solution is unacceptable. However, high-contrast fringes can be obtained for any rotation angle by means of phase-compensated roof prisms.[326]

Because it allows the recording of all the object Fourier components simultaneously under exactly the same seeing conditions, rotational shearing interferometry is generally the preferred technique. In the following, only rotational shear interferometers are considered.

1.3.1.3 Interferometric Imaging and Incoherent Holography.

Let us first consider a monochromatic point source at infinity in a direction α with respect to the telescope optical axis. For the time being we assume no perturbation (no turbulence) along the optical path. On the telescope aperture plane, the source produces a plane wave tilted at an angle α that is a complex amplitude of the form

$$A_0(\mathbf{r}) = \Omega \exp\left(-2i\pi \frac{\alpha \cdot \mathbf{r}}{\lambda}\right) , \qquad (1.103)$$

where λ is the wavelength of the source. A rotational shear interferometer produces two images of the entrance aperture, one rotated with respect to the other by an angle β. Hence, the complex amplitude at the interferometer output is

$$A(\mathbf{r}) = \frac{1}{\sqrt{2}}[A_0(\mathbf{r}) + A_0(\mathscr{R}\mathbf{r}) \exp(i\theta)] , \qquad (1.104)$$

where $\mathscr{R}$ is a rotation operator with rotation angle β. The angle θ represents a possible phase delay between the two interferometer arms. Putting Eq. (1.103) into Eq. (1.104) gives

$$A(\mathbf{r}) = \frac{\Omega}{\sqrt{2}} \left[\exp\left(-2i\pi \frac{\boldsymbol{\alpha} \cdot \mathbf{r}}{\lambda} \right) + \exp\left(-2i\pi \frac{\boldsymbol{\alpha} \cdot \mathscr{R}\mathbf{r}}{\lambda} \right) \exp(i\theta) \right] \ .$$

(1.105)

The illumination $I(\mathbf{r})$ at the interferometer output is given by the squared modulus of the complex amplitude

$$I(\mathbf{r}) = |\Omega|^2 \left\{ 1 + \mathrm{Re} \left\{ \exp\left[2i\pi \frac{\boldsymbol{\alpha} \cdot (\mathscr{R}\mathbf{r} - \mathbf{r})}{\lambda} \right] \right\} \exp(-i\theta) \right\} \ ,$$

(1.106)

where Re denotes a real part. It requires only simple geometry to show that

$$\mathscr{R}\mathbf{r} - \mathbf{r} = 2|\sin\beta/2|\mathscr{R}'\mathbf{r} \ ,$$

(1.107)

where $\mathscr{R}'$ is a rotation operator with rotation angle $(\beta + \pi)/2$. Introducing the new rotated coordinates $\mathbf{r}' = \mathscr{R}'\mathbf{r}$, Eq. (1.106) can be written

$$I(\mathbf{r}') = |\Omega|^2 \left[1 + \cos\left(4\pi|\sin\beta/2|\frac{\boldsymbol{\alpha} \cdot \mathbf{r}'}{\lambda} - \theta \right) \right] \ .$$

(1.108)

The cosine term in Eq. (1.108) describes interference fringes perpendicular to the direction $\boldsymbol{\alpha}$ with fringe spacing $\lambda/(2|\sin\beta/2|\|\boldsymbol{\alpha}\|)$.

Let us now consider the case of an extended object. Assuming that the object is an incoherent light source, the illumination at the interferometer output will be the sum of illuminations produced by each of its points. For a brightness distribution $|\Omega|^2 = O(\boldsymbol{\alpha})$, Eq. (1.108) becomes

$$I(\mathbf{r}') = \int O(\boldsymbol{\alpha}) \left[1 + \cos\left(4\pi|\sin\beta/2|\frac{\boldsymbol{\alpha} \cdot \mathbf{r}'}{\lambda} - \theta \right) \right] d\boldsymbol{\alpha} \ .$$

(1.109)

It is the sum of a constant term plus a cosine transform of the source brightness distribution. By recording the illumination $I(\mathbf{r}')$ for different values of θ one can obviously obtain both the cosine and the sine transform of the object brightness distribution $O(\boldsymbol{\alpha})$ and recover it by means of an inverse Fourier transform. Under certain conditions one can also recover it from a single interferogram. Assuming $\theta = 0$, Eq. (1.109) can be written

$$I(\mathbf{r}') = \hat{O}(0) + \mathrm{Re}[\hat{O}(k\mathbf{r}')] \ ,$$

(1.110)

where $\hat{O}(\mathbf{r}')$ denotes the Fourier transform of $O(\boldsymbol{\alpha})$ and

$$k = \frac{2|\sin\beta/2|}{\lambda} \ .$$

(1.111)

Taking the Fourier transform $\hat{I}(\mathbf{u}')$ of the illumination $I(\mathbf{r}')$ in the interferogram gives

$$\hat{I}(\mathbf{u}) = \hat{O}(0)\delta(\mathbf{u}) + \frac{1}{2|k|}[O(\mathbf{u}/k) + O(-\mathbf{u}/k)] , \qquad (1.112)$$

where $\delta(\mathbf{u})$ is a Dirac impulse at the origin. If the object is sufficiently off axis, the three terms in Eq. (1.112) do not overlap and its brightness distribution can be recovered.

This is the basis of incoherent holography. The technique consists of recording the illumination $I(\mathbf{r})$ on photographic film to produce a hologram, the behavior of which is very similar to that of coherent holograms recorded under coherent illumination; that is, an image of the original object can be seen in the diffraction pattern produced by the hologram when illuminated with a laser beam. This is demonstrated in Fig. 1.21, which shows the diffraction pattern produced by a hologram of the word "fourier" recorded through a rotational shear interferometer using an incandescent white light source. Note the bright central peak represented by a delta impulse in Eq. (1.112) together with the word and its mirror image.

Goodman[327] first pointed out the analogy between interferometric imaging of incoherent sources and holography. Incoherent holography was developed[328-330] in the 1960s as a means to extend holographic methods to incoherently illuminated objects. The success of incoherent holography was limited because of the steep decrease of the SNR as the number of resolved points increases in
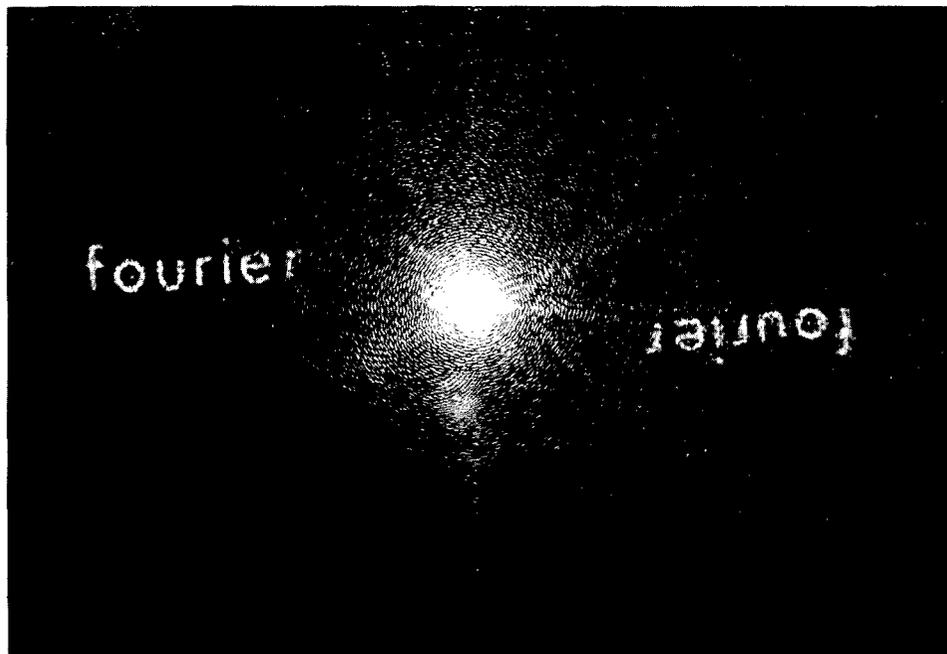
Fig. 1.21   Diffraction pattern produced by an incoherent hologram illuminated with a laser beam. Note the two symmetrical images of the object "fourier."

the object.[331,332] This is due to the dc term in Eq. (1.110), which contributes to the noise without contributing to the signal. The SNR decreases as the number of coherence areas increases inside the aperture. For our application, this degradation must be compared to that produced by turbulence in the image plane. It turns out that incoherent holography still compares favorably as long as the object is smaller than the seeing disk. Indeed in this case the coherence area is determined by the wavefront aberrations rather than by the object and is independent of the object size. The technique starts losing its advantage when the object gets bigger than the seeing disk.[333,334]

1.3.1.4 Effect of Atmospheric Turbulence on Interferograms.

Let $\Psi(\mathbf{r})$ be the instantaneous complex transmission of the turbulent atmosphere. Instead of $A_0(\mathbf{r})$, the source will produce on the telescope entrance aperture a complex amplitude

$$A_1(\mathbf{r}) = A_0(\mathbf{r})\Psi(\mathbf{r}) \ . \tag{1.113}$$

Replacing $A_0(\mathbf{r})$ with $A_1(\mathbf{r})$ in Eqs. (1.103) and (1.104) and assuming for simplicity that $\theta = 0$ gives

$$A(\mathbf{r}) = \frac{\Omega}{\sqrt{2}}\left[\Psi(\mathbf{r})\exp\left(-2i\pi\frac{\boldsymbol{\alpha}\cdot\mathbf{r}}{\lambda}\right) + \Psi(\mathcal{R}\mathbf{r})\exp\left(-2i\pi\frac{\boldsymbol{\alpha}\cdot\mathcal{R}\mathbf{r}}{\lambda}\right)\right] \ . \tag{1.114}$$

Taking the squared modulus of Eq. (1.114) gives the illumination at the interferometer output:

$$I(\mathbf{r}) = \frac{1}{2}|\Omega|^2\left\{|\Psi(\mathbf{r})|^2 + |\Psi(\mathcal{R}\mathbf{r})|^2\right.$$

$$\left. + \Psi(\mathbf{r})\Psi^*(\mathcal{R}\mathbf{r})\exp\left[-2i\pi\frac{\boldsymbol{\alpha}\cdot(\mathcal{R}\mathbf{r}-\mathbf{r})}{\lambda}\right] + \mathrm{cc}\right\} \ , \tag{1.115}$$

where * and cc denote complex conjugates. We assume that the effect of the atmosphere described by $\Psi(\mathbf{r})$ is independent of the direction $\boldsymbol{\alpha}$ of the source (isoplanicity condition). Then, for an extended object with brightness distribution $O(\boldsymbol{\alpha})$, Eq. (1.115) becomes

$$I(\mathbf{r}) = \frac{1}{2}[|\Psi(\mathbf{r})|^2 + |\Psi(\mathcal{R}\mathbf{r})|^2]\int O(\boldsymbol{\alpha})\,d\boldsymbol{\alpha}$$

$$+ \frac{1}{2}\Psi(\mathbf{r})\Psi^*(\mathcal{R}\mathbf{r})\int O(\boldsymbol{\alpha})\exp\left[-2i\pi\frac{\boldsymbol{\alpha}\cdot(\mathcal{R}\mathbf{r}-\mathbf{r})}{\lambda}\right]d\boldsymbol{\alpha} + \mathrm{cc} \ , \tag{1.116}$$

or, in terms of the Fourier transform $\hat{O}(\mathbf{r})$ of $O(\boldsymbol{\alpha})$,

$$I(\mathbf{r}) = \frac{1}{2}[|\Psi(\mathbf{r})|^2 + |\Psi(\mathcal{R}\mathbf{r})|^2]\hat{O}(0) + \frac{1}{2}\Psi(\mathbf{r})\Psi^*(\mathcal{R}\mathbf{r})\hat{O}\left(\frac{\mathcal{R}\mathbf{r}-\mathbf{r}}{\lambda}\right) + \mathrm{cc} \ . \tag{1.117}$$

As shown in Sec. 1.3.1.3 the second term on the right-hand side of Eq. (1.117) can be isolated to reconstruct an image of the object. However, the object Fourier transform is now multiplied by a complex factor $\Psi(\mathbf{r})\Psi^*(\mathfrak{R}\mathbf{r})$. Turbulence essentially produces phase aberrations. If we neglect amplitude fluctuations (stellar scintillation), then $|\Psi| = 1$ and Eq. (1.117) shows that the modulus of the object Fourier transform is not affected by pure phase aberrations as expected from a nonredundant beam recombination technique. Only the phase of the object Fourier transform is distorted. This is illustrated in Fig. 1.22, which shows the effect of a wavefront phase distortion on a rotational shear interferogram taken with a double point source. Because the fringes produced by each of the source points undergo the same distortion (isoplanicity condition), the beat pattern (fringe amplitude) is the same as if there were no distortion. From such a single interferogram one can determine both the angular separation and the orientation of the double source, which are given by the beat frequency vector. A speckle interferometric technique would require processing thousands of speckle frames to obtain the same result. This demonstrates a clear advantage of aperture-plane interferometry over image-plane interferometry.

When observing through the atmosphere, amplitude fluctuations produced by high-altitude turbulence may affect the accuracy of the estimation of $|\hat{O}|$. In this case one still has to compute averages taken over statistically independent interferograms. A good accuracy can be obtained with typically 100 interferograms. One estimate[335] consists of averaging the squared modulus of the object Fourier transform term in Eq. (1.117):
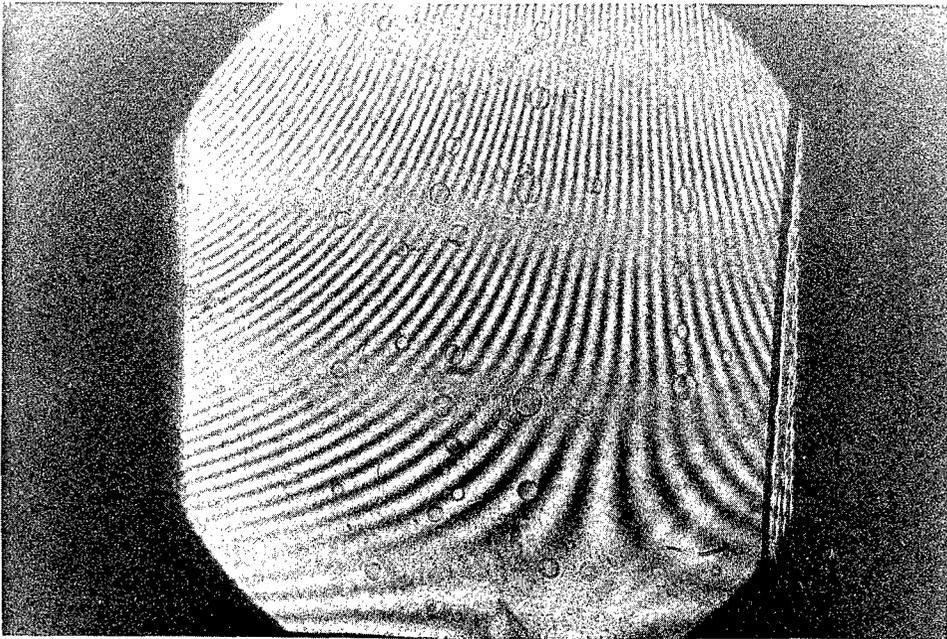
Fig. 1.22 Interferogram of a double point source distorted by severe optical aberrations. Because fringes produced by each object point undergo the same distortion, the beat pattern (fringe amplitude) remains the same as if there were no distortion.

$$\left\langle \left| \Psi(\mathbf{r})\Psi^*(\mathcal{R}\mathbf{r})\hat{O}\left(\frac{\mathcal{R}\mathbf{r} - \mathbf{r}}{\lambda}\right) \right|^2 \right\rangle = \langle|\Psi(\mathbf{r})\Psi^*(\mathcal{R}\mathbf{r})|^2\rangle \left| \hat{O}\left(\frac{\mathcal{R}\mathbf{r} - \mathbf{r}}{\lambda}\right) \right|^2 . \quad (1.118)$$

Because amplitude fluctuations are correlated only over small distances (typically less than 10 cm), the squared modulus of the object Fourier transform is multiplied everywhere by a factor equal to unity, except very close to the origin ($\mathbf{r} = 0$):

$$\langle|\Psi(\mathbf{r})\Psi^*(\mathcal{R}\mathbf{r})|^2\rangle = \langle|\Psi(\mathbf{r})|^2\rangle \cdot \langle|\Psi(\mathcal{R}\mathbf{r})|^2\rangle = 1 . \quad (1.119)$$

Another possible estimate consists of averaging the log rather than the square of the modulus[336]:

$$\left\langle \log\left| \Psi(\mathbf{r})\Psi^*(\mathcal{R}\mathbf{r})\hat{O}\left(\frac{\mathcal{R}\mathbf{r} - \mathbf{r}}{\lambda}\right) \right| \right\rangle = \langle\log|\Psi(\mathbf{r})|\rangle + \langle\log|\Psi^*(\mathcal{R}\mathbf{r})|\rangle$$

$$+ \left\langle \log\left| \hat{O}\left(\frac{\mathcal{R}\mathbf{r} - \mathbf{r}}{\lambda}\right) \right| \right\rangle . \quad (1.120)$$

To a good approximation stellar scintillation is a log-normal process,[335] and

$$\langle\log|\Psi(\mathbf{r})|\rangle = \langle\log|\Psi^*(\mathcal{R}\mathbf{r})|\rangle = 0 . \quad (1.121)$$

Taking the antilog of the result will therefore produce an unbiased estimate of the modulus of the object Fourier transform.

The SNR performance of aperture-plane interferometry has been analyzed by several authors[333,334,337–342] under different assumptions. We summarize here the results for the modulus of the object Fourier transform (fringe visibility), assuming that the object is smaller than the seeing disk. One must distinguish atmospheric noise and detection noise. The latter can be either signal dependent, such as photon shot noise in the visible, or signal independent, such as detector or background noise in the infrared. At very low light levels, when detection noise dominates over atmospheric noise, both aperture-plane and image-plane interferometry have similar performance. Aperture-plane interferometry performs slightly better near the diffraction limit, whereas image-plane interferometry performs slightly better at lower spatial frequencies. In both cases the SNR grows quickly in proportion to the signal in the visible and to the square of the signal in the infrared. At higher light level, atmospheric noise starts to dominate over detection noise, and for image-plane interferometry the SNR saturates to a constant value, whereas for aperture-plane interferometry it keeps growing as the square root of the signal in the visible and as the signal in the infrared.

1.3.2   Rotation Shearing Interferometry in the Visible

1.3.2.1   Instrumentation.   Many different configurations have been proposed for a rotation shearing interferometer.[311,313,316,317,320,343,344] Compact designs[316,320,343,344] are preferable for three reasons. First, the number of necessary adjustments is kept to a minimum, making the interferometer easier
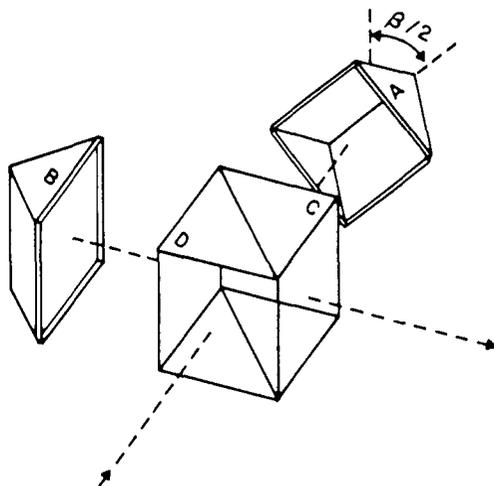
Fig. 1.23 Schematic design of a rotation shearing interferometer. C and D: beamsplitter cube; A: rotatable roof prism; B: fixed roof prism; β: shear angle.

to use. Second, it greatly reduces the sensitivity of the interferometer to mechanical vibrations; third, it eliminates stray light and light losses produced by internal reflections at air-glass interfaces.

Figure 1.23 shows a widely used configuration. It consists of a beamsplitter cube (C–D) and two roof prisms (A and B), one in each arm of the interferometer. A rotation of prism A by an angle β/2 produces a rotation of the associated image by an angle β. In Breckinridge's design,[316] the angle β/2 is equal to 90 deg producing a fixed 180-deg rotation of the associated image. Prism A and C form a single piece of glass. Prism B and D form a second piece of glass. Both pieces were cut out of cube corner reflectors. They are held in contact with an oil film at the beamsplitting interface. The optical path difference can be zeroed by sliding one prism against the other. This is the only adjustment. A drawback of this design is a mismatch in the polarization of the two interfering beams, which reduces the fringe contrast.

A similar configuration was used by Roddier et al.[320] but split into three pieces rather than two. Prism C alone is one of the three pieces. It is held at a fixed position in the mount. Roof prism A is the second piece. It is pressed against prism C with an oil film between. It is free to rotate over any angle β/2 and can also be translated. Prisms B and D form the third piece. They adhere together by optical contact and form a single block. This third piece is also pressed against prism C through another oil film. Optical path difference is adjusted by translating it against prism C. It can also slightly rotate. There are four adjustments: two translations and two rotations. Polarization mismatch can be a severe problem in this case. Although high-contrast fringes are observed at zero shear angle, the fringe visibility decreases when the shear angle increases, and fringes eventually disappear and reappear with an inverted contrast. The problem has been solved by putting quartz phase plates of appropriate thickness on the entrance face of each roof prism so that the

state of polarization at the prism output is the same as that at the prism input.[326] With this modification, high-contrast fringes are observed at any rotation angle. Figure 1.24 shows the optical parts of the interferometer, and Fig. 1.25 shows the interferometer in its mount.

A drawback of these configurations is that only one of the two interferometer outputs is used. Half of the light is reflected back to the telescope and lost. The use of both interferometer outputs doubles the system efficiency.[345] It also helps in the determination of the continuum (see Sec. 1.3.2.2.1). Mertz[343] recently proposed a design for a compact rotation shearing interferometer with two outputs, but a fixed 180-deg wavefront shear. Figure 1.26 shows a possible configuration for a compact interferometer with two outputs and a variable rotational shear.[344] It consists of two reversion prisms inserted between two Köster prisms. Rotating the reversion prisms symmetrically would change the magnification of the object Fourier transform without rotating it. Although feasible, these designs are technically challenging; no such interferometer has yet been built.
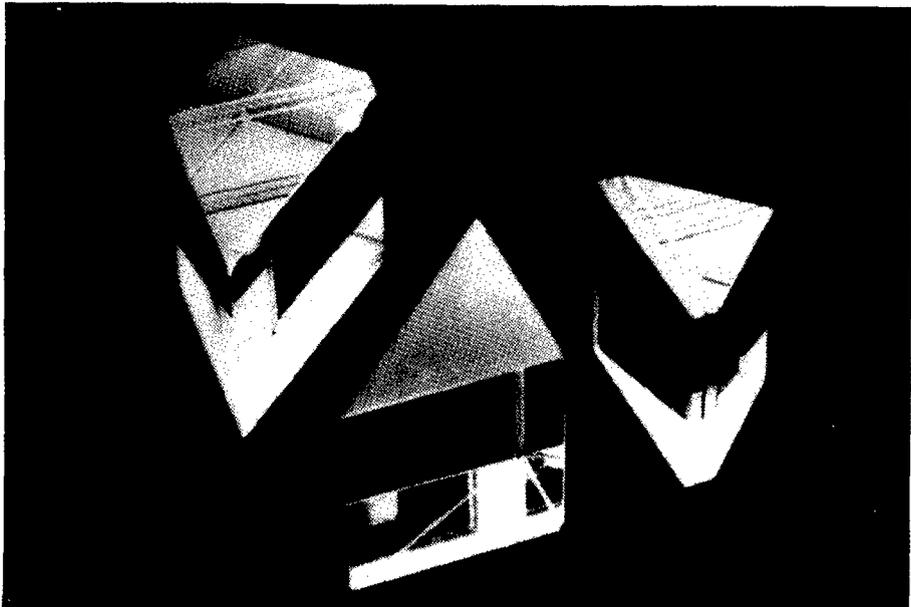
Fig. 1.24 Optical parts of a rotation-shearing interferometer. Half of the beamsplitter cube adheres to the fixed roof prism. Note the quartz phase plates on the entrance face of the roof prisms.
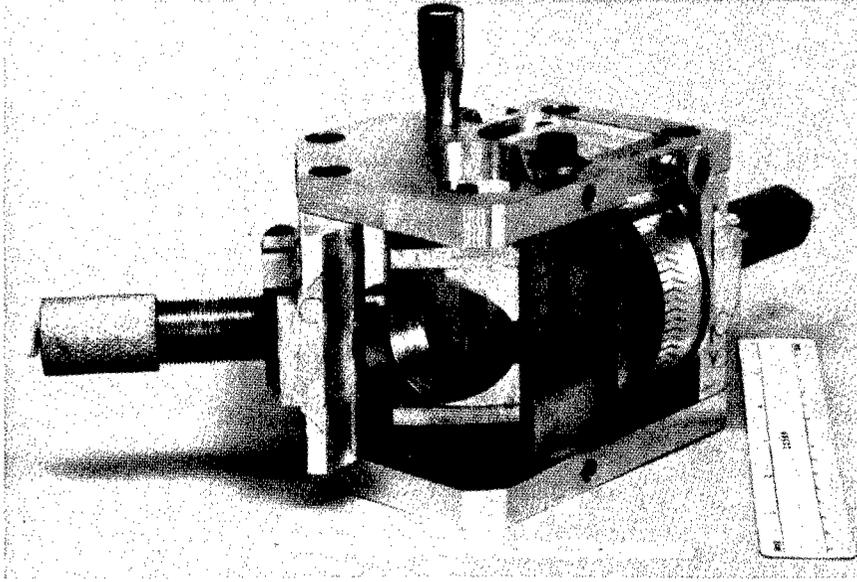
Fig. 1.25 A rotation-shearing interferometer in its mount. Long cylinders with spring-loaded pistons press together the optical parts (see on Fig. 1.24) through a steel ball, allowing the parts to move freely. Spacers are used to maintain a small space interval between the parts, which is filled with oil.
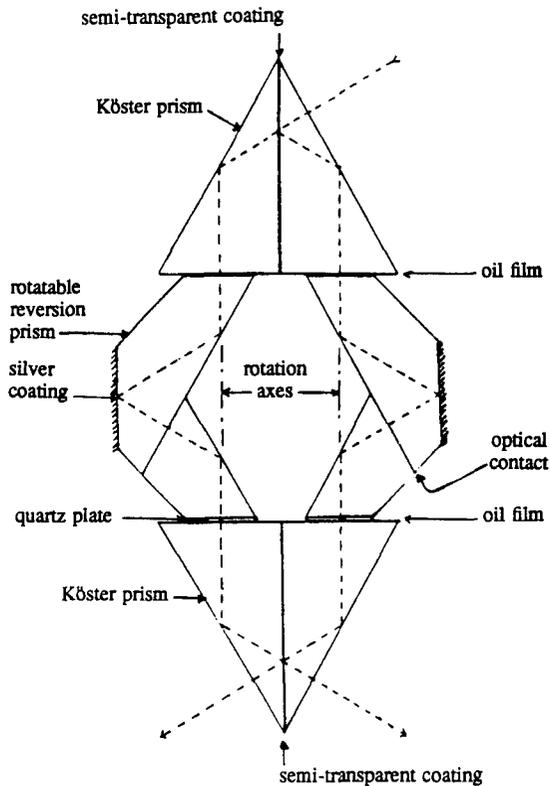
Fig. 1.26 Design for a compact rotational-shearing interferometer with a variable shear and two outputs.

1.3.2.2   Data Processing

1.3.2.2.1   Estimation of the Modulus of the Object Fourier Transform.   We saw in Sec. 1.3.1.4 that turbulence essentially affects the phase of the fringes. Hence the modulus of the object Fourier transform can in principle be estimated from the amplitude of the fringes recorded on a single interferogram. We describe here how to do so. In practice, it is advisable to average the estimated object energy spectrum over several interferograms. This will improve the SNR and reduce any effect of turbulence on the fringe amplitude as shown in Sec. 1.3.1.4.

We assume that a properly sampled and digitized off-axis interferogram has been obtained and photometrically calibrated. Any background scattered light has to be carefully estimated and subtracted. The problem is to map the fringe visibility over the interferogram. This is best done by means of Fourier transform techniques as described in Ref. 346. A two-dimensional Fourier transform produces an array of complex numbers with a central peak and two sidelobes. A window is set to isolate one of the sidelobes and this reduced data set is inverse Fourier transformed, yielding a map of complex numbers whose modulus and argument represent the amplitude and phase of the fringes. Here we consider only the amplitude. The square of the amplitude is an estimate of the object energy spectrum. When dealing with faint light sources, this estimate is biased by the photon noise spectrum. The photon noise bias can be estimated from an interferogram taken with differently oriented fringes as described in Ref. 346. The estimated noise energy (square of the amplitude) is then subtracted from the estimated object energy spectrum, yielding a bias-free estimate. The shape of the photon noise spectrum also provides an estimate of the system optical modulation transfer function, which can be calibrated out. Because the illumination in the interferogram is generally not uniform, the *dc level* or *continuum* must also be similarly mapped by setting a window on the central peak and taking the inverse Fourier transform. At each point on the pupil, the object amplitudes can then be corrected for the continuum nonuniformities.

Errors arise because the central peak and the two sidelobes always partially overlap. Indeed the illumination in the interferogram can be written

$$I(\mathbf{r}) = D(\mathbf{r})A(\mathbf{r})\{1 + \text{Re}[C(\mathbf{r})\exp(2i\pi\mathbf{f}_0 \cdot \mathbf{r})]\} \tag{1.122}$$

$$= D(\mathbf{r})A(\mathbf{r})\left[1 + \frac{1}{2}C(\mathbf{r})\exp(2i\pi\mathbf{f}_0 \cdot \mathbf{r}) + \frac{1}{2}C*(\mathbf{r})\exp(-2i\pi\mathbf{f}_0 \cdot \mathbf{r})\right] ,$$

where $D(\mathbf{r})$ describes the domain over which the interferogram extends,

$$D(\mathbf{r}) = \begin{cases} 1 & \text{inside the domain} \\ 0 & \text{outside the domain} \end{cases} ,$$

$A(\mathbf{r})$ is the dc level or continuum, and $C(\mathbf{r})$ is the complex fringe visibility with modulus equal to the fringe amplitude and argument equal to the fringe phase. The spatial frequency vector $\mathbf{f}_0$ represents the reference fringe frequency. Taking the Fourier transform of Eq. (1.122) gives

$$\hat{I}(\mathbf{f}) = \hat{D}(\mathbf{f}) * \hat{A}(\mathbf{f}) * \left[\delta(\mathbf{f}) + \frac{1}{2}\hat{C}(\mathbf{f} - \mathbf{f}_0) + \frac{1}{2}\hat{C}^*(-\mathbf{f} - \mathbf{f}_0) \right] . \qquad (1.123)$$

In the ideal case of an infinite interferogram with uniform continuum [$A(\mathbf{r})$ = $D(\mathbf{r})$ = 1], the right-hand side of Eq. (1.123) reduces to the factor between brackets. Since $\hat{C}$ has a finite support the interferogram Fourier transform displays nonoverlapping terms as long as $\mathbf{f}_0$ is greater than the cutoff frequency of $\hat{C}$. The trouble comes from the smearing introduced by the $\hat{D}$ and $\hat{A}$ convolution factors.

On high SNR data, good results have been obtained by estimating the continuum from interferogram Fourier transforms in which the side peaks have been removed and replaced with equivalent data recorded without fringes. The illumination $I(\mathbf{r})$ in the interferogram is then divided by the estimated continuum $A(\mathbf{r})$ on a frame-by-frame basis. Clearly, the use of an interferometer with two opposite phase outputs is helpful in providing directly both the ac and the dc parts of the signal.

Because $D(\mathbf{r})$ is a discontinuous function, its Fourier transform $\hat{D}$ extends widely to infinity and has a disastrous smearing effect in Eq. (1.123). A solution is to extrapolate the interferogram with a Gershberg-type algorithm.[346] In the interferogram Fourier transform, data are set to zero outside the assumed support of the $\hat{C}$ functions. The result is inverse Fourier transformed, producing an interferogram with extrapolated fringes. The original data are put back inside the domain while keeping the extrapolated data outside the domain, and the result is Fourier transformed. After a few iterations well-isolated sidelobes are obtained. One of them can easily be selected and Fourier transformed back, yielding an improved map of the fringe complex visibility.

1.3.2.2.2 *Estimation of the Phase of the Object Fourier Transform.*
Methods have been developed to determine also the phase of the object Fourier transform from atmospherically distorted pupil-plane interferograms. As we have seen, the phase obtained from a single interferogram is severely affected by turbulence. The most straightforward method consists of averaging this phase over a large number of statistically independent interferograms. However, direct averaging is not possible because the phase obtained from a Fourier transform is only defined modulo 2π. It is called a *wrapped* phase. On high SNR data, the phase can be unwrapped using a two-dimensional phase unwrapping algorithm, and the unwrapped phase can be averaged. This method was successfully used in the 1970s to produce the first image reconstruction from rotational shear interferograms taken through laboratory-generated turbulence.[319,347,348] A similar experiment was performed by Itoh and Ohtsuka[307,317,349,350] in the early 1980s. In this case the phase at each frequency was followed and unwrapped as a function of time using a one-dimensional unwrapping algorithm described by Mertz.[351] The result was then averaged.

Rather than averaging unwrapped phases, an algorithm that is more robust to noise consists of considering the phase difference $\Delta\phi$ between two consecutive samples in the frequency plane and averaging the phasor $\exp(i\Delta\phi)$ over a large number of interferograms. The argument of the average phasors is an estimate of the gradient of the object phase. The phase is reconstructed from its gradient by integration.[352] Least-squares integration algorithms have been

widely described in the literature especially in relation to wavefront estimation in adaptive optics.[353,354] These algorithms have to be modified because the phase of an object Fourier transform is likely to contain poles.[355] This technique is now being applied to infrared interferograms (see Sec. 1.3.3).

The preceding methods have the drawbacks of requiring averages over many independent frames and being sensitive to telescope aberrations. The same procedure has to be repeated on a reference point source to calibrate out telescope aberrations. As noted by Ribak,[356, 357] phase closure relations can be extracted from a pupil-plane interferogram when 360 deg is an integral multiple of the shear angle. These relations can be used to enhance the image reconstruction process,[357–359] but they are not sufficient to retrieve the object phase entirely. A method was proposed by Hofmann and Weigelt to fully retrieve the object phase from a single snapshot.[360,361] It is called *triple shearing interferometry* and consists of recording interference produced by three overlapping pupil images with different angular shears. An equivalent but simpler method,[344,362] which has been tested in the laboratory, consists of recording at the same time two interferograms, one with a shear α, the other with a shear 2α. The object phase can be recovered either by solving the phase closure equations explicitly or by using *self-calibration techniques* similar to those developed by radioastronomers.[363–365]

1.3.2.3 Astronomical Results

Rotation shearing interferometry has been applied to the study of the red supergiant star Betelgeuse (α-Ori). This star has the largest angular diameter in the sky and in 1921 was the first star to be resolved interferometrically by Michelson and Pease.[366] Rotation shearing interferograms of Betelgeuse were taken in 1980 at the Canada-France-Hawaii telescope on Mauna Kea. Interferograms were recorded on photographic films through a four-stage magnetically focused EMI image intensifier. Figure 1.27 shows an example of such an interferogram. Exposure times of $1/30$ and $1/60$ of a second were used to freeze the effect of atmospheric turbulence. Different spectral windows were used in the continuous spectrum and in the TiO bands. Evidence was found for variations of the stellar diameter with wavelength.[367]

A total of 71 frames, taken at wavelength $\lambda = 535$ nm, were scanned with a PDS microdensitometer and photometrically calibrated. A two-dimensional map of the fringe visibility was obtained as described in Sec. 1.3.2.2. It shows evidence for a deviation from circular symmetry. The result was interpreted as being produced by the combined effect of a bright spot on the stellar disk plus a halo of scattered light, indicating the presence of a close dust envelope.[368] The existence of a bright spot on the disk was recently confirmed by the nonredundant pupil masking method.[283,285]

An attempt was made to reconstruct an image of the star using a phaseless maximum entropy algorithm.[369] The reconstructed star image is displayed in Fig. 1.28. It shows evidence for an asymmetric dust condensation at about 2.5 stellar radii from the star center. Because the noisy high spatial frequency structures were filtered out in the reconstruction process, the bright spot did not appear on the stellar disk.

Further astronomical observations[370] made with the rotation shearing interferometer revealed a time evolution of the dust envelope and the possible existence of a companion to α-Ori.

Fig. 1.27 Short-exposure rotational-shear interferogram recorded in the visible (534-nm) region at the CFH telescope. The source was α-Ori and the shear angle was 30 deg.
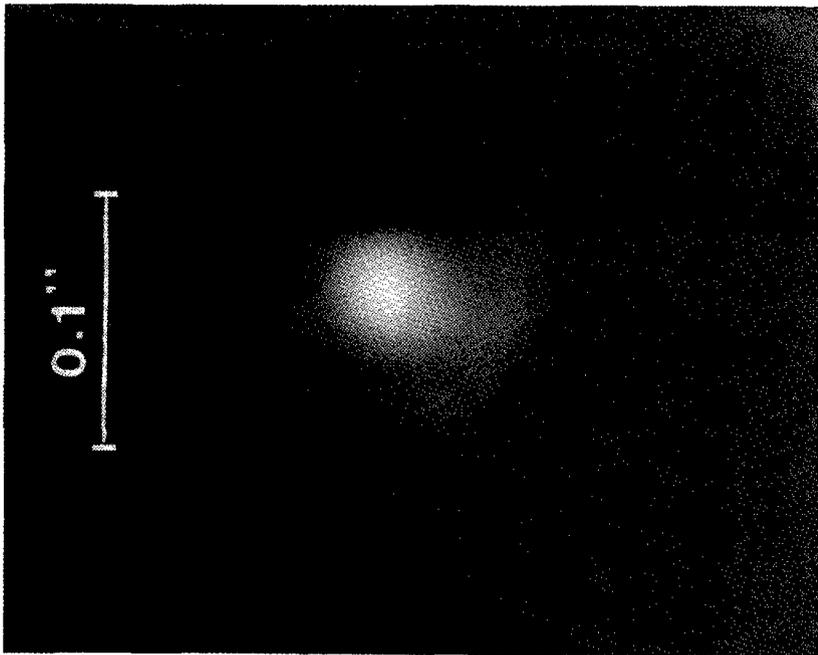
Fig. 1.28 Image of α-Ori reconstructed from rotational-shear interferograms. The reconstruction was made without phase information, using a maximum entropy algorithm. It reveals light scattered by a highly asymmetric dust envelope.

1.3.3   Rotation Shearing Interferometry in the Infrared

1.3.3.1   Instrumentation

1.3.3.1.1   The Infrared Case.   Theoretically, the near infrared (NIR) and thermal infrared (TIR) domains are well suited to exploit the benefits of aperture-plane interferometry. This is due to two main reasons:

1. In contrast to the visible domain, the dominant source of noise is no longer the photon noise from the source, but rather the detector noise (NIR) or the photon noise of the thermal background (TIR). The noise is now independent of the source. In the speckle-dominated regime (number of photons per frame and per coherence cell $\gg$ 1), this implies that the SNR is proportional to the number of photons per frame and per coherence cell, whereas for the visible domain it is proportional to the square root of this number of photons (see Sec. 1.3.1.4).

2. It is well known that the effect of turbulence on image formation is strongly wavelength dependent.[335] Both the Fried parameter r_0 and the coherence time τ_0 increase with the wavelength, which considerably favors the IR domain with respect to the visible.

However, the detection of the interferometric signal is far less efficient in the infrared. In the TIR, one faces a physical limit, namely, the photon noise of the huge thermal background: The number of photons per frame and per coherence area delivered by the source sets the instantaneous SNR and there is no way to improve this quantity significantly, at least for ground-based observations. On the other hand, the situation is different in the NIR: A critical parameter is σ_r/η, the ratio of the readout noise to the quantum efficiency of the detector. Modern IR array detectors feature excellent quantum efficiency ($>70\%$) and readout noise of the order of a few tens of electrons per pixel, although this performance can be degraded when the array is operated with the very short exposure times mandatory in interferometry. Yet, this is far from the photon counting operating mode usually reached in the visible.

As is already the case for high angular resolution image-plane techniques, the limit between the two domains, NIR and TIR, is shifted for aperture-plane interferometry toward longer wavelengths, typically 5 µm. This shift is the consequence of the large reduction of the background level due to the limited spectral bandwidth, the very short exposure time ($\ll \tau_0$) and the very limited beam étendue (throughput) per pixel ($\ll \lambda^2$). Since at the wavelength of 5 µm and above, even the largest optical telescopes are close to being naturally diffraction limited, one can see that infrared aperture-plane interferometry works mostly in the detector noise-limited regime.

1.3.3.1.2   The Time Encoding Setup.   The immediate consequence is that, in order to keep the detector noise as low as possible, IR pupil-plane interferometers should be implemented so as to minimize the number of pixels used to analyze the interferometric signal. An attractive solution is to fluff out the fringes, that is, to widen the fringe spacing until they disappear. This is done by setting the source on the axis of the interferometer. Defining p as the equivalent pixel pitch, i.e., the equivalent size of a pixel at the telescope pupil,

the sampling condition for the wavefront roughly corresponds to $p \leq r_0/2$. With respect to the off-axis setup for which we assume that the off axis yields about three fringes per coherence area, the gain in the total number of pixels is 3×3.

For this setup, of course, the estimation of the degree of coherence of the source through the two-dimensional analysis of the fringe contrast is no longer possible. A third dimension is required to encode the interferometric signal, namely, the time. This leads to the concept of phase shifting interferometry (PSI): the optical path difference (OPD) δ inside the interferometer is modulated in discrete and equal fractional increments of the wavelength

$$\delta_i = \delta_0 + \frac{i-1}{n}\lambda \qquad \text{for } i = 1, 2, ..., n \ . \tag{1.124}$$

For each step i, a frame is recorded. Then the process is repeated in sequence. From each set of n successive frames, it is straightforward to recover the modulus and the phase at each point of the interference field, i.e., in this case, at each spatial frequency. The sampling theorem imposes $n > 2$. On the other hand, n should not be too large since the total readout noise of the sequence is proportional to n, and since the whole sequence has to be completed in a total time shorter or of the order of τ_0. Hence, $n = 3$ or 4 appear as reasonable compromises.

A prototype infrared rotation-shearing phase shifting interferometer for astronomical observations has recently been described.[322–325] The implementation of this prototype is based on a Michelson interferometer that provides the pupil rotation with roof mirrors installed on both arms. Figure 1.29 shows the optical setup of this interferometer. One of these mirrors can be rotated in order to control the shear angle. In practice, a 180-deg shear is used in order to have access to the full spatial frequency coverage up to the diffraction limit of the telescope. It turns out that for this shear angle the contrast of the fringes is high (>0.9), thus avoiding the need for polarization compensating plates. When a different shear angle is used, the reduction of the fringe contrast is indeed observed. A grid polarizer can be used to restore the full contrast by selecting one polarization.

The other roof mirror is mounted on a piezo-stack, which provides the OPD modulation. For this instrument, the range is covered in four steps, i.e., the successive frames are in phase quadrature.

The optical elements of the interferometer are all standard components. Figure 1.30 shows the Michelson interferometer and the dewar containing the infrared camera. This camera uses a 32×32 CdHgTe diode array bump-bonded to a standard silicon CCD (IRCCD). One important advantage of using a CCD for 2-D interferometry is its ability to capture one complete image at a time, therefore freezing the turbulence in the same state over the whole image, without using any mechanical shutter in front of the detector. The detector's quantum efficiency is rather high ($\approx 60\%$ at 3.5 μm) and it can store $\approx 2 \times 10^7$ electrons, allowing long integration times to be used. A description of the IR camera can be found in Monin et al.[371] where it is used in a classical imaging version.
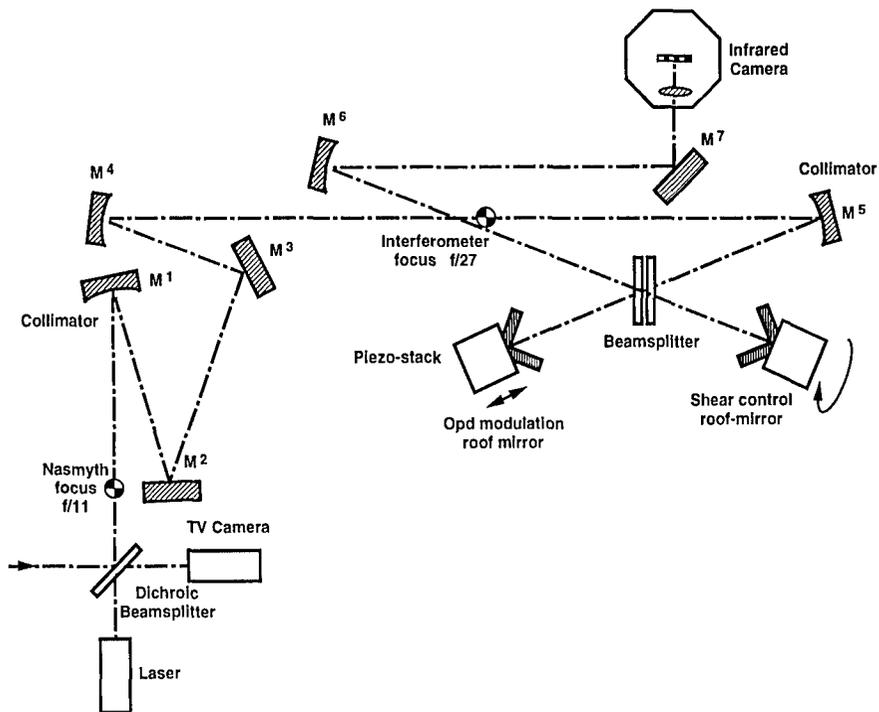
Fig. 1.29 Optical setup for an infrared rotation-shearing phase shifting interferometer. (After Ref. 325)
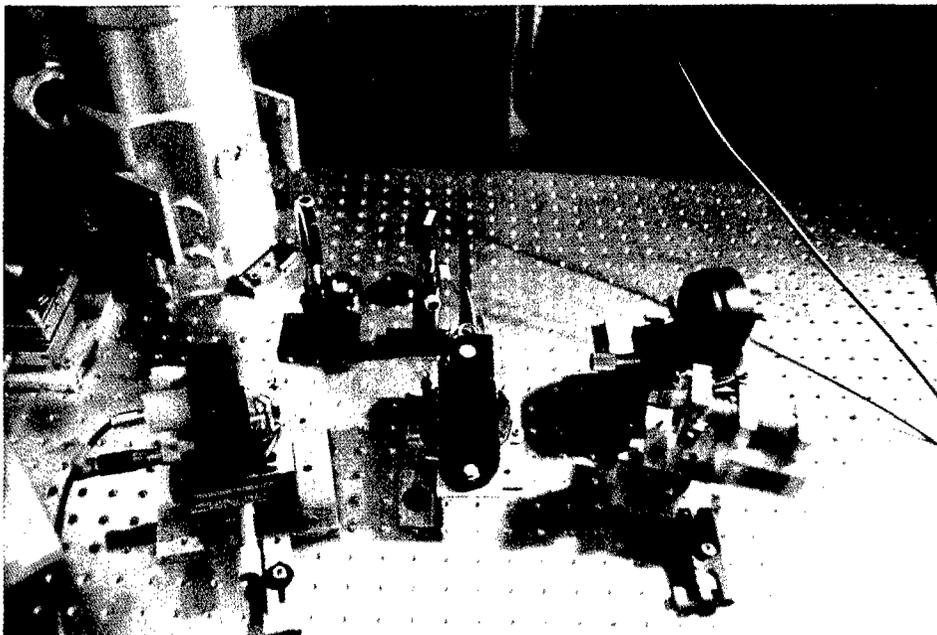
Fig. 1.30 The central part of the infrared rotation-shearing interferometer. The beamsplitter stands in the middle and the IR camera in the background at left.

1.3.3.1.3   The Space Encoding Setup.   The previous setup has been designed to work in the flat-tint mode. If the source is set slightly off-axis, equal thickness fringes appear in the pupil image, and the interferometer can be operated as in the visible. The only condition to meet is a sampling condition: For a 4-m telescope and a Fried parameter $r_0 = 40$ cm (a typical value at 2.2 μm), and assuming that we choose to sample a coherence cell with three fringes, we need to image the pupil onto approximately 60×60 pixels. Since r_0 is only a statistical expectation for the mean size of a coherence cell, a 128×128 detector array would be a reasonable choice. State-of-the-art infrared detector arrays now exist in this format and even 256×256. Hence, spatial encoding of the interferometric signal is possible for infrared rotation-shearing interferometers with telescope diameters up to 8 m.

Except for the detector, an interferometer of that kind would be very similar to its visible counterpart, including the data reduction methodology. To our knowledge, no such system has yet been built and we do not discuss it further.

1.3.3.2   Data Reduction.

Data reduction procedures for infrared rotation-shearing phase shifting interferometry have been described by Mariotti et al.[325] and Monin et al.[372] The first step, of course, is to apply the classical corrections mandatory for infrared imaging, e.g., flat-fielding, corrections for bad pixels, and estimation of the level of the thermal background. Then the procedures of PSI can be applied.

1.3.3.2.1   Estimation of the Modulus.   It is straightforward to see that if the interference field is stable, a simple combination of the n phase-shifted frames yields an estimation of the phasor amplitude A for each pixel of the detector, i.e., for each spatial frequency. For instance, in the case $n = 4$, we have[325]

$$A = \left\{ \left[\frac{1}{2}(I_1 - I_3) \right]^2 + \left[\frac{1}{2}(I_4 - I_2) \right]^2 \right\}^{1/2}, \tag{1.125}$$

where I_i denotes the signal in the i'th frame. Figure 1.31 illustrates the estimation of the modulated interferometric signal from the raw data: It displays side-by-side series of two *instantaneous quadrature images* $(I_1 - I_3)$ and $(I_4 - I_2)$. They correspond to the *sine* and the *cosine* components of the interferometric signal from which A is estimated.

Similarly, the mean number of photons is given by a mere summation, so that the modulus of the visibility function is simply given by

$$V = \frac{A}{\frac{1}{n} \sum_{i=1}^{n} I_i}. \tag{1.126}$$

Hence, in theory, the estimation of the source visibility is extremely simple and should be performed easily in real time. Note also that the experiment is self-calibrated: Variations in the transmission of the setup, or the differences in sensitivity of the detector pixels cancel out in the expression for V. The only assumption is that the response of each pixel of the detector versus flux is linear near its operating point. However, it is often wise to make an estimation of the actual MTF of the setup, for instance, by performing observations of
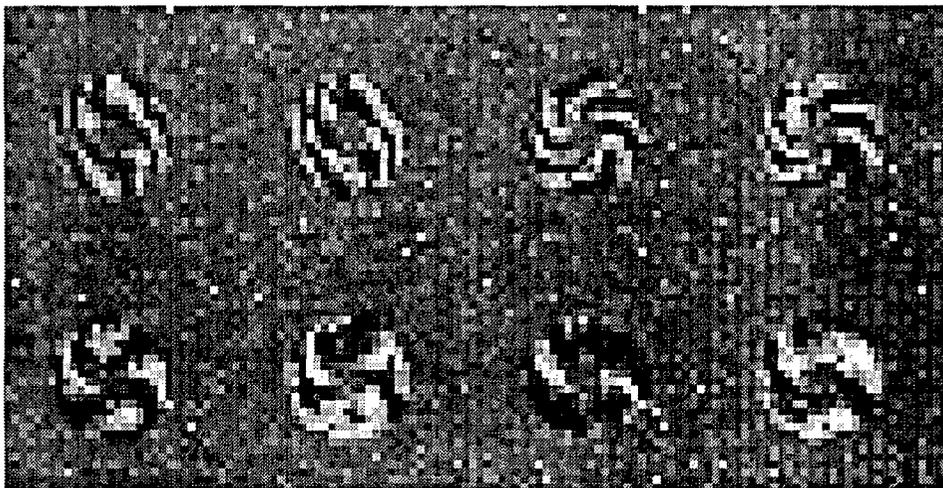
Fig. 1.31   Pairs of fringe patterns taken in quadrature at four different instants at the wavelength of 3.8 μm on the star β Peg. (After Ref. 325)

unresolved sources. Note, however, that the conditions for this calibration are far less stringent than for speckle interferometry.

In practice, the final data reduction is more complex, especially when the instantaneous SNR is weak.[325] Both the above expressions for A and V are biased, the first one by the variance of the additive noise, the second by the number of photons emitted by the background. These biases are easily estimated and corrected for by performing an observation of the nearby sky close to the source in the same experimental conditions, a procedure also mandatory for speckle interferometric observations.

A more complex matter to handle is the detection of undersampled frames. The deformations of the wavefront have a typical size of r_0, but can be much smaller for some frames, possibly smaller than twice the equivalent pixel pitch, in which case they are not correctly sampled by the detector. This phenomenon creates another bias leading to an underestimation of the fringe contrast. This bias is difficult to correct. The best solution is to detect and discard these spatially undersampled frames. Similarly, although each phase shifting sequence is scanned in a time shorter than τ_0, for some sequences the frames can be decorrelated, leading to an erroneous estimation of the contrast. Again, the best solution is to reject these undersampled time sequences. Quality estimators of the data have been developed, which allow one to solve these problems for the final data reduction.[325] Unfortunately, the procedure is iterative and cannot be easily implemented in real time.

1.3.3.2.2   Estimation of the Phase.   While the MTF of a pupil-plane interferometer is not affected by the turbulence, the estimation of the phase is. In the case of the phase shifting interferometer with four steps, it is straightforward to estimate an instantaneous phase Y at each spatial frequency

$$Y = \tan^{-1}\left(\frac{I_4 - I_2}{I_1 - I_3}\right) , \qquad (1.127)$$
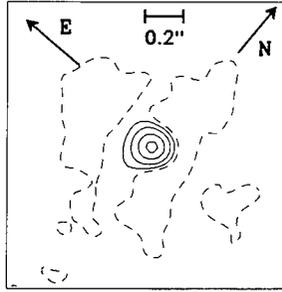
Fig. 1.32   Image of NML Cygni at 3.8 μm obtained by IR rotation-shearing interferometry and deconvolved with a regularization method. Contours are $-\sigma$, 3σ, 6σ to peak value by steps of 6σ (σ is the rms noise value on a sky image containing no object). (After Ref. 352)

where Y is the wrapped value of "true" phase Φ. Even if it were possible to unwrap the estimator Y, for instance, by following it in time and detecting the $\pm 2\pi$ jumps, Φ is strongly dominated, for spatial frequencies larger than f_0, by the atmospheric-induced phase errors. Fortunately, several solutions for the estimation of the phase exist. The method used for the previously described infrared interferometer relies on the spatial correlation of the phase for adjacent frequencies.[372] The idea is to average the spatial gradients of the phase in a spirit similar to the Knox and Thompson method.[373] The contribution of the atmospheric phase in the phase gradients is small because $p < r_0$ and averages out rapidly. The next step involves the reconstruction of the average phase from its gradients by a minimal norm least-squares fit. Finally, the "internal" phase errors corresponding to the stable aberrations of the setup and the telescope are calibrated out by subtraction of the phase map obtained in similar experimental conditions on an unresolved astronomical source. Again, the conditions for this calibration are far less stringent than for image-plane interferometry. The reader is referred to Refs. 352 and 372 for a detailed description of the phase restoration and image reconstruction procedures. Figure 1.32 shows the reconstructed image of a giant star with a circumstellar shell. As illustrated in this figure, this interferometer is primarily used for observations of marginally resolved sources. In such conditions, neither the lack of low spatial frequency information due to the telescope central obscuration, nor the trace of the edges of the roof mirrors in the visibility map are problematic. In most cases, it is not even necessary to interpolate the complex visibility data to fill these gaps, since the image reconstruction algorithm is sufficiently constrained by the high spatial frequencies to converge to a correct solution.

1.3.3.3 Astronomical Results.   As already stated, the only rotation-shearing infrared interferometer is a prototype instrument. Its aim was to demonstrate the feasibility of aperture-plane interferometry at infrared wavelengths and its ability to produce diffraction-limited images. Both these goals have been reached, opening the way to the application of the technique to the observation of astronomical sources. Since the instrument was not optimized for sensitivity, it cannot compete with image-plane techniques (e.g., speckle-interferometry, adaptive optics) in terms of high-resolution imaging of faint sources. Rather,

its advantage over these techniques lies in its high-valued transfer function near the cutoff frequency D/λ. Aperture-plane interferometry is therefore well adapted to observe objects for which the important information lies at high spatial frequencies, i.e., to measure the characteristic size of marginally resolved sources.

Some evolved stars undergo a process of large mass loss, creating a thick circumstellar envelope. Dust grains can form in this envelope at a few stellar radii from the surface of the star, a zone where the temperature has dropped below the sublimation point. This dust radiates in the infrared, creating the so-called "IR excess." The observation of the spatial structure of the shell and of the temperature and the density of the dust can help us to understand the mass loss phenomenon, the dynamics of the envelopes, and the interaction with the interstellar medium. At typical galactic distances, the expected size for these IR circumstellar emissions ranges from 1 to 10^{-2} arcsec, and falls close to the diffraction limit of a 4-m class telescope at 3.8 μm. Indeed, a dozen of these shells have been recently observed at the 4.2-m William Herschel Telescope (La Palma, Canary Islands). Several have been resolved[374] at 3.8 μm, and diffraction-limited images of NML Cyg (see Fig. 1.32) and O Cet have been reconstructed. The astrophysical interpretation of these images is in progress.

1.3.4 Summary and Future Prospects

Although more difficult to implement than speckle interferometry, aperture-plane interferometry has been successfully demonstrated at both visible and infrared wavelengths. The technique was found to be notably effective on bright sources, when in the image plane speckle noise dominates over detector noise. Because information is recorded in the frequency domain, the transfer function of the atmosphere and of the telescope are better estimated and calibrated out. Aperture-plane beam recombination techniques have also been considered in long-baseline interferometry.[336,342,375]

Aperture-plane interferometry applies to the reconstruction not only of turbulence-degraded images, but also of images degraded by permanent aberrations.[364] Large lightweight telescopes are now being envisaged in space for millimetric and infrared applications. Aperture-plane interferometry could be applied to extend their resolution toward shorter wavelengths.[341,376,377] This interferometric imaging technique might as well be called Fourier transform imaging by analogy with Fourier transform spectroscopy. Indeed the two techniques are similar and can be combined to provide both angular and spectral resolution with the same interferometer. The method involves three-dimensional Fourier transforms. It has been described and extensively investigated in the laboratory by Itoh and Ohtsuka.[378–381] It has been independently proposed and investigated on a telescope by Mariotti and Ridgway[382,383] using a modified Fourier transform spectrometer.

A real-time image compensation technique called adaptive optics is now starting to be applied to astronomical telescopes and one might wonder whether postdetection processing techniques such as speckle-interferometry or aperture-plane interferometry have any future. Adaptive optics is a very efficient way to compensate for the effect of large-scale wavefront errors, but it

requires very bright on-axis reference sources or guide stars to produce fully diffraction-limited images. Most of the time only partial compensation will be achieved and postdetection processing techniques will still be required to further enhance image resolution. Rotation-shearing interferometry appears to provide a promising means to calibrate accurately the effects of turbulence on partially compensated images and reach the diffraction limit with the highest SNR. Moreover, whenever adaptive optics brings the wavefront rms phase error sufficiently below 2π, it becomes possible to record and process long-exposure interferograms. This may well open for the first time the possibility of producing very high resolution images of very faint sources.

1.4 PASSIVE INTERFEROMETRIC RANGE-ANGLE IMAGING
Written by Anthony M. Tai

Passive interferometric sensors have long been employed for the imaging of celestial objects. Radio telescope is a prime example. Interferometric method allows the synthesis of a large imaging aperture using widely separated pairs of receiving apertures.[384] The effective image resolution is determined by the separation of the apertures instead of the size of the individual apertures. In recent years, substantial progress has also been achieved in the implementation of interferometric sensors at optical wavelengths as described in the previous section. Most interferometric sensors, whether they operate in the optical or microwave regions, are ground-based sensors and they produce images in angle-angle coordinates.

Another well-developed sensing technique to create high-resolution images utilizes an active coherent sensor as exemplified by the synthetic aperture radar (SAR), which has been operational for more than three decades. This form of sensor actively transmits a coherent beam and detects the complex amplitude of the returned electromagnetic field. Unlike passive interferometric sensors, SAR systems are typically operated from airborne platforms to image ground targets, and the images they produce are in range-angle coordinates. In this section, the two synthetic aperture imaging concepts are combined to produce a class of passive interferometric imaging sensors that image in range-angle coordinates. Unlike other interferometric sensors, which are designed primarily for astronomical imaging, this new class of sensors can also be used to image ground targets from airborne sensor platforms.

A passive interferometric sensor measures the mutual coherence of the received light field emanating from a far-field incoherent source. The mutual coherence or cross-correlation function is related to the intensity distribution of the source via a Fourier transformation. In Sec. 1.4.1, we see that the Fourier transform relationship is three-dimensional in nature. Depending on how the imaging aperture is synthesized, the resulting image can be in two-dimensional angle-angle or range-angle coordinates. If enough degrees of freedom are avail-

able in the sensor imaging geometry, passive three-dimensional synthetic aperture interferometric imaging is also possible.

A synthetic aperture radar[385,386] typically operates in a side-looking mode with a fairly shallow grazing angle. It produces high-resolution images in range-angle coordinates. The two-dimensional image resembles an angle-angle image of the scene, which is illuminated from the side and viewed from above. In Sec. 1.4.2, an airborne two-aperture interferometric imaging sensor operating in a similar side-looking imaging mode is described. Extension to three-dimensional interferometric imaging is also discussed.

A sensor designed to image ground targets must have a substantially wider field of view than a sensor intended for astronomical imaging. With a conventional interferometer, the spectral bandwidth of the cross-correlation measurements must be restricted to accommodate the path difference from an off-axis source to the two receiving apertures. If the field of view is large, the measurement spectral bandwidth becomes very narrow. With the source-emitting broadband light, the optical efficiency of a conventional sensor is very low. In Sec. 1.4.3, the implementation of a grating-based two-aperture interferometric sensor is described. The achromatic grating interferometer allows the cross-correlation of broadband radiation from a relatively large object scene to be measured. In addition, the grating interferometer allows cross-correlation measurements to be performed simultaneously at different spectral bands, which, for a fixed baseline interferometric sensor, correspond to different spatial frequencies. This feature is crucial to the implementation of the passive range-angle interferometric imaging concept.

Finally, the issues associated with the implementation of an airborne sensor such as phase errors induced by air turbulence and platform motion are addressed in Sec. 1.4.4. While technically challenging, the implementation of an airborne passive synthetic aperture interferometric imaging sensor is shown to be feasible and can be accomplished with much less difficulty than a coherent optical synthetic aperture laser radar.

1.4.1 Basic Concept

The underlying theory of interferometric imaging is well known and it has long been applied in the field of astronomical imaging. The basic concept is illustrated in Fig. 1.33. Consider a two-aperture interferometer. The electro-
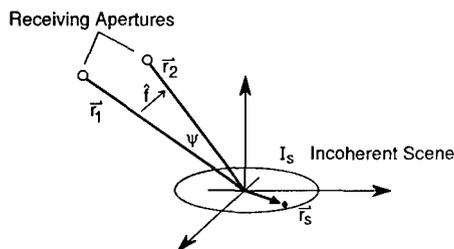
Fig. 1.33 Imaging geometry with a two-aperture interferometric sensor.

magnetic far-field wavefronts $u_1(\mathbf{r}_1)$ and $u_2(\mathbf{r}_2)$ received at the sensor input apertures can be related to the source scene field $u_s(\mathbf{r}_s)$ via

$$u_1(\mathbf{r}_1) = \frac{\exp[(i2\pi r_1)/\lambda]}{i\lambda r_1} \int u_s(\mathbf{r}_s) \exp\left[\frac{-i2\pi(\hat{r}_1 \cdot \mathbf{r}_s)}{\lambda}\right] d\mathbf{r}_s , \qquad (1.128)$$

and

$$u_2(\mathbf{r}_2) = \frac{\exp[(i2\pi r_2)/\lambda]}{i\lambda r_2} \int u_s(\mathbf{r}_s) \exp\left[\frac{-i2\pi(\hat{r}_2 \cdot \mathbf{r}_s)}{\lambda}\right] d\mathbf{r}_s . \qquad (1.129)$$

The coordinate geometry shown has its origin located at a point in the scene that is defined as the *phase reference point*. Position vector $\mathbf{r}_s$ is for an arbitrary point in the scene and vectors $\mathbf{r}_1$ and $\mathbf{r}_2$ define the location of the interferometer apertures as related to the phase reference point; $\hat{r}$ denotes the unit vector and r the magnitude of $\mathbf{r}$. If the fields received at the two apertures are correlated (or made to interfere), the cross-correlation output of the received field samples can be expressed as[387]

$$\mu_{12} = \langle u_1(\mathbf{r}_1) u_2^*(\mathbf{r}_2) \rangle$$

$$= \frac{\exp\{[i2\pi(r_1-r_2)]/\lambda\}}{\lambda^2 r_1 r_2} \int I_s(\mathbf{r}_s) \exp\left[\frac{-i2\pi(\hat{r}_1-\hat{r}_2) \cdot \mathbf{r}_s}{\lambda}\right] d\mathbf{r}_s , \qquad (1.130)$$

where $\langle \cdot \rangle$ represents the ensemble average over all source realizations and $I_s(\mathbf{r}_s) = \langle |u_s(\mathbf{r}_s)|^2 \rangle$.

The expression describes a three-dimensional Fourier transform relationship between the intensity distribution of a spatially incoherent source and the correlation output of the two-aperture interferometer. The correlation output measured at wavelength λ represents a single Fourier coefficient of the three-dimensional object spectrum, evaluated at spatial frequency

$$\mathbf{f} = \frac{(\hat{r}_1 - \hat{r}_2)}{\lambda} = \frac{2\hat{f}}{\lambda} \sin\left(\frac{\alpha}{2}\right) , \qquad (1.131)$$

where α is the angle subtended by range unit vectors $\hat{r}_1$ and $\hat{r}_2$. The direction of the spatial frequency vector $\mathbf{f}$ in the three-dimensional Fourier space is defined by the sensor line of sight to the phase reference point in the object scene and the orientation of the baseline that connects the two apertures of the interferometer. It can be changed by varying the orientation of the baseline and/or the line of sight of the interferometric sensor to the phase reference point. The magnitude of the spatial frequency vector $f = |\mathbf{f}|$, on the other hand, can be altered by increasing or decreasing the separation of the sensor apertures and/or the measurement wavelength. If enough degrees of freedom are available to vary the direction and length of the spatial frequency vector, a full three-dimensional synthetic imaging aperture can be generated. However, complete freedom in changing the spatial frequency vector is usually not available, and aperture synthesis is limited to two dimensions. In astronomical
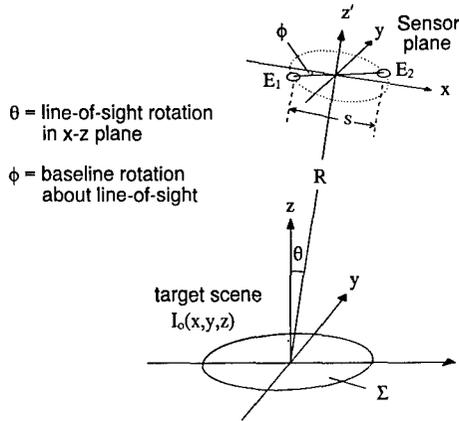
Fig. 1.34 Imaging geometry expressed in Cartesian coordinates.

imaging, for example, the coherence measurements at various spatial frequencies are obtained with the use of different baselines or shear distances as described in the previous section. The direction of the spatial frequency vector is changed by rotating the sensor baseline about the optical axis. The rotation may be accomplished with a rotation shearing interferometer[388-390] or, in the case of long baseline interferometric imaging, by utilizing the earth's rotation.[391] The result is a two-dimensional imaging aperture in angle-angle coordinates.

Of special interest in this chapter is the use of an airborne two-aperture interferometric sensor for passive synthetic aperture imaging. Unlike a sensor for astronomical imaging where the line of sight to the object is fixed, an airborne sensor can fly around the object scene. With this added freedom in changing the direction of the spatial frequency vector, a synthetic aperture in range-angle coordinates can be obtained.

The generation of a synthetic aperture in the range direction is more clearly shown by expressing the three-dimensional Fourier transform relationship described in Eq. (1.130) in Cartesian coordinates. Consider the imaging geometry illustrated in Fig. 1.34. The angle between the sensor baseline and the x'-z' plane is denoted as ϕ, and the angle between the sensor line of sight and the y-z plane is given by θ. The Fourier transform relationship in Eq. (1.131) can then be written as

$$\mu_{12} = K \int\int\int_{\Sigma} I_S(x,y,z) \exp[-i2\pi f(\cos\theta \ \cos\phi\hat{x}$$

$$+ \cos\theta \ \sin\phi\hat{y} + \sin\theta \ \cos\phi\hat{z})] \ dx \ dy \ dz \ , \tag{1.132}$$

where K is a complex constant and Σ represents the instantaneous field of view (IFOV) of the sensor. For $R \gg S$, $f = (2/\lambda) \sin(\psi/2) \cong S/R\lambda$ where R is the distance to the object scene and S is the length of the baseline. If the sensor

baseline lies on the x'-z' plane, then $\phi = 0$ and the cross-correlation of the fields at the two apertures can be simplified to

$$\mu_{12} = K \iiint\limits_{\Sigma} I_S(x,y,z) \exp(-i2\pi x f \cos\theta) \exp(-i2\pi z f \sin\theta) \, dx \, dy \, dz \ .$$

(1.133)

In the Fourier transform kernel, the spatial frequency is a function of $\sin\theta$ in the z direction. For small θ, $\sin\theta \approx \theta$, the spatial frequency is approximately a linear function of θ. The viewing angle θ can be changed by moving the sensor along an arc of radius R in the x-z plane. Rotating the viewing angle from $-\Delta\theta/2$ to $+\Delta\theta/2$ and continuously measuring the correlation outputs of the sensor, a synthetic aperture in the range direction can be generated. The synthesized imaging aperture spans spatial frequencies from $S \sin(-\Delta\theta/2)/R\lambda$ to $S \sin(\Delta\theta/2)/R\lambda$ in the z or range direction. In the following section, we examine a specific imaging geometry that produces a range-angle fine-resolution image similar in character to that of a spotlight synthetic aperture radar.[386]

1.4.2   Imaging Mode

There are practical restrictions on the size and weight of a sensor that is installed on an airborne platform. Imaging at long-wave infrared from a long distance, it may not be possible to construct a real aperture sensor with adequate resolution due to the large imaging aperture required. The implementation of a multiple-aperture interferometer on an airborne platform is also impractical because of the number of openings required on the fuselage. In this section, a side-looking staring imaging mode is described with which high-resolution two-dimensional images in range-azimuth coordinates can be produced with a two-aperture interferometric sensor.

The sensor operates in an imaging geometry similar to that of a spotlight SAR. The airborne two-aperture sensor flies at a stand-off distance and stares at a fixed patch of ground as illustrated in Fig. 1.35. As the sensor platform flies past the object scene, the viewing angle to a fixed point in the object scene is changed and the spatial frequency vector is rotated. Assuming that the sensor platform flies along a straight line and the sensor images the target broadside, the target range as a function of the viewing angle is equal to $R(\theta) = R_0/\cos(\theta)$. If the baseline of the sensor is also rotated synchronously such that it remains perpendicular to the line of sight, the measurement spatial frequencies in the x and z directions are then given by $S\cos(\theta)/R_0\lambda$ and $S\cos(\theta)\sin(\theta)/R_0\lambda$, respectively. Baseline rotation, however, is practical only if the aperture separation is small enough to fit the entire sensor within a gimbal mount. For a sensor with a baseline longer than 0.5 m, the orientation of the baseline may have to be fixed. In such cases, individual scanning mirrors can be used to steer the sensor line of sight and a pathlength adjuster can be utilized to equalize the optical paths from the two apertures to the phase reference point in the scene. The effective baseline of the sensor becomes $S\cos(\theta)$ and the measurement spatial frequencies in the x and z directions are equal to $S\cos^2(\theta)/R_0\lambda$ and $S\cos^2(\theta)\sin(\theta)/R_0\lambda$, respectively.
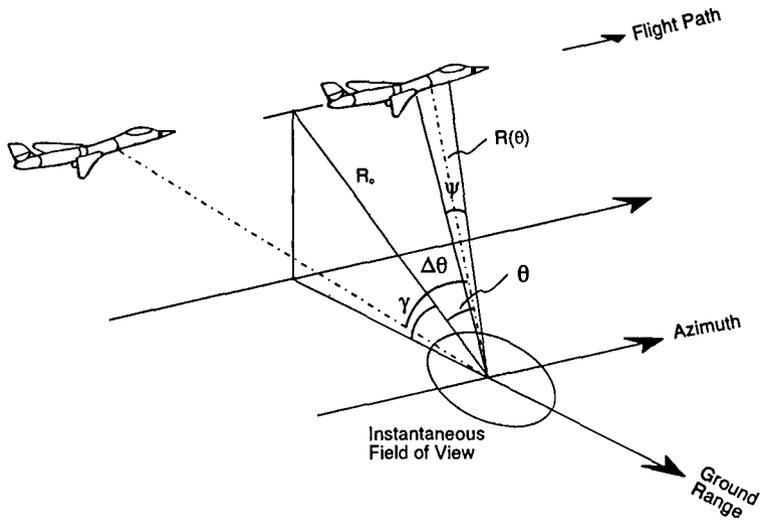
Fig. 1.35   Data gathering geometry for side-looking imaging.
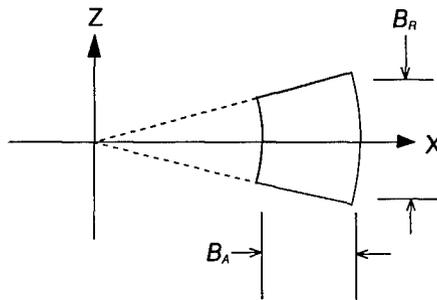
Fig. 1.36   Aperture synthesized by combining angle and wavelength diversities.

For simplicity, let us consider the case where the sensor baseline is rotated with the viewing angle, such that it remains perpendicular to the line of sight. The spatial frequency of the cross-correlation measurement in the x direction is given by $S \cos(\theta)/R_0\lambda$. For small θ, the change in viewing angle has little effect on the measurement spatial frequency in the x direction. Measurements at different spatial frequencies in the x direction can be accomplished by a diversity of wavelengths. An imaging aperture is synthesized by performing simultaneous cross-correlation measurements at a set of wavelengths that spans a spectral bandwidth of $\Delta\lambda$. The rotation of the viewing angle results in an aperture fill along an arc while the change in measurement wavelength produces a radial aperture fill.

For small θ, the spatial frequency bandwidths of the synthetic aperture generated by combining viewing angle and wavelength diversity are illustrated in Fig. 1.36. The spatial frequency bandwidths of the synthesized ap-

erture are approximately equal to $B_R = S \sin(\Delta\theta)/R_0\lambda$ in the z or range direction and $B_A = S(\Delta\lambda/\lambda)/\lambda R_0$ in the x or azimuth direction. Projecting the two-dimensional range-angle image to the ground plane, the ground range image resolution of the passive synthetic aperture sensor is given by

$$\rho_R = \frac{1}{B_R} \approx \frac{\lambda R_0}{S \sin(\Delta\theta) \cos(\gamma)} \, , \tag{1.134}$$

where γ is the grazing angle with values typically between 20 and 30 deg. The azimuth or cross-range resolution is a function of the percentage spectral bandwidth according to

$$\rho_A = \frac{1}{B_A} \approx \frac{\lambda R_0}{S(\Delta\lambda/\lambda)} \, . \tag{1.135}$$

We note that the range and cross-range resolutions are not dependent on the real collection aperture of the sensor. The image resolution in both dimensions is determined instead by the aperture separation S. The size of the synthetic aperture in the range direction is approximately equal to $S\Delta\theta$ and in the azimuth direction, $S\Delta\lambda/\lambda$. As an example, if $\Delta\theta = 14$ deg and $\Delta\lambda/\lambda = 1/4$, the size of the synthetic imaging aperture is $(S/4) \times (S/4)$. The spatial resolution of the interferometric is limited only by the largest aperture separation that can be implemented on the sensor platform.

We have shown that a synthetic imaging aperture in the z direction is synthesized by changing the viewing angle θ and in the x direction by making correlation measurements at different wavelengths. By adding another degree of freedom in the imaging geometry, a full three-dimensional synthetic imaging aperture can be synthesized. From Eq. (1.132), the spatial frequency vector in the y direction is a function of $\sin\phi$. A synthetic aperture in the y or elevation direction can also be generated by rotating the sensor baseline about the line of sight. Therefore, by rotating the sensor baseline rapidly about the optical axis over an angular range of $\Delta\phi$ as the sensor flies past the object scene, a full three-dimensional synthetic aperture can be generated. The spatial resolution in the y direction provided by the synthesized aperture is given by

$$\rho_E = \frac{1}{B_E} \approx \frac{\lambda R_0}{S \sin(\Delta\phi)} \, . \tag{1.136}$$

In all three dimensions, the spatial resolutions that can be achieved are determined by the separation of the apertures, S, but are not dependent on the size of the individual apertures.

1.4.3   Implementation with a Grating Interferometer

The cross-correlation of the fields at the two input apertures can be accomplished by measuring the complex fields at the individual apertures via heterodyne detection and then correlating the detector outputs electronically. The viability of such an approach, however, is hindered by the limitations of available hardware. The instantaneous spectral bandwidth of the heterodyne system

is restricted by the bandwidth of the detector, which, at this time, is limited to tens of gigahertz (with a 30-GHz bandwidth, the equivalent spectral bandwidth is only 10 nm at 10 μm). The narrowness of the instantaneous spectral bandwidth severely limits the amount of signal energy that can be detected. Moreover, to generate the needed laser lines to implement the aperture fill in the azimuth direction by wavelength diversity requires a very stable laser that is tunable over a broad range of wavelengths.

The cross-correlation can also be performed by interfering the fields optically with an interferometer.[388–390] The spectral bandwidth of the measurement is no longer restricted by the detector temporal bandwidth, but it must still be restricted unless the sensor field of view is very small. The optical path difference from the two apertures to a source at the edge of the field of view must be less than the coherence length provided by the measurement spectral bandwidth. For astronomical imaging, this restriction is not very severe because the object extent is typically very small. For the imaging of ground targets, however, the required sensor field of view is significantly larger and the spectral bandwidth of the correlation measurement must be made proportionally narrower. Since natural emission is spectrally wide band, the small instantaneous spectral bandwidth makes conventional interferometric imaging sensors very inefficient. In this section, the operation of a grating interferometer-based sensor[392] is described. The achromatic grating interferometer offers a means to increase substantially the spectral bandwidth of the correlation measurements. In addition, it allows multiple spectral channels to be implemented efficiently for aperture synthesis using wavelength diversity.

The spatial frequency of the correlation measurement in the direction of the baseline is given by $f = S/R\lambda$, where S is the lateral separation of the fields being correlated, R is the distance between the object and detection planes, and λ is the wavelength used in the measurement. To increase the spectral bandwidth, the grating interferometer varies the effective shear of the field being correlated such that $f = S(\lambda)/R\lambda \simeq$ constant. The spatial frequency of the correlation measurement becomes independent of wavelength. A much wider instantaneous spectral bandwidth can be used without a loss in the strength of the correlation signal. This is achieved with the grating interferometer by using the dispersive effect of the gratings.

The basic design of the grating interferometer is illustrated in Fig. 1.37. The incident radiation is diffracted by the two input gratings G_1 and G_2 and combined at G_3. After diffraction by G_3, which has the same spatial frequency as that of G_1 and G_2, the combined beams propagate coincidentally, and interfere constructively or destructively depending on the phase difference between the combined beams. For simplicity, a two-dimensional model will be used to describe the grating interferometer. Let us consider an ideal situation where the object is so far away that the field originating from any single point in the object field can be considered to be a plane wave when it impinges on the grating interferometer, and the spatial frequency of the plane wave is determined by the distance of the point source from the optical axis.

For a single point source of intensity $I_0(\alpha)$ located at a distance α from the optical axis, the field due to the point source just prior to being diffracted by the third grating G_3 is equal to
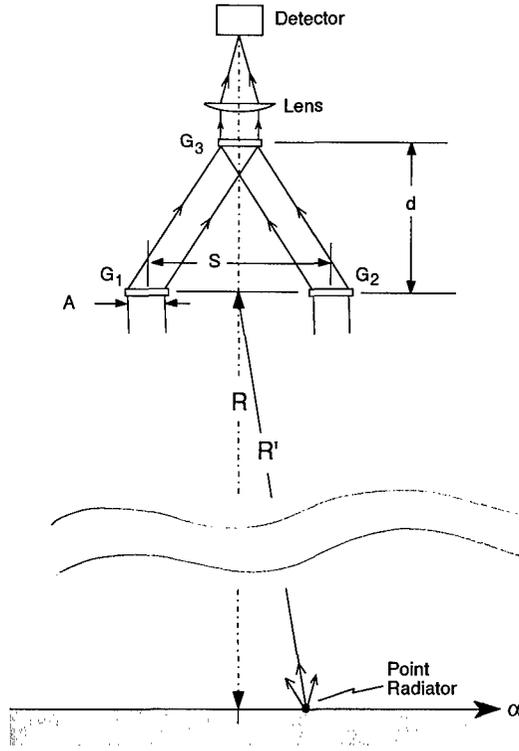
Fig. 1.37 Passive synthetic aperture imaging with a grating interferometer.

$$U_{G-}(\alpha) = [I_0(\alpha)]^{1/2}\Bigg(\exp\{-i2\pi(p - F)x - ikR(1 - \lambda^2 p^2)^{1/2}$$

$$- ikd[1 - \lambda^2(p - F)^2]^{1/2}\}$$

$$+ \exp\{-i2\pi(p + F)x - ikR(1 - \lambda^2 p^2)^{1/2}$$

$$- ikd[1 - \lambda^2(p + F)^2]^{1/2}\}\Bigg) , \tag{1.137}$$

where F is the spatial frequency of the diffraction gratings, $p = \sin\{\tan^{-1}[(r/R)]\}/\lambda$ is the spatial frequency of the coherence measurement, d is the separation between gratings G_3 and $G_1 - G_2$, and $k = 2\pi/\lambda$. The corresponding intensity distribution can be written as

$$I_{G-}(\alpha) = CI_0(\alpha)\Bigg[1 + \cos\Bigg(4\pi Fx + \frac{2\pi d}{\lambda}\{[1 - \lambda^2(p + F)^2]^{1/2}$$

$$- [1 - \lambda^2(p - F)^2]^{1/2}\}\Bigg)\Bigg] , \tag{1.138}$$

where C is an appropriate constant. Using a grating G_3 with spatial frequency F to recombine the beams diffracted by G_1 and G_2, the interfering fields propagate coincidentally and the intensity distribution right after being diffracted by G_3 becomes

$$I_{G+}(\alpha) = CI_0(\alpha)\left[1 + \cos\left(\frac{2\pi d}{\lambda}\{[1 - \lambda^2(p + F)^2]^{1/2}\right.\right.$$

$$\left.\left. - [1 - \lambda^2(p - F)^2]^{1/2}\}\right)\right] \ . \tag{1.139}$$

If we expand the square-root term into a binomial series and keep only the first-order term, we obtain

$$I_{G+}(\alpha) = CI_0(\alpha)\left[1 + \cos 2\pi\left(\frac{2Fd\alpha}{R}\right)\right] \ . \tag{1.140}$$

We see that the output intensity, to the first order, is not a function of wavelength, and the spatial frequency is a linear function of the position of the point target.

The object scene within the FOV of the interferometer can be considered as a superposition of incoherent point sources. The output intensity due to the entire scene can then be expressed as

$$I(\omega) \cong C\int_\Sigma I_0(\alpha)\,d\alpha + C\int_\Sigma I_0(\alpha)\cos(\omega\alpha)\,d\alpha \ , \tag{1.141}$$

where $\omega = 4\pi Fd/R$ and Σ denotes the area within the sensor FOV. Equation (1.141) gives a cosine transform relationship between the interferometer output and the scene intensity together with a bias term.

The quadrature component of the correlation output can be obtained by translating G_3 laterally by $\frac{1}{8}$ of a grating groove spacing which introduces $\pm\,\pi/4$ phase shifts to the interfering beams. The output of the interferometer becomes

$$I(\omega) \cong C\int_\Sigma I_0(\alpha)\,d\alpha + C\int_\Sigma I_0(\alpha)\sin(\omega\alpha)\,d\alpha \ . \tag{1.142}$$

Alternatively, complex data can be acquired by translating grating G_3 laterally at a uniform speed in a direction normal to the optical axis. The translation of the grating adds a temporal carrier and the output becomes

$$I(t) = \text{bias} + |G(\omega)|\cos[2\pi FVt + \theta(\omega)] \ , \tag{1.143}$$

where F is the spatial frequency of the grating, V is the translation speed,

$$\text{bias} = C\int_\Sigma I_0(\alpha)\,d\alpha$$

and

$$G(\omega) = |G(\omega)| \exp[i\theta(\omega)] = \int_{\Sigma} I_0(\alpha) \exp(-i\omega\alpha) \, d\alpha \ .$$

The Fourier coefficient $G(\omega)$ can be obtained by Fourier transforming the detector temporal output $I(t)$ and measuring the magnitude and phase at frequency FV.

Besides providing a means to measure the magnitude and phase of the object spectrum, the addition of a temporal carrier also improves the SNR of the measurement. Up to 1 kHz, detector noise decreases linearly with 1/frequency. By choosing a carrier frequency ≥ 1 kHz, the sensor SNR is optimized. Otherwise, a chopper has to be inserted in front of the detector to modulate the optical signal. The use of a chopper is inefficient because 50% of the input photons are thrown away. In addition, grating translation produces a temporal modulation only on the signal term. The bias term, which is typically much larger than the signal, remains a constant, and it can be removed by ac coupling the detector output. The result is lower quantization noise and the elimination of any degradation due to dc sidelobes.

As explained in the previous section, aperture fill in the range direction is achieved with viewing angle diversity and the aperture fill in the azimuth direction is accomplished through the use of wavelength diversity. With the grating interferometer, the change in measurement spatial frequency in the azimuth direction is obtained by varying the separation d between $G_1 - G_2$ and G_3.

For given values of F, d, and R, the grating interferometer measures the source spectrum at a spatial frequency of $\omega = 4\pi Fd/R$. The measurement spatial frequency is therefore proportional to the longitudinal grating separation d. Changing the grating separation d effectively alters the measurement wavelength as illustrated in Fig. 1.38. The grating G_3 at a given longitudinal
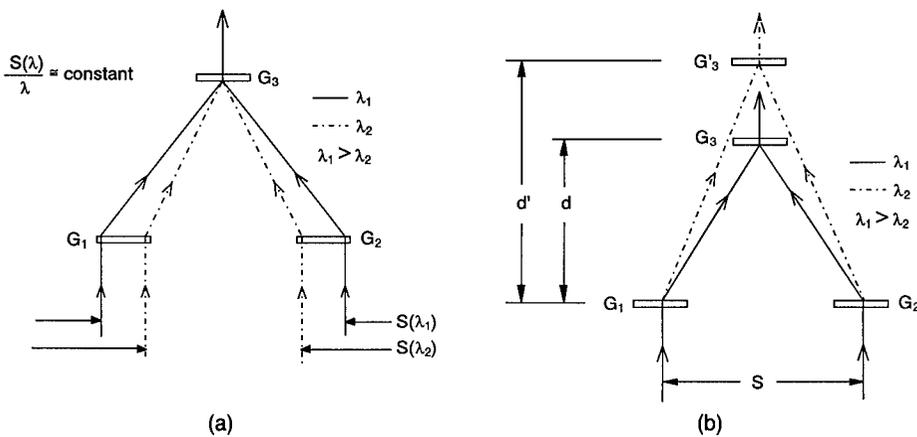
Fig. 1.38 (a) Achromaticity of grating interferometer and (b) selection of measurement spatial frequency by varying the longitudinal separation of the gratings.
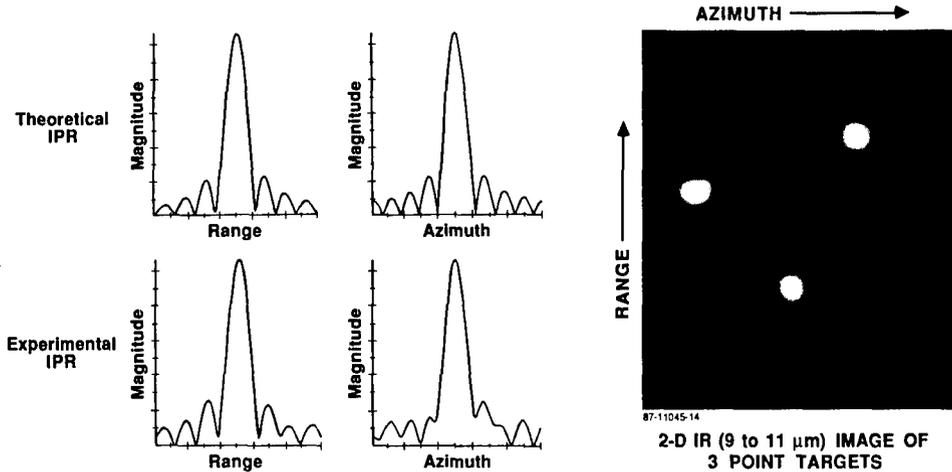
Fig. 1.39 Synthetic aperture image of three-point radiators obtained with a grating-based passive interferometric sensor.

position intercepts and combines a band of wavelengths diffracted by G_1 and G_2. Due to the wavelength dependency of diffraction, the shear varies with wavelength, as illustrated in Fig. 1.38(a) such that $S(\lambda)/\lambda$ = constant to the first order. At a different position along the optical axis, grating G_3 intercepts a different band of wavelength as shown in Fig. 1.38(b). Changing grating separation d is therefore equivalent to changing the measurement wavelength, and within the wavelength bands, the radiations interfere achromatically. Measurement at multiple spatial frequencies can be made simultaneously by utilizing a parallel set of output gratings (G_3) at different longitudinal positions, or sequentially by moving grating G_3 along the optical axis.

A range-angle image obtained with the imaging approach described in this section is presented in Fig. 1.39. The grating-based interferometric sensor operated in the 9- to 11-μm region. The ground-based imaging experiment was performed with a fixed sensor and the target was placed on a rotary platform. Such an imaging geometry is similar to the one depicted in Fig. 1.35 except that the target range R is constant instead of proportional to $1/\cos(\theta)$. The simple target was composed of three small thermal "point" sources arranged in the form of a triangle and the grazing angle was approximately 15 deg. The image in Fig. 1.39 is shown in ground range-azimuth coordinates and thus appears as if the sensor were looking straight down at the ground target. The image presentation is therefore much more akin to those from a synthetic aperture radar than a forward-looking infrared (FLIR) sensor.

1.4.4 Applications and Implementation Issues

In the preceding discussion, we have concentrated on imaging ground targets from an airborne platform. However, the change in look angle can be created by either sensor or target motion. The same imaging geometry also applies when using a ground-based sensor to image airborne targets or even low earth orbit space objects. If the object also rotates (as some satellites do) as it travels

across the sky, a three-dimensional fine-resolution image can be obtained. The sensor concept can also be implemented, at least theoretically, in space on two satellites connected by an optical link that performs both signal transfer for cross-correlation measurements and optical ranging to determine accurately and continuously the aperture separation.

In Eq. (1.130), which describes the Fourier transform relationship between the cross-correlation of the sheared field and the source intensity distribution, there is a phase term outside the Fourier integral described by $\exp[i2\pi(r_1 - r_2)/\lambda]$. To properly form an image, this phase term must be precisely known. Any uncertainty in the knowledge of $(r_1 - r_2)$, the optical path difference from a fixed phase reference point in the object scene to the two receiving apertures, will produce phase error in the synthesized aperture. The uncertainty in the optical path difference $(r_1 - r_2)$ can be caused by several factors, including (1) atmospheric turbulence, (2) uncompensated random aircraft motion, (3) sensor component motion due to, for example, vibration or thermal effects, and (4) uncertainties in platform position, heading, and velocity. Factors 3 and 4 affect the passive interferometric and active coherent sensors equally, but factors 1 and 2 pose less of a problem for the passive interferometric sensor. With an active coherent laser radar, the phase term is given by $\exp[i4\pi\hat{r}\cdot\mathbf{r}_s/\lambda]$, which is determined by the double pass optical path from the sensor to the object scene. For a down-looking airborne sensor, for example, the optical path transverses the atmosphere twice. Any fluctuation in the optical path due to atmospheric turbulence will produce significant phase errors. With a passive interferometric sensor, on the other hand, the optical paths from any point in the scene to the two receiving apertures are nearly identical near the ground where atmospheric turbulence is the strongest. Atmospheric turbulence therefore has a relatively small effect on the path difference and, in most cases, results only in a small phase error in the correlation measurements.

Similarly, the passive interferometric sensor is much less sensitive to aircraft motion than an active coherent sensor. With an active sensor, the position of the airborne sensor in the $\hat{r}$ direction must be known to within a fraction of a wavelength, which is extremely difficult to accomplish. With a passive interferometric sensor, it is only necessary to know the path difference to a fraction of a wavelength. The passive sensor is therefore more sensitive to the sensor attitude than its absolute position. While it is very difficult to measure the position of a sensor with micrometer accuracy, the attitude of a sensor can be measured directly with very high precision.

1.4.5 Summary

For some very long range imaging applications, the aperture required to achieve the needed resolution may be so large that it is impractical to implement the sensor as a conventional real-aperture focal plane imager. For such applications, a synthetic aperture imaging sensor can provide the desired image resolution with much smaller physical sensor apertures. In this section, we have combined two well-known but fundamentally different imaging concepts to produce a new class of passive sensors for high-resolution imaging. As with a long baseline interferometric sensor, the new sensor synthesizes a large imaging aperture by passively measuring the mutual intensity of the received

light field at different spatial frequencies. But unlike interferometric sensors that are designed for astronomical imaging, the new sensor can be used to image ground targets from an airborne platform to produce high-resolution imagery in range-angle coordinates, using an imaging geometry similar to that of an active synthetic aperture imaging radar or ladar. Under some limited conditions where enough degrees of freedom are available in altering the sensor/target parameters and geometry, extension to three-dimensional synthetic aperture imaging is also possible. The implementation of a passive synthetic aperture sensor on an airborne platform is technically challenging. Nevertheless, its implementation is feasible with current state-of-the-art hardware technology, at least for some specific imaging applications.

1.5   IMAGE FORMATION FROM NONIMAGED LASER SPECKLE PATTERNS
Written by James R. Fienup

In most conventional imaging systems, an image is formed by an optical system and the electromagnetic radiation (light, when at optical wavelengths) is detected in the image plane directly. However, if the scene or object being imaged is illuminated by coherent radiation, as from a laser, then several alternative approaches can be used to form an image without the use of image-forming optics. Images can be formed from data collected in a plane where the aperture of an imaging system would ordinarily be located. This is possible because the electromagnetic field in the plane of the receiver is related to the field reflected by the scene by an invertible transform. The transform is the Fresnel transform if the scene is in the near field of the receiver, and the Fourier transform if in the far field.[393] If field measurements are made at an array of points in an aperture plane (i.e., without a focusing system) by heterodyne or holographic methods, then the image formation can be accomplished by an inverse Fourier (or Fresnel) transform in a digital computer. This is commonly done for microwave synthetic aperture radar (SAR) systems.

The symbols used in this section are defined in Table 1.2.

The resolution limit imposed by diffraction is

$$\rho = \lambda R/D_a \; , \tag{1.144}$$

where ρ = resolution at the object, λ = wavelength, R = range to target, and D_a = diameter of the aperture or collecting array.

The familiar $1.22\lambda R/D_a$ is replaced by $\lambda R/D_a$ because we are considering a square aperture of length D_a. To image distant objects with very fine resolution, this limit on resolution forces the use of short wavelengths in order to allow for apertures of practical size. However, for shorter wavelengths, particularly in the ultraviolet through the midinfrared, heterodyne measurements of the field are very difficult. Furthermore, at these wavelengths the turbulence of

Table 1.2   Symbols, Nomenclature, and Units

Symbol	Nomenclature	Units
c	Speed of light	m/s
C	Autocovariance of aperture-plane intensity	W^2/m^2
d_a	Detector spacing	m
$d_a^2 \eta_d$	Area of one detector	m^2
D_a	Diameter of receiver aperture	m
D_o	Object diameter	m
f	Optical field in object plane	$\sqrt{W/m^2}$
f_I	Incoherent object brightness	W/m^2
F	Optical field in aperture plane	$\sqrt{W/m^2}$
F_I	Fourier transform of incoherent object	W/m^2
g	Optical field in image plane	$\sqrt{W/m^2}$
h	Planck's constant	J s
i	$\sqrt{-1}$	
L_c	Coherence length of laser	m
L_p	Length of laser pulse	m
M	Number of detectors per aperture-plane speckle	
n_{bp}	Number of resolved bright image points at a given resolution	
n_d	Average number of photons per detector	
n_s	Number of speckles in the aperture	
N	Number of independent realizations of the speckle pattern	
N_m	Maximum number of independent realizations available	
p	Coherent impulse response	
P	Pupil function	
r	Autocorrelation of coherent object	W/m
r_I	Autocorrelation of incoherent object	W^2/m^2
r_o	Object average reflectivity	
R	Range to object	m
s	Point spread function, $|p|^2$	
S	(Normalized) optical transfer function	
S_o	(Unnormalized) optical transfer function	
SBP_{max}	Maximum 1-D space-bandwidth product	
SNR	Signal-to-noise ratio	
(u,v)	Spatial-frequency coordinates	m
u_r	$(u^2 + v^2)^{1/2}$	m
v_a	Speckle velocity in aperture plane	m/s
v_o	Object velocity	m/s
(x,y)	Image-plane coordinates	m
(x_a,y_a)	Aperture-plane coordinates	m
(x_o,y_o)	Object-plane coordinates	m
Greek:		
Δt_1	Time for pattern to move across one-quarter of the speckle width	s

Table 1.2 *(continued)*

Symbol	Nomenclature	Units
Δt_a	Time for pattern to move across aperture	s
Δt_m	Time before object rotates too much	s
Δz	Object depth along line of sight	m
η_d	Detector duty cycle	
η_q	Quantum efficiency of the detector	
λ	Wavelength	m
μ	Complex coherence factor	
ρ_a	Speckle width at aperture	m
ρ_o	Potential resolution at object	m
τ_{atm}	One-way atmospheric transmittance	
τ_{opt}	Transmittance of receiver optics (if any)	
τ_{pol}	Fraction of light reflected in desired polarity	
ϕ	Estimate of aperture-plane phase	rad
ψ	Aperture-plane phase	rad
ω	Object rotation rate (axis perpendicular to line of sight)	rad/s
Operators:		
$\mathscr{F}$	Fourier transformation	
$*$	Convolution	
$\otimes$	Autocorrelation	

the atmosphere causes phase errors in the field propagated through it, which can severely limit the resolution of a heterodyne system. If it is possible to reconstruct an image from the intensity of the field in the aperture plane, then these limits to resolution caused by aberrations can be overcome. Figure 1.40 shows examples of configurations that can be used to gather the aperture-plane intensity. These laser intensity patterns are speckled if the scene is rough in comparison to the wavelength of illumination. In this section, some methods for forming images from nonimaged, aperture-plane speckle patterns are discussed. First some terms are defined and a few methods are briefly described. Then one particular method, imaging laser correlography, is described in more detail, to serve as an example of a nonconventional imaging method using aperture-plane speckled intensity patterns.

These techniques generally require that the scene be of limited extent—that it be a bright object on a dark background. They have been developed primarily for the applications of fine-resolution imaging of space objects by ground-based systems and of discriminating decoys from reentry vehicles for the Space Defense Initiative (SDI) midcourse phase. Another application could be the imaging of a ground scene from an overhead platform if only a well-defined portion of the ground were illuminated by a coherent source. Relying on powerful sources of short-wavelength, coherent radiation, these techniques have been in development only since the 1960s. The first concerted effort[394-396] to develop these ideas began at the Woods Hole Summer Study in 1966.

Laser   Detector array   (b)   (c)

(a)   Alternative detector array geometries
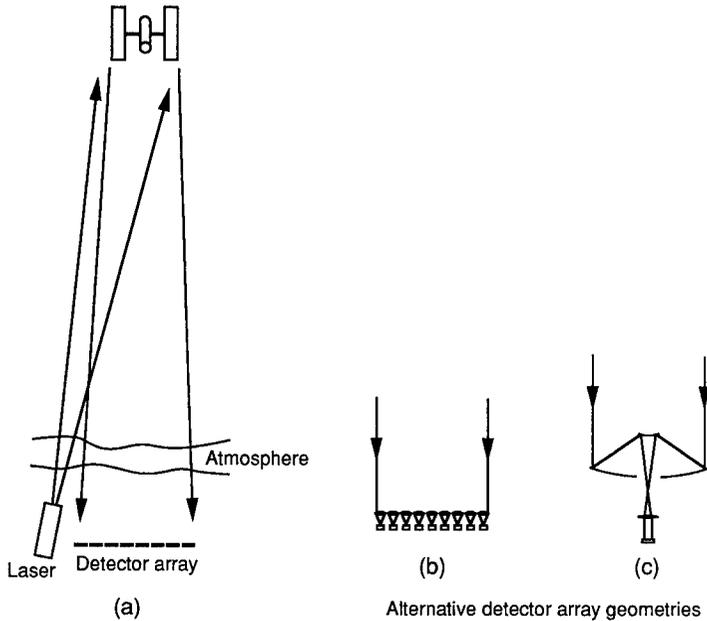
Fig. 1.40 Imaging geometries: (a) The intensity of coherent radiation scattered from an object is measured by an array of detectors without intervening optics. (b) Alternatively, a collection lens can be used to gather the light onto each of the individual detectors, which can have smaller areas. (c) Another alternative is to reimage the pupil plane of a telescope onto a detector array of small dimensions, such as a CCD array.

1.5.1   Basic Imaging Equations

Because of the mixture of coherent and incoherent imaging concepts inherent in imaging correlography, it is necessary to define here the basic imaging equations that will be needed later in this section.

The object is assumed to be illuminated by radiation that is both spatially and temporally coherent over the volume of the object. Here we refer to the illuminated scene as the object. Let the complex field reflected by the object be $f'(x_o, y_o)$ in a plane perpendicular to the line of sight (LOS). Then the field in the aperture plane at the detector array, also for simplicity assumed to be perpendicular to the LOS, is given by the Fresnel transform[393]:

$$
\begin{aligned}
F'(x_a, y_a) = {} & \frac{1}{i\lambda R} \exp\left(\frac{i2\pi}{\lambda R}\right) \exp\left[\frac{i\pi}{\lambda R}(x_a^2 + y_a^2)\right] \\
& \times \iint f'(x_o, y_o) \exp\left[\frac{i\pi}{\lambda R}(x_o^2 + y_o^2)\right] \\
& \times \exp\left[\frac{-i2\pi}{\lambda R}(x_a x_o + y_a y_o)\right] dx_o \, dy_o \ .
\end{aligned}
\tag{1.145}
$$

In this section, all integrals are over $(-\infty, \infty)$. For the case of a diffusely scattering object, the intensity of this aperture-plane optical field is a speckle pattern. We are assuming that in front of the detector array is a polarizer that selects a single polarization, which is necessary for the methods described here to work. Hence, throughout this section we are considering radiation of a single polarization, enabling us to employ scalar diffraction theory.

If we absorb the constant phase term $-2\pi/\lambda R$, the quadratic phase term $(-\pi/\lambda R)(x_a^2 + y_a^2)$, and the factor i into the phase of $F'(x_a, y_a)$ to form $F(x_a, y_a)$, and absorb the quadratic phase term $(\pi/\lambda R)(x_o^2 + y_o^2)$ into the phase of $f'(x_o, y_o)$ to form $f(x_o, y_o)$, then we have the simpler Fourier transform relationship

$$F(x_a, y_a) = |F(x_a, y_a)| \exp[i\psi(x_a, y_a)] = \mathscr{F}[f(x_o, y_o)]$$

$$= \frac{1}{\lambda R} \int\int f(x_o, y_o) \exp\left[\frac{-i2\pi}{\lambda R}(x_a x_o + y_a y_o) \right] dx_o \, dy_o \qquad (1.146)$$

between the field $f(x_o, y_o)$ at the object plane and the field $F(x_a, y_a)$ at the receiver plane. This Fourier relationship is valid for the near-field conditions[393] normally allowed only for the Fresnel transform. If the complex-valued aperture-plane field, $F(x_a, y_a)$, could be measured without any phase errors, then an inverse Fourier transform performed in a computer would yield a diffraction-limited complex-valued, coherent image. This image would be given by the convolution (with appropriate scaling of the coordinates to account for the image magnification)

$$g(x, y) = f * p(x, y) = \frac{1}{\lambda R} \int\int f(x_o, y_o) p(x - x_o, y - y_o) \, dx_o \, dy_o \ , \qquad (1.147)$$

where $*$ denotes convolution, $p(x, y) = \mathscr{F}^{-1}[P(x_a, y_a)]$ is the system (coherent) impulse response, and $P(x_a, y_a)$ is the pupil function (unity over the extent of the aperture and zero outside). The corresponding point spread function for incoherent imaging through the same aperture would be

$$s(x, y) = |p(x, y)|^2 = \mathscr{F}^{-1}[S_o(u, v)] \ , \qquad (1.148)$$

where $S_o(u, v)$ is the unnormalized optical transfer function, given by the autocorrelation of the pupil function:

$$S_o(u, v) = P \otimes P(u, v) = \frac{1}{\lambda R} \int\int P(x_a, y_a) P^*(x_a - u, y_a - v) \, dx_a \, dy_a \ , \qquad (1.149)$$

and P^* is the complex conjugate of P, and (u, v) is a spatial-frequency coordinate. The (normalized) optical transfer function is given by

$$S(u, v) = \frac{S_o(u, v)}{|S_o(0, 0)|} \ . \qquad (1.150)$$

Like any coherent image, $g(x,y)$ would be speckled, assuming that the surface of the object is rough compared to the wavelength of the radiation.[397] As the object translates or rotates, a new part of the back-scattered field falls on the receiver; effectively, the object takes on a different realization of the phase in the object plane. When we wish to distinguish these different fields coming from the object, we denote the n'th realization of the object field as $f_n(x_o,y_o)$ and the corresponding field at the aperture plane as $F_n(x_a,y_a)$. The image of each $f_n(x_o,y_o)$ would have a different realization of the speckle pattern across it. If the field in the aperture plane translates by the width of the aperture between the time the two images are collected, then, for the case of a diffusely scattering object, the two field distributions are independent and the complex images are uncorrelated. Conditions for achieving this are discussed in Sec. 1.5.4.1. An average over many such object realizations would give

$$f_I(x_o,y_o) \;=\; \langle |f_n(x_o,y_o)|^2 \rangle_n \;=\; \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} |f_n(x_o,y_o)|^2 \;, \tag{1.151}$$

where $\langle \cdot \rangle_n$ denotes the ensemble average over the speckle realizations or aspect angles and N is the number of independent realizations. This noncoherent average of many coherent objects yields the ordinary incoherent reflectivity (or brightness distribution) of the object, which is a real-valued, non-negative distribution. We denote the Fourier transform of the incoherent object, $f_I(x_o,y_o)$, as $F_I(u,v)$. The coherent images $p*f_n(x,y)$ are speckled versions of $s*f_I(x,y)$, the image of the incoherent object. The normalized Fourier transform,

$$\mu(u,v) \;=\; \frac{F_I(u,v)}{F_I(0,0)} \;, \tag{1.152}$$

is the complex coherence factor (or the complex visibility function),[398] and $|\mu(u,v)|^2$ is the normalized spatial power spectrum (i.e., the normalized squared modulus of the Fourier transform) of the incoherent object.

1.5.2 Imaging Methods

The quantity that we assume to be measured, and from which we wish to reconstruct an image, is the aperture-plane intensity, $|F(x_a,y_a)|^2 = |F'(x_a,y_a)|^2$. The square root of the intensity, $|F(x_a,y_a)|$, is the magnitude (commonly called the *modulus* or *amplitude*) of the field and $\psi(x_a,y_a)$ is the phase. Without the phase of $F(x_a,y_a)$, the image cannot be computed simply by inverse Fourier transformation. Since $|\mathscr{F}[f^*(-x,-y)]| = |\mathscr{F}[f(x,y)]| = |F(x_a,y_a)|$, from the aperture-plane intensity it is impossible to distinguish between the object $f(x,y)$ and its twin $f^*(-x,-y)$. This twofold ambiguity is usually acceptable since the twin looks just like the object, but rotated by 180 deg, and it usually suffices for identifying the object. In addition, $|\mathscr{F}[f(x-x_1,y-y_1)]| = |\mathscr{F}[f(x,y)]|$, and so the absolute position of the object cannot be determined from the intensity $|F(x_a,y_a)|^2$ alone.

We next briefly describe some of the methods that can be used to form images from this seemingly insufficient aperture-plane intensity data.

1.5.2.1 Holography. By inverse Fourier transforming $|F(x_a,y_a)|^2$, we arrive at the autocorrelation of the object:

$$r(x,y) = f \otimes f(x,y) = \frac{1}{\lambda R}\int\int f(x',y')f^*(x-x',y-y')\,dx\,dy$$

$$= \mathscr{F}^{-1}[|F(x_a,y_a)|^2] \ . \tag{1.153}$$

If the object consists of an extended part of the object plus a glint, separated from the extended part by at least the width of the extended part, then $|F(x_a,y_a)|^2$ is the equivalent of a hologram[399] of the extended part. When this holographic condition is met, the autocorrelation consists of four terms, one of which is an image of the extended part of the object. Since $|F(x_a,y_a)|^2$ is measured only over the aperture $P(x_a,y_a)$ of width D_a, the autocorrelation and the reconstructed image are limited in resolution (in object space) to $\rho = \lambda R/D_a$. This imaging modality was analyzed for the case of imaging space objects from the earth by Refs. 394 and 395.

1.5.2.2 Object Support from Autocorrelation Support. The diameter of the object $f(x,y)$ along any direction is just half the diameter of the autocorrelation $r(x,y)$ along the same direction. The support of the object is defined as a set of points (or an area) outside of which the object has a value of zero, consistent with the assumption that we are considering bright objects on a dark background. The support of the autocorrelation allows us to determine bounds on the support of the object. For the sake of brevity these methods are not detailed here, except to say that they involve taking the intersection of two or more translated versions of the support of the autocorrelation. The reader is referred to Refs. 401 and 402 for the details. For favorable cases, such as a highly nonconvex object or an object consisting of two or more separated parts, the upper bound on the support of the object can serve as an approximate outline of the object. This outline may be sufficient information about the object for some applications. For less favorable cases, such as for a convex object, the upper bound only gives the general size of the support of the object. In all cases, the upper bound on the support can be used as a support constraint for phase retrieval, as is described in the next section.

1.5.2.3 Phase Retrieval. An approach to image formation from coherent aperture-plane intensity is *phase retrieval*. The approach is to use a computer algorithm to determine a phase $\phi(x_a,y_a)$, which, when combined with $|F(x_a,y_a)|$ from the measured intensity and Fourier transformed, yields an image

$$\hat{g}(x,y) = \mathscr{F}^{-1}\{|F(x_a,y_a)|\exp[i\phi(x_a,y_a)]\} \tag{1.154}$$

that is consistent with the support constraint (i.e., has a value close to zero in the region outside the upper bound on the support of the object). Dozens of phase retrieval algorithms have been proposed—see, for example, Ref. 403. Phase retrieval algorithms have been particularly successful for real-valued,

non-negative images such as those one would have for incoherent illumination. However, from $|F(x_a,y_a)|$ the quantity being reconstructed is the field, which is complex valued. Then the only constraint available is the support constraint. In this case it is possible to reconstruct an image, particularly for objects consisting of two or more separated parts; but it is difficult and is not yet reliable for general objects.[404,405]

1.5.2.4 Additional Measurements.

If additional measurements are made, then the problem of reconstructing a complex-valued image becomes much easier. One example for a space-based system is to have a diffraction-limited aperture of moderate size embedded within a larger array of detectors. By use of a beamsplitter, the intensity passing through the moderate-sized aperture is measured simultaneously in two different planes: in the focal plane, which contains a diffraction-limited but moderate-resolution image, and in the plane of the aperture.[406] Then the reconstruction of the phase over the diffraction-limited aperture can be accomplished by an iterative algorithm that is robust.[407,408] This phase can then be used to help to reconstruct the phase over the larger array, where only intensity measurements were made, in a boot-strapping version of a phase retrieval algorithm.[406]

A second approach using additional measurements is to combine aperture-plane intensity methods with wavefront sensor data. The imaging wavefront (Hartmann-Idell) sensor[409] combines wavefront slope data from a Hartmann array with the aperture-plane intensity data to reconstruct the optical field, from which an image is reconstructed by Fourier transformation. By temporally averaging the wavefront slopes over many speckle realizations during a time interval less than the atmospheric correlation time, it is also possible to estimate the aberrations due to the turbulent atmosphere, which could then be subtracted to form a diffraction-limited image.

Other schemes using additional measurements are also possible.

1.5.3   Imaging Correlography

The diffraction-limited autocorrelation, $p*r_n(x,y)$, which can be computed from the aperture-plane intensity $P(x_a,y_a)|F_n(x_a,y_a)|^2$, is a complex-valued, speckled autocorrelation. As the object translates or rotates, the aperture-plane intensity pattern moves. After sufficient motion at the receiver a new realization of the speckled intensity pattern appears, which yields a new realization of the speckled autocorrelation. The diffraction-limited autocorrelation of the incoherent image, $|p|^2*r_I(x,y)$, can be determined by averaging the squared magnitudes of the speckled autocorrelations over many realizations of the speckle pattern[410–412]:

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N}|p*r_n(x,y)|^2 = b_1|p(x,y)|^2 + b_2|p|^2*r_I(x,y) \ , \tag{1.155}$$

where

$$b_1 = b_2\left[\int\int |f_I(x',y')|^2 \ dx' \ dy'\right]^2 \tag{1.156}$$

is the square of the total measured intensity, b_2 is a constant, and N is the number of independent realizations. This result assumes a rough object whose coherent autocorrelation function is of the form of a delta function. The first term on the right-hand side of Eq. (1.155) is a point-spread function of known shape located at the origin (the dc term) that can be subtracted, leaving $|p|^2 * r_I(x,y)$, the diffraction-limited autocorrelation of the incoherent object. That is, the noncoherent average of the autocorrelations of the *coherent* images yields the autocorrelation of the *incoherent* image. Alternatively, processing the same intensity measurements $P(x_a,y_a)|F_n(x_a,y_a)|^2$ in a different way,[413] we compute the autocovariance of the intensity, which is given by

$$C(u,v) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \iint P(x_a + u, y_a + v) P(x_a,y_a)$$

$$\times \ [|F_n(x_a + u, y_a + v)|^2 |F_n(x_a,y_a)|^2 - \bar{I}^2] \ dx_a \ dy_a \ , \qquad (1.157)$$

where $\bar{I}$ is the average intensity over the aperture plane. Using the moment factoring theorem for circular-complex Gaussian fields (Ref. 414, p. 44), it can be shown that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} [|F_n(x_a + u, y_a + v)|^2 |F_n(x_a,y_a)|^2 - \bar{I}^2] = |F_I(u,v)|^2 \ , \qquad (1.158)$$

and so

$$C(u,v) = S(u,v)|F_I(u,v)|^2 \ , \qquad (1.159)$$

where $S(u,v)$ is the optical transfer function for the aperture. That is, the autocovariance of the intensity of the *coherent* aperture-plane field is proportional to the squared magnitude of the Fourier transform of the *incoherent* object. This is related to the incoherent autocorrelation $r_I(x,y)$ by a Fourier transform. Computation of the autocovariance of the aperture-plane intensity is completely equivalent to intensity interferometry,[415] but using realizations of reflected laser speckle patterns rather than incoherent light emitted by the object.

Each pair of detectors over the aperture contributes to a measure of $C(u,v)$ and therefore to $|F_I(u,v)|^2$. If the pair of detectors is at locations (x_{a1},y_{a1}) and (x_{a2},y_{a2}) in the aperture plane, then they contribute to a measurement of $|F_I(u,v)|^2$ at spatial frequency $(u,v) = \pm(x_{a1} - x_{a2}, y_{a1} - y_{a2})$. [The $\pm$ is due to the fact that $|F_I(-u,-v)|^2 = |F_I(u,v)|^2$, and both are measured simultaneously.] In radio interferometry, the separations (u,v) are referred to as *baseline separations*. Redundant measurements—different pairs of detectors with the same baseline separations—contribute to increasing $S(u,v)$ and increasing the SNR of the estimate of $|F_I(u,v)|^2$.

Having computed $|F_I(u,v)|$ as described, an incoherent image can be fairly reliably reconstructed using a phase retrieval algorithm, even though the reconstruction of the individual coherent images from the intensities $|F_n(x_a,y_a)|^2$ is difficult. The reconstruction is easier for incoherent images because, in addition to the support constraint, we have the additional constraint that the image is a real-valued, non-negative function.

Prior to image reconstruction, it is usually advantageous to perform a filtering or a weighting of $|F_I(u,v)|$. This is needed because the SNR for the higher spatial frequencies is much poorer than that at lower spatial frequencies. For the highest spatial frequencies, the noise usually dominates. The spatial frequencies at which SNR $\leq$ 1 should be discarded or at least greatly deemphasized. This can be accomplished, for example, by multiplying $|F_I(u,v)|$ by a filter function

$$W(u,v) = \frac{S(u,v)|\hat{F}_I(u,v)|^4}{|S(u,v)|^2|\hat{F}_I(u,v)|^4 + \Phi(u,v)} , \tag{1.160}$$

where $|\hat{F}_I(u,v)|^2$ is an estimate of $|F_I(u,v)|^2$ (which could be an average power spectrum for objects of an appropriate class) and $\Phi(u,v)$ is the variance of the noise in the estimate of the autocovariance function. This filter function is related to the Wiener-Helstrom filter.[416] Other variations on the filter function have also been used—anything that suppresses the highest spatial frequencies, where SNR $\leq$ 1, can be beneficial.

Figures 1.41 and 1.42 show an example of a computer simulation of imaging laser correlography.[412] In Fig. 1.41 we see that as more realizations of the
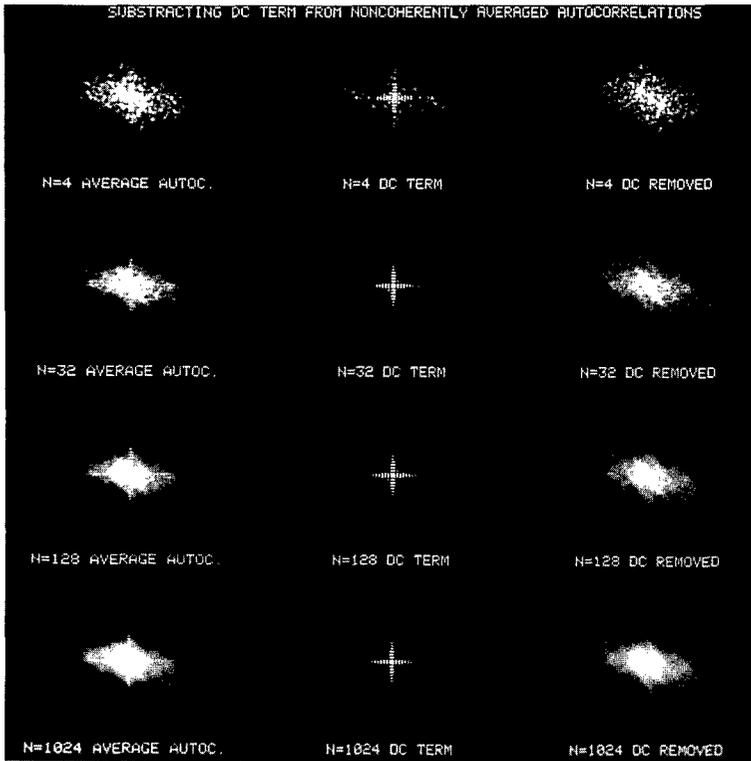
Fig. 1.41 Noncoherent averages of coherent image autocorrelations.[412] First row, left to right: noncoherent average of $N = 4$ coherent image autocorrelations, dc term, and averaged autocorrelation with dc term subtracted; second row: $N = 32$; third row: $N = 128$; and fourth row: $N = 1024$.
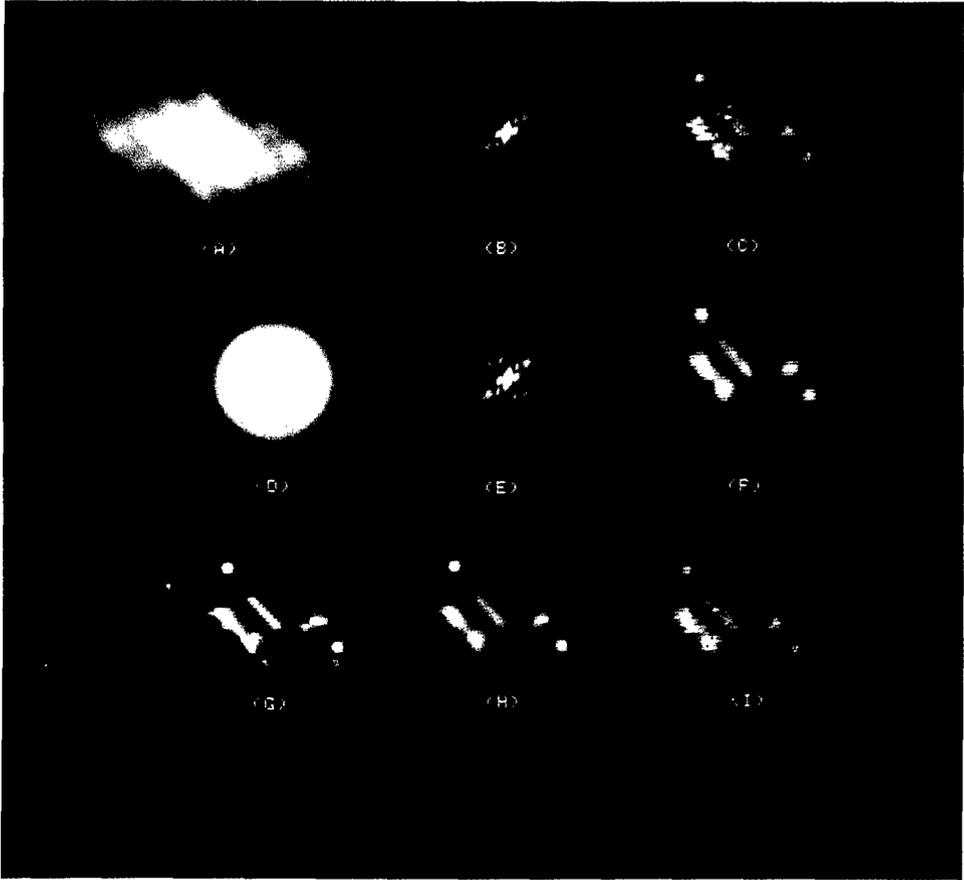
Fig. 1.42 Image recovery from simulated correlography data.[412] (A) Estimate of the autocorrelation of the incoherent image from $N = 10,000$ coherent image autocorrelations; (B) estimate of the Fourier modulus of the incoherent object; (C) image reconstructed from (B) by means of the iterative-transform phase retrieval algorithm; (D) Wiener filter; (E) filtered Fourier modulus; (F) image reconstructed from (E); (G) original incoherent object; (H) Wiener-filtered incoherent object; and (I) result of Wiener filtering of (C).

speckled autocorrelations are averaged together, it approaches the autocorrelation of the incoherent object. Figure 1.42 shows an example of image reconstruction from simulated correlography data using the iterative transform phase retrieval algorithm.[408] Filtering the data is necessary to suppress noise at the higher spatial frequencies. Figure 1.42(F), which shows the image reconstructed from the filtered data, compares favorably with Fig. 1.42(H), which shows the ideal image passed through the same filter. Figure 1.43 shows an example of imaging laser correlography in a laboratory experiment.[400,417]
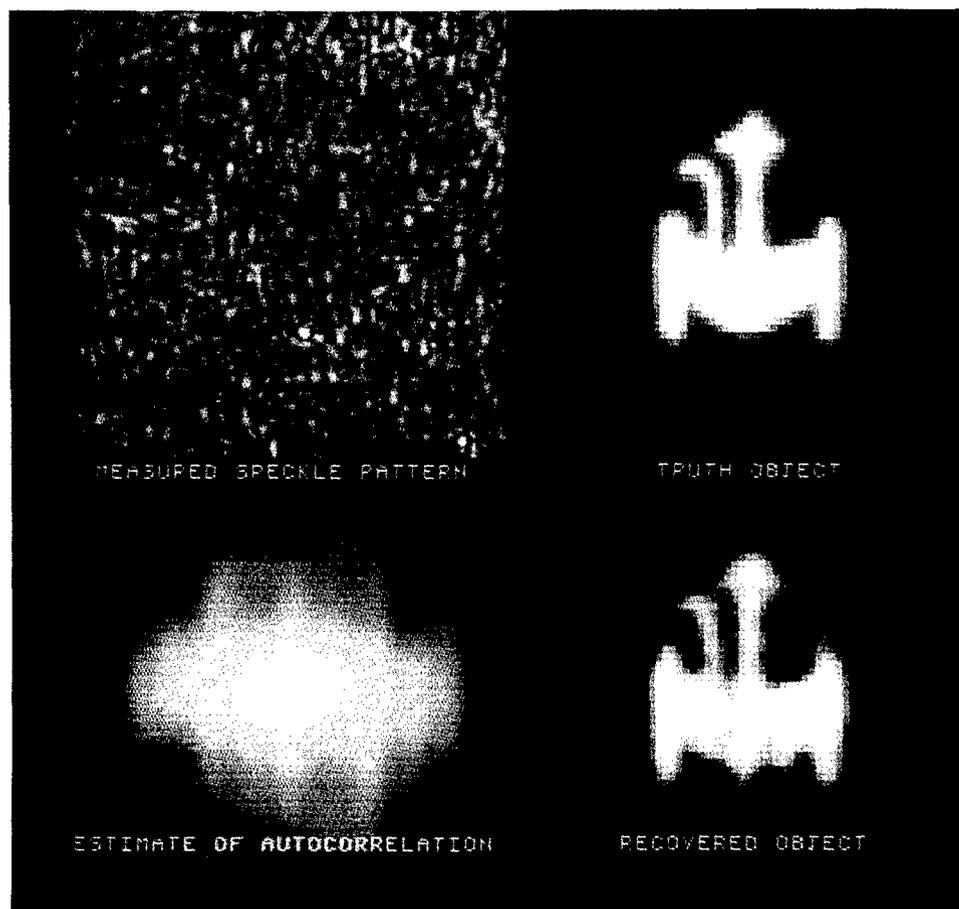
Fig. 1.43 Example of reconstruction from data gathered in the laboratory. *Top left:* one realization of the measured speckle intensity patterns; *top right:* incoherent image of the object; *bottom left:* autocorrelation estimate computed using 100 realizations of the speckle pattern; and *bottom right:* the reconstructed image. (Image courtesy of J. D. Gonglewski, P. S. Idell, and D. G. Voelz, Phillips Laboratory)

1.5.4   System Parameters

With the basic method of imaging laser correlography in hand, we now describe the requirements on the system—the transmitter (laser) and the receiver (detector array). For the sake of an example calculation, we use the parameters given in Table 1.3.

1.5.4.1   Receiver Requirements.   To obtain a resolution at the object of ρ_o ($= 0.2$ m for the example), an aperture with a diameter of at least

$$D_a = \frac{\lambda R}{\rho_o} \tag{1.161}$$

Table 1.3   Example of Parameters for an Imaging Scenario

Symbol	Example Value	Description
D_o	5 m	Object diameter
Δz	5 m	Object depth along LOS
ω	1 rad/s	Object rotation rate (axis perpendicular to LOS)
v_o	5 km/s	Object velocity perpendicular to LOS
λ	1 μm	Wavelength
R	1000 km	Range to object
D_a	5 m	Receiver aperture diameter
ρ_o	0.2 m	Resolution at object $= \lambda R/D_a$
ρ_a	0.2 m	Speckle size in aperture $= \lambda R/D_o$

(= 5 m for the example) is required. Limits on resolution that have to do with the SNR will be discussed later.

A fundamental quantity needed to explain the system requirements is the characteristic diameter ρ_a of the speckles in the detected intensity pattern, $|F_n(x_a,y_a)|^2$. It is given by

$$\rho_a = \frac{\lambda R}{D_o} \qquad (1.162)$$

(= 0.2 m for the example). The maximum possible space-bandwidth product of the image in each dimension is

$$\text{SBP}_{\max} = \frac{D_o}{\rho_o} = \frac{D_o D_a}{\lambda R} = \frac{D_a}{\rho_a} \qquad (1.163)$$

(= 25 for the example). This is the maximum number of resolution elements across the object, which is equal to the number of speckles across the aperture. The actual number of resolution elements across the object may be fewer because the resolution can be limited by noise, as is described later. In order for $|F_n(x_a,y_a)|^2$ to be measured in the aperture plane without aliasing (to avoid undersampling), the center-to-center spacing of the detectors must satisfy

$$d_a \leq \frac{\rho_a}{2} = \frac{\lambda R}{2 D_o} \qquad (1.164)$$

(= 0.1 m for the example), which gives two samples per speckle in each dimension. For a filled array, D_a/d_a (= 50 for the example) detectors are needed in each dimension. However, since the data used for reconstruction, $|F_I(u,v)|^2$, are given by the autocorrelation of the measured data, $|F_n(x_a,y_a)|^2$, sampling of the measured data on a regular filled grid is not necessary. In fact, sparse arrays of detectors may be used.[413] The effective transfer function of the system is the autocorrelation of the aperture function, which is equivalent to the

optical transfer function. Issues regarding filling of the (u,v) plane (Fourier aperture) are similar to those encountered in radio interferometry.[418]

The intensity pattern moves across the aperture plane, due to the object's rotation, at a velocity $2\omega R$ (= 2000 km/s for the example), and so it moves by a speckle diameter in time $\rho_a/(2\omega R) = \lambda/(2\omega D_o)$ (= 100 ns for the example). Therefore the integration time for a single speckle pattern (without some form of speckle tracking) should be limited to about

$$\Delta t_1 \le \frac{\lambda}{8\omega D_o} \tag{1.165}$$

(= 25 ns for the example). The intensity pattern also moves across the aperture plane at a velocity $2v_o$ (= 10 km/s for the example) due to the velocity of the object. Note that at long ranges the velocity of the aperture-plane speckles due to object rotation are much greater than that due to object velocity, except for very slow rotation rates. If the detectors have an integration time or a readout time that is long compared with Δt_1, then the detectors must be gated over an interval of time Δt_1 (or the speckles must be tracked). For a pulsed laser of pulse length L_p, a long-readout detector could only use the fraction $c\Delta t_1/L_p$ of the energy of the pulse, where c is the speed of light.

The time it takes the speckles to pass over the receiver aperture of diameter D_a is

$$\Delta t_a = \frac{D_a}{2\omega R} \tag{1.166}$$

(= 2.5 μs for the example). This is the time necessary to obtain an independent realization of the speckle pattern and defines the maximum rate at which independent realizations can be collected (400,000 independent realizations per second, for the example). The maximum integration time for a single speckle realization is less than Δt_a even with speckle tracking. Since there are $SBP_{max} = D_o/\rho_o$ (= 25 for the example) diffraction-limited resolution elements across the object, a rotation of the object by roughly $\lambda R/(D_o D_a)$ radians will cause the object brightness distribution to begin to change substantially. This would occur in time

$$\Delta t_m = \frac{\lambda R}{\omega D_o D_a} \tag{1.167}$$

(= 40 ms for the example). Unless multiple revolutions of the object are used, this is also the maximum amount of time available for gathering an image of the object for a single aspect angle. During time Δt_m, the maximum number of independent realizations available is

$$N_m = \frac{\Delta t_m}{\Delta t_a} = \frac{2\lambda R^2}{D_o D_a^2} \tag{1.168}$$

(= 16,000 realizations for the example).

A synthetic-aperture collection approach may also be taken. The approach described above is to gather N snapshots of the laser speckle pattern over a two-dimensional (possibly sparse) array of detectors. Using a synthetic-aperture approach, only a one-dimensional array of detectors is required. For simplicity, suppose that the 1-D array is along the y_a axis and that the speckles move in the x_a direction with velocity $v_a = 2\omega R$ (= 2000 km/s for the example). At any instant in time t, the array is collecting a 1-D cut, $|F_n(v_a t, y_a)|^2$, through the speckle pattern. As the speckle pattern sweeps across the 1-D array, additional slices of speckle pattern fall on the detector array. For the case of cw illumination, the detectors must read out each sample of the intensity in a time $\Delta t_1 \leq \lambda/(8\omega D_o)$ (= 25 ns for the example). Then the total data rate is $(D_a/d_a)/\Delta t_1$ (= 2×10^9 samples per second for the example). After time $\Delta t_a = D_a/v_a$ (= 2.5 μs for the example), a two-dimensional array of values of $|F_n(x_a, y_a)|^2$ has been read out, equivalent to a square area of length D_a. After time $N\Delta t_a$, a rectangular array of values of $|F_n(x_a, y_a)|^2$ has been read out of size $ND_a \times D_a$. This rectangular array can be subdivided into N square arrays of size $D_a \times D_a$, each equivalent to one of the snapshots (realizations) of the speckle pattern that would have been measured if we had a filled 2-D array. Even better than dividing the rectangle into squares would be to compute the covariance function $C(u,v)$ by summing over the products of all pairs of intensity measurements separated by the distance (u,v). This would increase the SNR somewhat for spatial frequencies having components in the u direction, since separations would be used that would otherwise have crossed the boundaries between the squares of data forming the snapshots. This idea of synthesizing a 2-D array from a 1-D array has the advantage of requiring many fewer detectors than a real 2-D array, but the detectors must be read out very fast and must be oriented appropriately with respect to the direction of the speckle velocity.

1.5.4.2 Transmitter Requirements.   The received illumination is assumed to be quasimonochromatic, i.e., the intensity pattern $|F_n(x_a, y_a)|^2$ should be constant over the wavelength band in use. The requirements on the coherence length of the laser almost always supersede this requirement, which can thereby be ignored. The coherence length required of the laser illuminator is

$$L_c \geq 2\Delta z \tag{1.169}$$

(= 10 m for the example), where Δz is the depth of the object along the LOS.

If the laser is pulsed, then the pulse repetition frequency is required to be fast enough in order to obtain a large number of speckle realizations before the object rotates too much.

The laser power requirements are driven by several factors, including the fraction of the transmitted energy that falls on a detector and the SNR of $|F_I(u,v)|^2$ for correlography.

1.5.4.3 Signal-to-Noise Ratio.   The SNR of $|F_I(u,v)|^2$ (the signal divided by the standard deviation of the noise), as limited by the number of realizations and photon noise, is given by[419]

$$\text{SNR}(u,v) = \frac{[Nn_s S(u,v)]^{1/2}|\mu|^2}{\left[3 + 14|\mu|^2 + 3|\mu|^4 + \dfrac{4(1 + 2|\mu|^2)}{Mn_d} + \dfrac{1 + |\mu|^2}{Mn_d^2}\right]^{1/2}} , \quad (1.170)$$

where

n_d = the average number of photons per detector
M = the number of detectors (pixels) per aperture-plane speckle area
Mn_d = the average number of photons per speckle
n_s = the number of speckles within the area of the aperture.

In this formula, $|\mu|^2$ is a function of (u,v). Figure 1.44 shows two families of plots of SNR(u,v) as a function of light level for two values of $|\mu|^2$ (0.25 and 0.01); and several values of Mn_d, the number of photons per speckle. The first three terms in the denominator of the expression for SNR(u,v) are due to the approximation of the ensemble average for the autocovariance by a finite sum over N realizations, and the last two terms are due to photon noise. For $|\mu|^2 \ll 1$, and for $M = 4$ detectors per speckle, these two contributions to the noise are equal when $n_d = 0.5$ photon per detector or $Mn_d = 2$ photons per speckle. Consequently, photon noise can be ignored when there are much greater than 2 photons per speckle. For the case of a photon-counting detector, any laser power that yields more than about 8 photons per speckle is wasting energy, since at that point it is the number of realizations, not the photon noise, that is limiting the SNR, as can be seen from Fig. 1.44.

For the higher spatial frequencies, for which $|\mu|^2 \ll 1$, and for moderate light levels (greater than two photons per speckle), the SNR expression reduces to

$$\text{SNR}(u,v) \approx [(N/3)n_s S(u,v)]^{1/2}|\mu(u,v)|^2 . \quad (1.171)$$

For a given object and resolution, this expression shows the trade-off that can be made between the number N of realizations and the number $n_s S(u,v)$ of redundant measurements of a given spatial frequency (u,v).
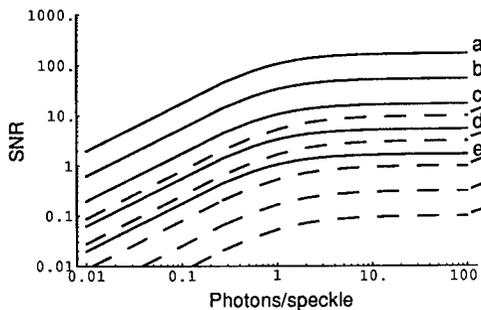
Fig. 1.44 The SNR versus number of photons per speckle. Solid lines: $|\mu|^2 = 0.25$; dashed lines: $|\mu|^2 = 0.01$. Number of realizations (top to bottom): (a) 10,000; (b) 1000; (c) 100; (d) 10; and (e) 1.
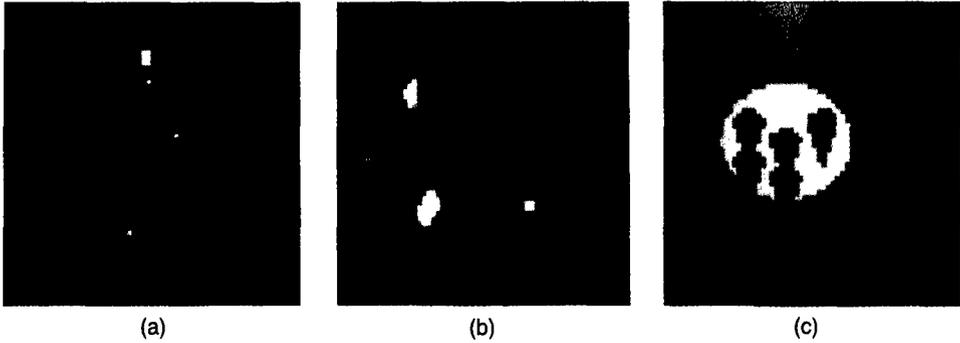
Fig. 1.45 Three types of objects: (a) four points—a 2 × 4-pixel rectangle plus three single-pixel points; (b) satellite—an object with glints; and (c) bus/RV—an extended diffuse object.
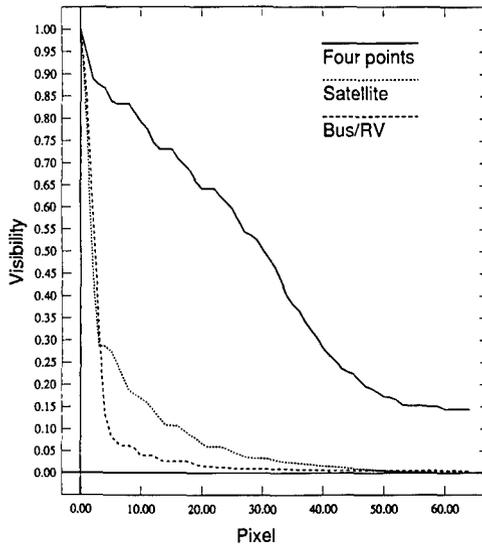
Fig. 1.46 Visibility (Fourier modulus) $|\mu|$ as a function of spatial frequency.

The SNR is approximately proportional to $|\mu(u,v)|^2$, which heavily depends both on (u,v) and on the structure of the object being imaged. At one extreme, an object consisting of a single unresolved glint (a delta function) has $|\mu(u,v)|$ = 1.0 for all (u,v). At the other extreme, a diffuse, constant-value object that fills the field of view will have $|\mu(u,v)|$ that is a delta function and is negligibly small for $(u,v) \neq (0,0)$. To give you a feel for the values that $|\mu(u,v)|$ takes on for different types of objects, Fig. 1.45 shows three types of objects and Fig. 1.46 shows $|\mu(u_r)|$, the spin-averaged (azimuthally averaged) values of the visibility $|\mu(u,v)|$, where $u_r = (u^2 + v^2)^{1/2}$. A very rough rule of thumb is that

$$|\mu(u_r)|^2 \approx \frac{1}{n_{bp}} \ , \tag{1.172}$$

where n_{bp} is the number of resolved bright image points at resolution u_r. For an extended object n_{pb} is the number of 2-D resolution elements within the area of the object. So if, as the resolution improves with increasing u_r, more points in the diffraction-limited image of an object are resolved, then $|\mu(u_r)|^2$ decreases correspondingly.

Ideally the laser should have sufficient power so that the number of detected photons per speckle exceeds two. For a simplified object consisting of a flat plate that is a Lambertian reflector, the average number of photons per speckle is given by

$$M n_d = E_L \tau_{atm}^2 \eta_{LA} (\cos\theta_i \ \cos\theta_o/\pi) r_o \tau_{pol} \tau_{opt} (\eta_d d_a^2/R^2) \eta_q (\lambda/hc) \ , \tag{1.173}$$

where

E_L = energy of the laser per pulse or per detector integration time
τ_{atm} = one-way atmospheric transmittance
η_{LA} = fraction of laser intensity falling on the area of the object
θ_i = angle between object surface normal and illumination
θ_o = angle between object surface normal and LOS
r_o = average reflectivity of object
τ_{pol} = fraction of light reflected in desired polarization
τ_{opt} = transmittance of receiver optics (if any)
$\eta_d d_a^2$ = area of one detector
η_q = quantum efficiency of the detector
hc/λ = energy of one photon (Planck's constant times frequency).

For example, for the parameters of Table 1.3 and for $E_L = 1$ J/pulse, $\tau_{atm}^2 \tau_{pol} \tau_{opt} = 0.2$, $\eta_{LA} = 0.1$, $\theta_i = \theta_o = 0$, $r_o = 0.5$, $\eta_d d_a^2 = 0.005$ m^2, $\eta_q = 0.1$, and $hc/\lambda = 2 \times 10^{-19}$ J/photon, we have $M n_s = 8$ photons per speckle per pulse, which is in the high light-level regime.

If the receiver is ground based, then atmospheric scintillation could be a limiting factor, since it directly changes the measured aperture-plane intensity. However, since many speckle patterns are averaged, the effects of scintillation tend to cancel and are not expected to be a limiting factor for imaging objects directly overhead.

1.5.5 Conclusion

The use of laser illumination opens the possibility for many novel ways to image objects with resolution one to two orders of magnitude finer than conventional imaging systems that are limited by the blurring effects due to atmospheric turbulence. These imaging methods, which spring from the heritage of holography, are especially useful for obtaining fine-resolution images of earth-orbiting satellites. These same methods can also be applied to other imaging scenarios, ranging from microscopy to imaging of the ground from space. While these methods have been investigated and demonstrated to varying degrees of sophistication, they are just now on the verge of becoming operationally feasible.

1.6   OPTICAL RANGE-DOPPLER IMAGING
Written by Carl C. Aleksoff

Range-Doppler imaging uses a combination of coherent ranging measurements and Doppler measurements to produce image-like displays of an object. The Doppler measurements might better be described as range-rate measurements that are available due to relative motion between the object, source(s), and receiver(s). Synthetic aperture imaging or inverse synthetic aperture imaging are also terms used to describe the technique. This type of coherent imaging technique has been used successfully with microwave radars for many years.[420–424] Recent experiments with laser radars (ladars) have shown that such techniques are also viable at optical wavelengths.[425–427]

In the following sections we consider two types of coherent range-Doppler imaging techniques. One type is characterized by ambiguity functions and the other by 3-D Fourier transforms. Each has its own requirements on the transmitted waveforms, relative motion, and geometry. Both will be developed only for imaging rotating targets.

Our objective is to present some of the unique characteristics of range-Doppler imaging. The classical aspects of ladars, such as power requirements and SNRs are found elsewhere in this Handbook.

1.6.1   Basic System Operation

The basic ladar system under consideration is illustrated in Fig. 1.47. This generalized system has separate lasers for the transmitter and receiver sections. The transmitter laser is the source for target illumination and the receiver laser is used to generate a reference signal for heterodyne detection. A more typical system would have a common laser oscillator with amplifiers to furnish the light required for both sections. These transmitter/receiver system variations are described in more detail in the laser radar section of this Handbook. For our purposes here, we will assume that the sources can be operated coherently, i.e., the lasers are phase stable with respect to each other.

Consider illumination of the target/scene by the transmitted waveform $E_T(\mathbf{r},t)$ where $\mathbf{r}$ is a position vector at the target and t is time. We assume that the waveforms are scalar analytic signals describing the electric field. In general, a complex vector (polarization) analysis would be more exact, but for our purposes of describing the range-Doppler technique, it is adequate for us to consider the waveforms as being of one polarization mode at all times, e.g., all waveforms are linearly polarized in the same direction.

Let the received waveform in the plane of the detector be described by $E_R(\mathbf{r}_D,t)$ where $\mathbf{r}_D$ is a position vector along the detector surface. A reference waveform given by

$$H(\mathbf{r}_D,t)\,\exp(2\pi i \nu_{\mathrm{IF}} t) \ , \tag{1.174}$$

is also propagated to the detector, where ν_{IF} is the intermediate frequency (IF) offset. The resulting instantaneous intensity at the detector is given by
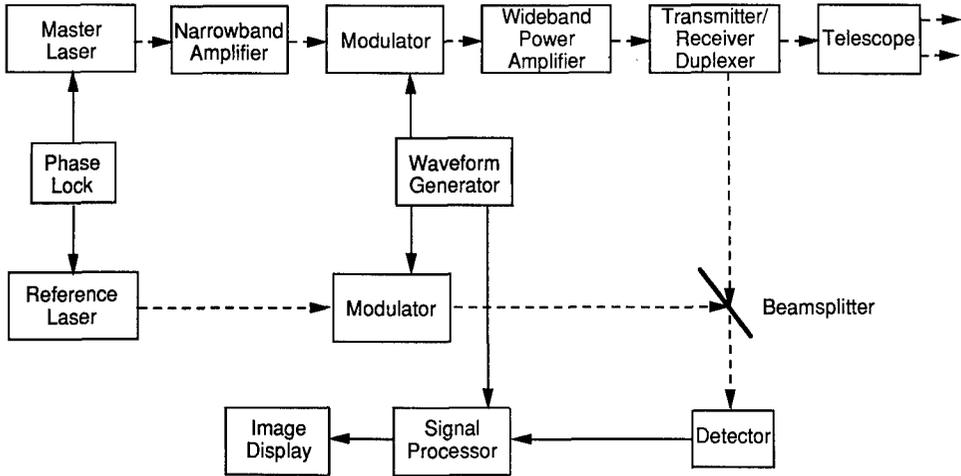
Fig. 1.47   Basic ladar system.

$$I(\mathbf{r}_D, t) = |E_R(\mathbf{r}_D, t) + H(\mathbf{r}_D, t) \exp(2\pi i \nu_{\mathrm{IF}} t)|^2$$

$$= |E_R|^2 + |H|^2 + E_R H^* \exp(-2\pi i \nu_{\mathrm{IF}} t)$$

$$+ E_R^* H \exp(2\pi i \nu_{\mathrm{IF}} t) \ .$$

(1.175)

The electrical signal from the detector is given by the integration of the instantaneous intensity over the surface of the detector convolved with its time response function. For our purposes, it is adequate to assume a point detector of sufficiently fast response (i.e., adequate bandwidth) to detect the instantaneous intensity, or equivalently, the difference frequencies between the received signal and the reference signal. Thus, the output electrical signal from the detector is taken to be proportional to the instantaneous intensity at a point, as given by Eq. (1.175). We again refer the reader to the laser radar chapter in this Handbook for information about actual system characteristics and the intricacies of heterodyne detection as implied by Eq. (1.175).

The detector output signal can be processed in a number of different ways, depending on the desired results. The processing can be a mix of analog and digital techniques to remove bias signals and to form correlation terms. The techniques used depend on the bandwidths and IF signals that need to be processed.

For phase-coherent processing, the signal is electronically mixed with a coherent electrical IF signal in order to isolate (bandpass filter) one of the sidebands with respect to the other parts of the signal. We take the processed output to be proportional to

$$S_{\mathrm{coh}}(t) = E_R H^* \ .$$

(1.176)

Here, we have implicitly assumed that the optically generated ν_{IF} is sufficiently greater than the received signal bandwidth so that we can cleanly isolate (filter) the signal to obtain the single-sideband signal described by Eq. (1.176).

We also have the possibility of envelope detection where the optical phase information is lost. The processed output signal is then given by

$$S_{env}(t) = |E_R H^*| \ . \tag{1.177}$$

Also, for direct detection where there is no reference wave, the output signal is then described by

$$S_{dir}(t) = |E_R|^2 \ . \tag{1.178}$$

1.6.2 Received Signal Description

Range-Doppler imaging depends on the relative motion of the transmitter/ receiver with respect to the target. Although images can be generated with a number of different multistatic arrangements where the transmitter(s) and/ or receiver(s) and target are all moving, the most basic arrangement, and the one for which range-Doppler imaging has been implemented, is a monostatic system imaging a rotating object, such as a satellite. It is the rigid rotation of a target that allows the range and Doppler (range-rate) information to be mapped out in a way to generate the desired image. We consider only this latter arrangement in detail, as indicated in Fig. 1.48, where a rotating target is to be imaged with a ladar and the ladar consists of a collocated transmitter and receiver.

Let the transmitted wave have the time dependence given by $U(t) \exp(2\pi i \nu_c t)$, where $U(t)$ is the complex time envelope of the signal and ν_c is the optical carrier frequency. At the target the electric field amplitude is described by

$$E_T(\mathbf{r},t) = A(\mathbf{r})U(t - \tau) \exp[2\pi i \nu_c(t - \tau)] \ , \tag{1.179}$$

where

   $\mathbf{r}$ $= r\hat{r}$ is the target position vector (fixed with respect to the target)
   $A(\mathbf{r})$ = spatial beam pattern weighting at the target
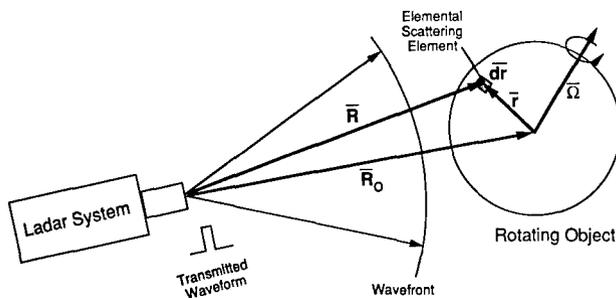
Fig. 1.48 Geometry for ladar range-Doppler imaging.

$\tau \quad = R/c$ is the time delay from the transmitter to the target

$\mathbf{R} \quad = R\hat{R}$ is the vector from the transmitter to the target position $\mathbf{r}$.

Consider an elemental scattering volume located at $\mathbf{r}$. The received field from this element is

$$dE_R = C_1 R^{-1}\sigma(\mathbf{r})E_T(\mathbf{r}, t - \tau) \, d\mathbf{r} \ , \tag{1.180}$$

where C_1 is a (complex) constant, and $\sigma(\mathbf{r})$ is the field scattering function describing the target. Notice that we have assumed for simplicity that the elemental scatterer produces an isotropic spherical wave.

Integrating over all the target scattering elements gives the received signal:

$$E_R(t) = C_1 \int A(\mathbf{r})R^{-1}\sigma(\mathbf{r})U(t - 2\tau) \exp[2\pi i v_c(t - 2\tau)] \, d\mathbf{r} \ . \tag{1.181}$$

We now make a number of simplifying assumptions. First, we consider the target to be in the far field such that we can use

$$R = R_0 + \mathbf{r} \cdot \hat{R}_0 \ , \tag{1.182}$$

where $\overline{\mathbf{R}}_0 = R_0\hat{R}_0$ is the LOS vector from the transmitter/receiver to the target coordinate center.

Consistent with far field operation, $A(\mathbf{r})$ will be made a constant, i.e., we are assuming uniform illumination, and R^{-1} is also assumed not to change significantly for the given geometry. Thus, we can write Eq. (1.181) as

$$E_R(t) = C_2 \exp[2\pi i v_c(t - 2\tau_0)] \int \sigma(\mathbf{r})U(t - 2\tau_0 - 2\tau_r)$$

$$\times \exp(-2\pi i v_c 2\tau_r) \, d\mathbf{r} \ , \tag{1.183}$$

where $\tau_0 = R_0/c$ is the time delay to the target coordinate origin, and $\tau_r = \mathbf{r} \cdot (\hat{R}_0/c)$ is the time delay or advance along the LOS direction to $\mathbf{r}$.

We now expand the time delay τ_r in the Taylor power series

$$\tau_r = \tau_{r0} + a(t - \tau_0) + \dots \tag{1.184}$$

where

$$\tau_{r0} = \tau_r \Big|_{t=\tau_0} \tag{1.185}$$

and

$$a = \frac{\partial \tau_r}{\partial t}\Big|_{t=\tau} \ . \tag{1.186}$$

Of course, τ_{r0} and a are functions of $\mathbf{r}$. The type of waveforms that are useful in range-Doppler imaging have a relatively slowly varying (complex) envelope in which we can replace τ_r by τ_{r0}. However, the rapidly varying exponential carrier phase requires that τ_r be replaced by $\tau_{r0} + a(t - \tau_0)$. The result is that Eq. (1.183) becomes

$$E_R(t) = C_2 \exp(2\pi i \nu_c t') \int \sigma(\bar{\mathbf{r}}) \exp(-2\pi i 2\nu_c \tau_{r0})$$

$$\times\; U(t' - 2\tau_{r0}) \exp[-2\pi i \nu_d(t - \tau_0)]\, d\mathbf{r} \;, \qquad (1.187)$$

where we have let

$$t' = t - 2\tau_0 \;, \qquad (1.188)$$

and where

$$\nu_d = 2\nu_c a \qquad (1.189)$$

is the Doppler frequency.

The reference signal is used to remove the optical carrier and insert the IF for electrical processing. We use

$$H^*(t) = \exp[-2\pi i \nu_c(t - 2\tau_0)] \;, \qquad (1.190)$$

in Eq. (1.175) to get the electrical signal $S_{\mathrm{coh}}(t)$ as given by Eq. (1.176). We point out that knowing the value of $2\tau_0$, the time delay or equivalently the distance to the target, accurately enough when the target is moving may require accurate tracking with another sensor or further processing to correct the error.

The $S_{\mathrm{coh}}(t)$ signal is then typically matched filtered. Matched filtering maximizes the SNR in the presence of additive white Gaussian noise. Matched filtering is basically a correlation of the received signal with a Doppler-shifted conjugated version of the transmitted waveform. Thus the matched filtered output signal is given by

$$S(\tau',\nu') = \int E_R(t) H^*(t) U^*(t' - \tau') \exp[2\pi i \nu'(t' - \tau')]\, dt \;, \qquad (1.191)$$

which, on substitution from Eq. (1.187), gives

$$S(\tau',\nu') = C_3 \int \sigma(\mathbf{r}) K \chi(2\tau_{r0} - \tau',\, \nu_d - \nu')\, d\mathbf{r} \;, \qquad (1.192)$$

where the so-called "ambiguity function" is given by[422–424]

$$\chi(\tau,\nu) = \int U(t) U^*(t + \tau) \exp(-2\pi i \nu t)\, dt \;, \qquad (1.193)$$

and

$$K = \exp\{-2\pi i[\nu'\tau' + (\nu_c - \nu_d - \nu')2\tau_{r0}]\} \;. \qquad (1.194)$$

The ambiguity function is a key function for determining the image characteristics. However, before we investigate its properties, we will consider how Eq. (1.192) applies to the rotating target.

Consider a rotating target described by the angular rotation vector $\mathbf{\Omega}$. Let the origin of the target position vector $\mathbf{r}$ lie on the rotation axis. Let the z axis lie along the LOS direction $\hat{R}_0$ as indicated in Fig. 1.49. We can also let the y axis be perpendicular to the z axis and the rotation axis. The x axis is then determined by being perpendicular to both the z axis and y axis. We note that the coordinate system (x,y,z) does not rotate with the target. That is, it is not fixed with respect to the target, but rotates to keep aligned with the LOS direction and the axis of rotation. In this coordinate system, the scatter at $\mathbf{r}$ is given by the parametric equations

$$x(t') = r \cos[\Omega t' + \theta_0(\mathbf{r})] \sin\eta \cos\varepsilon + r \cos\eta \sin\varepsilon \ , \tag{1.195}$$

$$y(t') = r \sin[\Omega t' + \theta_0(\mathbf{r})] \sin\eta \ , \tag{1.196}$$

$$z(t') = r \cos[\Omega t' + \theta_0(\mathbf{r})] \sin\eta \sin\varepsilon + r \cos\eta \cos\varepsilon \ , \tag{1.197}$$

where $\theta_0(\mathbf{r})$ is the position of the scatterer $\mathbf{r}$ at $t' = 0$ measured from the x-z plane, η is the polar angle for $\mathbf{r}$, and ε is the tilt of the rotation axis from the LOS (i.e., the z axis). The time delay is given by

$$\tau_{r0} = \left. \frac{z(t')}{c} \right|_{t'=0} = \frac{2z_0}{c} \ , \tag{1.198}$$

where $z_0 = z(0)$. The Doppler frequency is given by Eqs. (1.189) and (1.186) to be

$$\nu_d = 2\nu_c \left. \frac{\delta\tau_r}{\delta t'} \right|_{t'=0}$$

$$= \frac{2\nu_c}{c} \left. \frac{\delta z(t')}{\delta t'} \right|_{t'=0} \ , \tag{1.199}$$
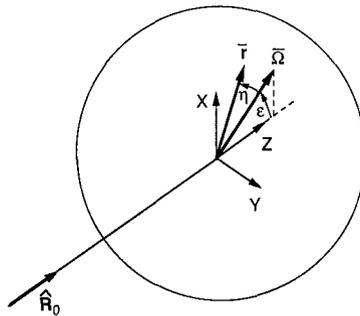
Fig. 1.49 Coordinates for imaging the rotating target.

which from Eqs. (1.196) and (1.197) we find that

$$v_d = \frac{2v_c \Omega \sin\varepsilon}{c} y_0 \; ,$$

(1.200)

where $y_0 = y(0)$. Equations (1.198) and (1.200) indicate that delay is proportional to the z axis coordinate of the scatterer and Doppler frequency is proportional to the y axis position of the scatterer. There is no x axis dependence, which means that we can integrate (at time $t' = 0$) over the x direction to get

$$\sigma(z_0, y_0) = \int \sigma(\mathbf{r}) \, dx \; .$$

(1.201)

Thus, in effect, all the scatterers are projected down to the y-z plane, which causes the so-called "layover" effect. Substitution of Eqs. (1.198), (1.200), and (1.201) into Eq. (1.192) gives

$$S(\tau', v') = C_2 \int \int \sigma(z_0, y_0) K \chi \left(\frac{2z_0}{c} - \tau', \frac{2\Omega v_c \sin\varepsilon}{c} y_0 - v' \right) dz_0 \, dy_0 \; .$$

(1.202)

This correlation equation is the basic result. It shows that *the ambiguity function plays the role of a point spread function for the range-Doppler image description*. Notice that the range-Doppler image is two dimensional and that the Doppler scale depends on the rotation rate. The K function describes a rapid phase variation across the object that will cause speckle-like images.

1.6.3   The Ambiguity Function

The ambiguity function, since it plays the role of the point spread function for the range-Doppler image, describes the basic characteristics of the image. It allows the evaluation of image resolution, ambiguities, and other image qualities. Hence, we will now consider the ambiguity function in more detail. However, we first point out that sometimes the function $|\chi(\tau,v)|^2$ is called the *ambiguity function* in the literature, but we will reserve the term *ambiguity function* for $\chi(\tau,v)$.

The ideal ambiguity function would be a delta function $\chi(\tau,v) = \delta(\tau)\delta(v)$, which would from Eq. (1.202) give the detected image

$$|S(\tau,v)|^2 = C \left| \sigma\left(\frac{c\tau}{2}, \frac{cv}{2\Omega v_c \sin\varepsilon} \right) \right|^2 \; .$$

(1.203)

However, the theory of ambiguity functions indicates that the ideal δ function response is not possible.[422-424] The fact that the ambiguity function is formed by the correlation of a waveform with a phase shifted version of itself limits the generality of the function and produces the following characteristics.[422-424] The first characteristic is that the peak value is always at the origin, i.e.,

$$\chi(0,0) \geq |\chi(\tau,\nu)| \ . \tag{1.204}$$

Next, the volume under $|\chi(\tau,\nu)|^2$ is a constant independent of the waveform shape, i.e.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |\chi(\tau,\nu)|^2 \ d\tau \ d\nu \ = \ |\chi(0,0)|^2 \ . \tag{1.205}$$

These characteristics imply that the best the waveform design can do is to move the inevitable volume around to other regions of the $\tau - \nu$ space so that they don't degrade the imaging characteristics desired. There are three general shapes of ambiguity functions: the thumbtack, bed-of-spikes, and knife edge, as illustrated in Fig. 1.50. A train of pulses can improve the Doppler resolution over that of a single pulse but at the expense of producing a bed of spikes that produces ambiguous images. A modulated pulse can produce the ridge response. A random phased pulse train can produce the thumbtack response.

Let us next consider the resolution in delay-Doppler space, which by the scale change implied in Eq. (1.202) can be related to actual image resolution. The classical definition of resolution would be defined by the full-width half-maximum (FWHM) of the central main lobe of the intensity (magnitude squared) response $|\chi(\tau,\nu)|^2$. Realizing that another characteristic of the ambiguity function is the symmetry $|\chi(\tau,\nu)| = |\chi(-\tau,-\nu)|$, then we can expand about the central point to get[422–424]

$$|\chi(\tau,\nu)| = |\chi(0,0)|\left[1 - \left(\frac{\tau^2}{\rho_\tau^2} + \frac{2\tau\nu}{\rho_{\tau\nu}^2} + \frac{\nu^2}{\rho_\nu^2}\right) + ...\right] \ . \tag{1.206}$$

The resolution ellipse defined by

$$\frac{\tau^2}{\rho_\tau^2} + \frac{2\tau\nu}{\rho_{\tau\nu}^2} + \frac{\nu^2}{\rho_\nu^2} = \frac{1}{4} \ , \tag{1.207}$$
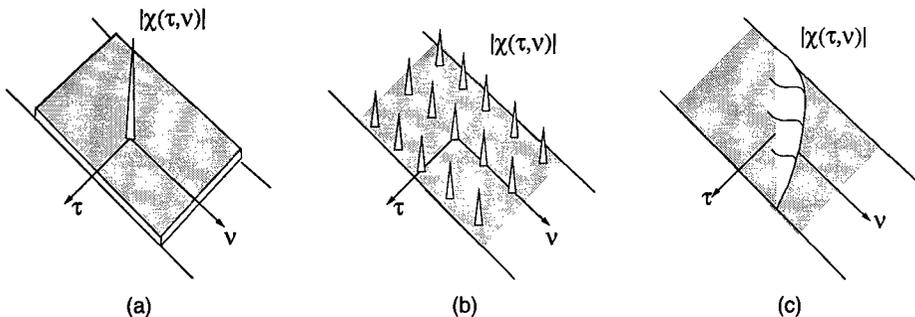
Fig. 1.50   General ambiguity function shapes[424]: (a) thumbtack, (b) bed-of-spikes, and (c) knife-edge.

$$\zeta = \text{Arctan} \frac{\rho_{\tau\nu}^{-2}}{\rho_{\tau}^{-2} - \rho_{\nu}^{-2}}$$

$$\rho_a^{-2} = \rho_{\tau}^{-2} \cos^2(\zeta) + \rho_{\tau\nu}^{-2} \cos(\zeta)\sin(\zeta) + \rho_{\nu}^{-2}\sin^2(\zeta)$$

$$\rho_b^{-2} = \rho_{\tau}^{-2} \sin^2(\zeta) - \rho_{\tau\nu}^{-2} \cos(\zeta)\sin(\zeta) + \rho_{\nu}^{-2}\cos^2(\zeta)$$
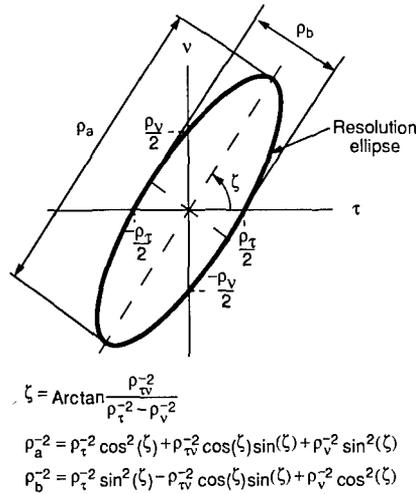
Fig. 1.51 The resolution ellipse parameters.

describes the directional aspects of the FWHM resolution (see Fig. 1.51). In general, it is a tilted ellipse at the angle ζ with major and minor axis resolutions of ρ_a and ρ_b, respectively. The delay resolution is given by

$$\frac{1}{\rho_{\tau}^2} = B_{\text{eff}}^2 = \frac{1}{2\chi(0,0)} \int_{-\infty}^{\infty} (2\pi\nu)^2 |u(\nu)|^2 \, d\nu \ , \tag{1.208a}$$

where B_{eff} is the effective bandwidth of the signal and the Fourier transform (spectrum) of the waveform is

$$u(\nu) = \int_{-\infty}^{\infty} U(t) \exp(-2\pi i\nu t) \, dt \ . \tag{1.208b}$$

The Doppler resolution is given by

$$\frac{1}{\rho_{\nu}^2} = T_{\text{eff}}^2 = \frac{1}{2\chi(0,0)} \int_{-\infty}^{\infty} (2\pi t)^2 |U(t)|^2 \, dt \ , \tag{1.208c}$$

where T_{eff} is the effective time duration of the signal. The cross-coupling resolution is given by

$$\frac{1}{\rho_{\tau\nu}} = \frac{-1}{2\chi(0,0)} \text{Re}\left[\int_{-\infty}^{\infty} (2\pi i t) U(t) \frac{\partial U^*(t)}{\partial t} dt \right] \ . \tag{1.208d}$$

where Re indicates the real-part operation. In the above equations we have assumed that the waveform $U(t)$ has a zero mean time and zero mean fre-

quency. Then a basic relationship for the resolution ellipse can be shown to be[422,423]

$$\rho_\tau^{-2}\rho_\nu^{-2} - \rho_{\tau\nu}^{-2} \geq \frac{\pi^2}{4} \, , \tag{1.209a}$$

or in terms of the principal resolutions

$$\rho_a\rho_b \leq \frac{2}{\pi} \, . \tag{1.209b}$$

These relationships give the upper limits on the resolution product or area of the resolution ellipse.

With respect to the rotating target, we can now say that the range resolution is given by

$$\rho_{R_0} = \frac{c}{2}\rho_\tau \, , \tag{1.210a}$$

and the cross-range resolution by

$$\rho_y = \frac{c}{2\Omega\nu_c \sin\varepsilon}\rho_\nu = \frac{\lambda_c}{2\Omega \sin\varepsilon}\rho_\nu \, . \tag{1.210b}$$

However, it should be clear from the preceding discussion that these resolutions, or even the resolution ellipse, are not the whole story, but that the overall character of the image is described by the entire ambiguity function.

Example 1.1: A Single Rectangular Pulse Waveform with Linear FM. A single rectangular pulse of duration T with linear FM (chirp) is defined by

$$U(t) = \frac{1}{\sqrt{T}} \text{rect}\left(\frac{t}{T}\right) \exp(i\pi\gamma t^2) \, , \tag{1.211}$$

where

$$\text{rect}(t) \triangleq \begin{cases} 1 & |t| < \frac{1}{2} \\ 0 & \text{elsewhere} \end{cases} \, , \tag{1.212}$$

and γt is the instantaneous modulation frequency.

The ambiguity function is given by

$$\chi(\tau,\nu) = \text{rect}\left(\frac{\tau}{2T}\right) \exp(i\pi\nu\tau)\frac{T - |\tau|}{T} \text{sinc}[(\nu + \gamma\tau)(T - |\tau|)] \, , \tag{1.213}$$

where $\text{sinc}(t) = \sin(\pi t)/(\pi t)$.

Figure 1.52 shows magnitude surface plots and contours for two cases of the rectangular pulse ambiguity function. The first case, shown in Fig. 1.52(a),
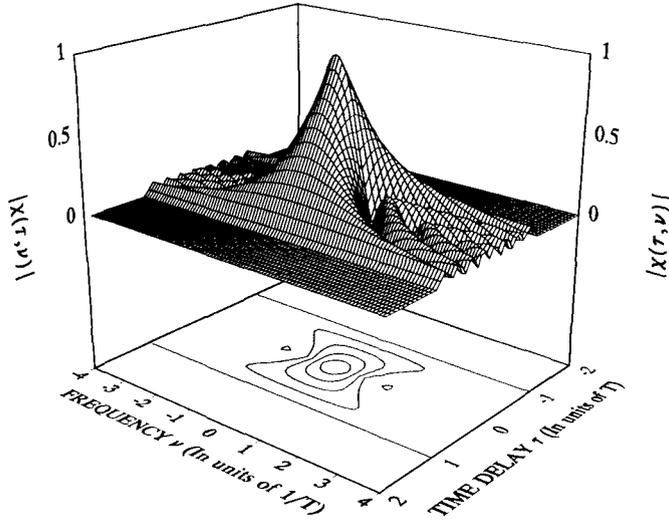
(a)
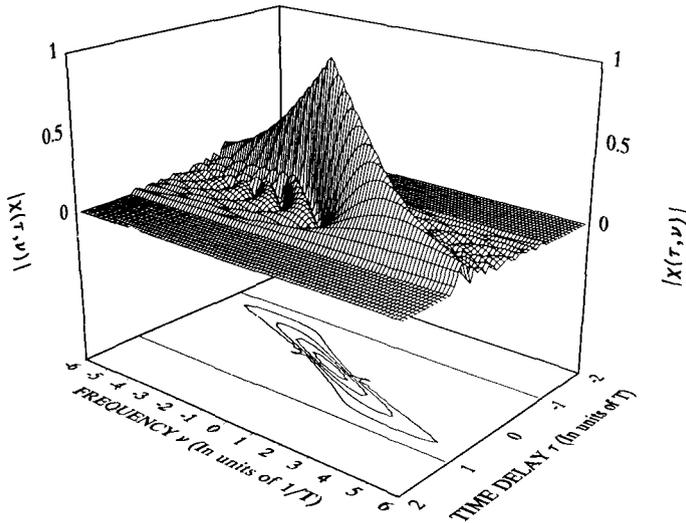
(b)

Fig. 1.52 Rectangular pulse ambiguity function: (a) The magnitude of the ambiguity function for an unmodulated rectangular pulse. The contours are separated by 0.2 from 0 to 1. (b) The magnitude of the ambiguity function for a linear FM rectangular pulse with $\gamma = 5/T^2$. The contours are separated by 0.2 from 0 to 1.

is a surface plot and a contour plot of the magnitude of the ambiguity function for no FM, i.e., $\gamma = 0$, and hence is a simple unmodulated rectangular pulse. For this case the FWHM delay resolution is about T and the FWHM Doppler resolution is about $1/T$. Hence, a longer pulse has finer Doppler resolution but a coarser delay resolution than that of a shorter pulse. But, of course, a shorter pulse has coarser Doppler resolution while exhibiting a finer delay resolution than that of a longer pulse. It is clear that indefinitely fine resolution in both delay and Doppler cannot be achieved with a single unmodulated rectangular pulse.

Figure 1.52(b) shows a surface plot and a contour plot for the magnitude of the ambiguity function where the FM linear modulation is characterized by $\gamma = 5/T^2$. It is clear that the resolution ellipse is now rotated from that of the case of no modulation. The principal axes of the ellipse are rotated as described by the line $\gamma t + \nu = 0$. The Doppler resolution has not changed from the unmodulated case, but the delay resolution has become finer by approximately the time-bandwidth product γT^2 at the expense of a tilted resolution ellipse.

Example 1.2: A Train of Identical Pulses.   Consider a pulse train of N identical pulses given by

$$
U(t) = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} V\left[t - \left(n - \frac{N+1}{2} \right) T_p \right] ,
\tag{1.214}
$$

where $V(t)$ describes a normalized single pulse of zero time and frequency mean, and limited time length of T. The term $\sqrt{N}$ normalizes the pulse train and the term $(N+1)T_p/2$ provides zero time mean for the pulse train. We also assume that $T_p > T$. The spectrum is given by

$$
u(\nu) = \frac{v(\nu)}{\sqrt{N}} \frac{\sin(N \pi \nu T_p)}{\sin(\pi \nu T_p)} ,
\tag{1.215}
$$

where $v(\nu)$ is the Fourier transform (spectrum) of $V(t)$. The ambiguity function is given by

$$
\chi(\tau,\nu) = \sum_{n=-(N-1)}^{N-1} \chi_v(\tau - n T_p, \nu)
$$

$$
\times \exp(i n \pi \nu T_p) \frac{\sin[(N - |n|)\pi \nu T_p]}{N \sin(\pi \nu T_p)} ,
\tag{1.216}
$$

where $\chi_v(\tau,\nu)$ is the ambiguity function for $V(t)$. We see that there are $2N - 1$ nonzero stripes parallel to the ν axis.

Thus, if χ_v is the ambiguity function for an unmodulated rectangular pulse as given in Example 1.1, then the ambiguity function for the train of five rectangular pulses has the character of a bed of spikes as shown in Fig. 1.53. The FWHM delay resolution around the central spike is still about T, but the Doppler resolution is now about $1/NT_p = 1/5T$. Thus, the Doppler resolution is improved by a factor of NT_p/T at the expense of producing ambiguities, i.e.,
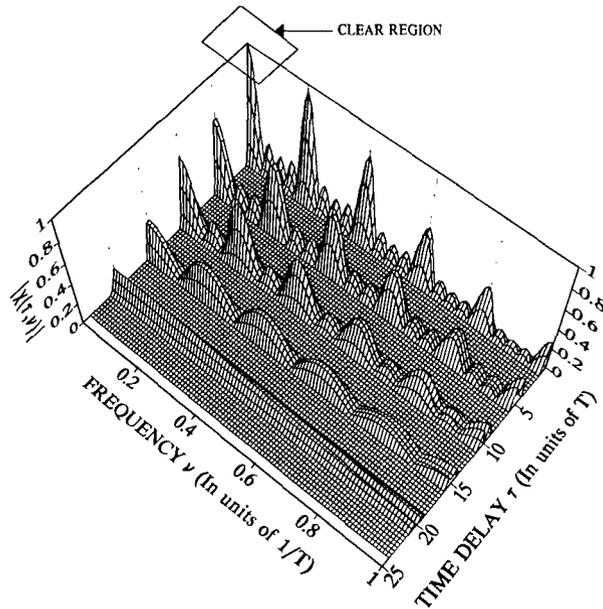
Fig. 1.53 The ambiguity function for a train of five unmodulated rectangular pulses. Only one quadrant is shown since the magnitude of the ambiguity function for this case is even symmetric in both the time delay and frequency directions. The interpulse period $T_p = 5T$.

an extra image is produced about each extra spike, and the extent of an uncontaminated image is limited to the clear region about the central spike. From Eq. (1.216), and as is evident in Fig. 1.53, we see that the ambiguity spikes are spaced at $1/T_p$ apart in the Doppler direction and T_p in the delay direction.

A large number of waveforms have been considered for range-Doppler radars.[422–424] However, only a few are used in practice. Table 1.4 lists some of the properties of potentially useful ladar waveforms.[427] Also, the effects of waveform errors on the range-Doppler ladar performance have been evaluated.[428] Such waveform error could come from the generation of the waveforms or from propagation in the atmosphere.

Table 1.4   Waveform Comparison

Waveform & (parameters)	Measurement	Performance		
		Resolution	Ambiguity Peaks Separation	Sidelobe to Mainlobe Ratio
unmodulated rect pulse (T)	Range	cT/2	none	0
	Doppler	1/T	none	-13 dB
unmodulated rect pulse train (T, N, T_p)	Range	cT/2	$cT_p/2$	0
	Doppler	$1/(NT_p)$	$1/T_p$	-13 dB
Linear FM chirp rect pulse (T, B)	Range	c/2B	none	-13 dB
	Doppler	1/T	none	-13 dB
Linear FM chirp rect pulse train (T, B, N, T_p)	Range	c/2B	$cT_p/2$	-13 dB
	Doppler	$1/(NT_p)$	$1/T_p$	-13 dB

T ≜ Individual pulse length

T_p ≜ Pulse period interval

N ≜ Number of pulses in pulse train

B ≜ Total sweep bandwidth of the linear FM chirp

1.6.4   Range-Doppler Imaging Limitations

In this section we consider some trade-offs between parameters as established by geometry and waveform constraints. Consider the rotating target space as being divided into range-Doppler resolution cells as illustrated in Fig. 1.54.

In the range direction the cell width is given by the range resolution

$$\rho_{R0} = \frac{c}{2B_{\text{eff}}} , \qquad (1.217)$$

where B_{eff} is the effective bandwidth. In the cross-range (Doppler) direction the width of the cells for $\varepsilon = \pi/2$ is given by

$$\rho_y = \frac{c}{2\Omega \nu_c T_{\text{eff}}} = \frac{\lambda_c}{2\Omega T_{\text{eff}}} , \qquad (1.218)$$

where T_{eff} is the effective time duration and λ_c is the carrier wavelength.

For the resolution cells to make sense, a target scatterer must not travel further than the assumed cell size during a waveform illumination time, otherwise image smearing will occur and the assumed resolution is not obtained. The maximum travel l over time T is approximately

$$l = r_{\max}\Omega T , \qquad (1.219)$$
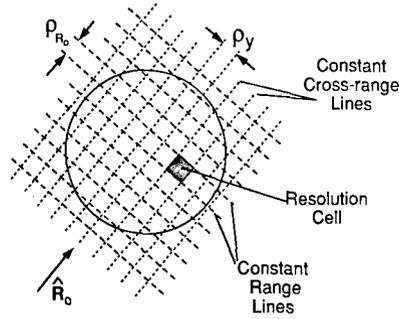
Fig. 1.54   Range-Doppler resolution cell division of the target.

where r_{max} is the radius of the most outer (from the center of rotation) scatterer. Hence, we require that $l < \rho_y$, or with substitution from Eqs. (1.218) and (1.219), and using $T_{eff} = T$,

$$\rho_y^2 > \frac{r_{max}\lambda_c}{2} \; . \tag{1.220}$$

Also, using $l < \rho_{R_0}$ and Eqs. (1.218) and (1.219), we have

$$\rho_{R_0}\rho_y > \frac{r_{max}\lambda_c}{c} \; . \tag{1.221}$$

Thus, we see the limitation and coupling between the size of a target and resolution. If we let the range and cross-range resolution both be equal to ρ, then we obtain the following from Eq. (1.221):

$$\rho^2 > \frac{r_{max}\lambda_c}{2} \; . \tag{1.222}$$

Notice that this resolution limit is independent of the rotation rate, although the time duration of the waveform required to obtain the limit must change accordingly. Figure 1.55 shows the resolution limit given by Eq. (1.222) with wavelengths as parameters. In the derivation of the basic range-Doppler imaging relationship of Eq. (1.202), it was assumed that the scattering element produced a uniform isotropic spherical wave. However, over the rotation angle associated with the assumed Doppler resolution, the wavefront may not be uniform but may be divided up into lobes due to specular reflection. We can coherently integrate over only one lobe. Also shadowing may limit the viewing time for any particular scatterer. Thus, the time duration in the cross-range resolution equations must be limited to the actual viewing time of any scatterer.
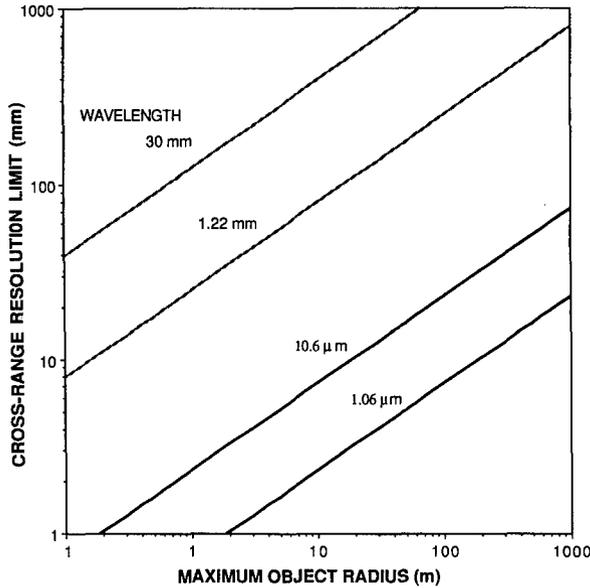
Fig. 1.55   Plot of resolution limit for range-Doppler imaging.

1.6.5   Fourier Transform Imaging

In the previous section we saw that limits exist on the obtainable resolution and target size as dictated by the ambiguity function. In this section we will see that these limitations can be removed and we obtain the ability to produce three-dimensional images rather than two dimensional ones if we transmit a long train of pulse waveforms where the Doppler shift is not significant during any pulse. We will call this technique *Fourier transform synthetic aperture imaging*.

1.6.5.1   Matched Filter Processing.   Consider a scenario in which we illuminate a rotating target with a long train of waveform pulses under the condition that the target rotation is slow enough that no Doppler frequency shift occurs during the scattering of any one of the pulses. The required significant motion takes place between the pulses. If we now correlate each received pulse as before, but drop any Doppler-dependent terms, Eq. (1.192) becomes

$$S(\tau) = C_2 \int \sigma(\mathbf{r}) \chi(2\tau_{r0} - \tau) \exp(-2\pi i \nu_c 2\tau_{r0}) \, d\mathbf{r} \ , \tag{1.223}$$

where the ambiguity function has degenerated into the one-dimensional signal

$$\chi(\tau) = \int U(t) U^*(t - \tau) \, dt \ . \tag{1.224}$$

The next step of processing is to take the Fourier transform of Eq. (1.223) with respect to τ to get

$$s(v) = C_2 \int \sigma(\mathbf{r}) |u(v)|^2 \exp[-2\pi i 2\tau_{r0}(v + v_c)] \, d\mathbf{r} \ , \tag{1.225}$$

where the lowercase letters represent Fourier transforms of functions represented by uppercase letters. This transformed signal can be written

$$s(v) = C_2 |u(v)|^2 \int \sigma(\mathbf{r}) \exp(-2\pi i \mathbf{f} \cdot \mathbf{r}) \, d\mathbf{r} \ , \tag{1.226}$$

where the spatial frequency vector is defined by

$$\mathbf{f} = \frac{2\hat{R}_0}{c} (v_c + v) \ . \tag{1.227}$$

Equation (1.226) shows that the transformed signal $s(v)$ is proportional to a 3-D Fourier transform of the scattering function $\sigma(\mathbf{r})$ weighted by the power spectrum of the transmitted waveform.

We can formally invert Eq. (1.226) to obtain

$$\tilde{\sigma}(\mathbf{r}) = C_2^{-1} \int \frac{s(v)}{|u(v)|^2} \exp(2\pi i \mathbf{f} \cdot \mathbf{r}) \, d\mathbf{f} \ , \tag{1.228}$$

where $\tilde{\sigma}(\mathbf{r})$ indicates an estimate (image) of $\sigma(\mathbf{r})$. The integration is over those positions of $\mathbf{f}$ for which data have been gathered, i.e., the support. The support grows as the sequence of pulses continues to be processed with target rotation. The support is often referred to as the synthetic aperture. Overall then, Eq. (1.228) says that *the image $\tilde{\sigma}(r)$ is obtained by taking the 3-D spatial Fourier transform of the synthetic aperture data normalized by the power spectrum of the transmitted pulse.*

Example 1.3: Repetitive Staircase Waveform Imaging. Consider the transmission of a train of rectangular pulses, which repetitively step through a staircase of frequencies, as illustrated in Fig. 1.56. The (n,k)'th pulse is given by

$$U_{nk}(t) = \frac{U_0}{T} \exp[-2\pi i(v_c + n\Delta v)t] \, \mathrm{rect}\left(\frac{t - nT_1 - kT_2}{T}\right) \ , \tag{1.229}$$

where

$$n = \frac{N-1}{2}, \dots, 0, \dots, \frac{N+1}{2}$$

$$k = \frac{K-1}{2}, \dots, 0, \dots, \frac{K+1}{2} \ ,$$

$$T_2 \geq NT_1 \ .$$

For convenience we have assumed that N and K are odd integers and U_0 is a constant. The power spectrum of the (n,k)'th pulse is
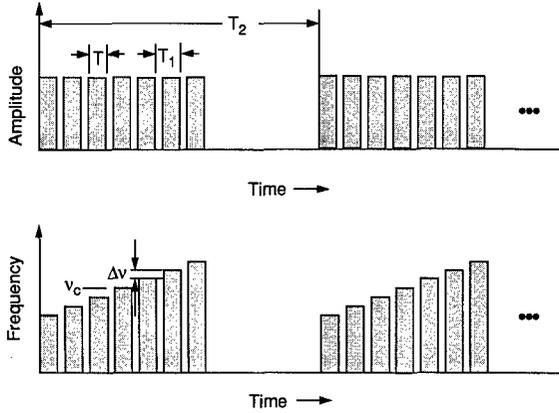
Fig. 1.56   Repetitive staircase waveform.

$$|u_{nk}(\nu)|^2 = U_0^2 \operatorname{sinc}^2[T(\nu - \nu_c - n\Delta\nu)] \ . \tag{1.230}$$

Let $s_{nk}(\nu)$ be the Fourier transform of the processed (n,k)'th pulse as given by Eq. (1.225). We take only one sample for each received pulse, namely, the value at $\nu = \nu_c + n\Delta\nu$. Then we can write Eq. (1.228) in the form

$$\tilde{\sigma}(r) = C_2^{-1}U_0^{-2} \sum_{n=1}^{N} \sum_{k=1}^{K} s_{nk}(n\Delta\nu) \exp(2\pi i \mathbf{f}_{nk} \cdot \mathbf{r}) \ , \tag{1.231}$$

where

$$\mathbf{f}_{nk} = \frac{2\hat{R}_{0nk}}{c}(\nu_c + n\Delta\nu) \ , \tag{1.232}$$

and $\hat{R}_{0nk}$ is the LOS direction with respect to the target coordinates when the (n,k)'th pulse scatters from the target. Thus, for a rotating target with its axis-of-rotation tilted by the angle ε from the LOS direction, the spatial frequency data lie on the surface of a cone with an ε apex angle, as illustrated in Fig. 1.57(a). The s_{nk} data sample is at a radius $\nu_c + n\Delta\nu$ and rotated by $\Omega(nT_1 - T_1 + kT_2)$ from the first sample, where Ω is the rotation rate. The synthetic aperture is the cone surface where the samples exist. The 3-D point response is the 3-D Fourier transform of this synthetic aperture. A conic synthetic aperture tends to give a conic-like spread function.

From the theory of Fourier transforms we know that the inverse of the extent of the data (i.e., the synthetic aperture) along a particular direction in spatial frequency space gives the nominal resolution along that direction in image space. Thus, if the data extent is over a small angle of the cone, as illustrated in Fig. 1.57(b), then the resolution in the z direction is given by
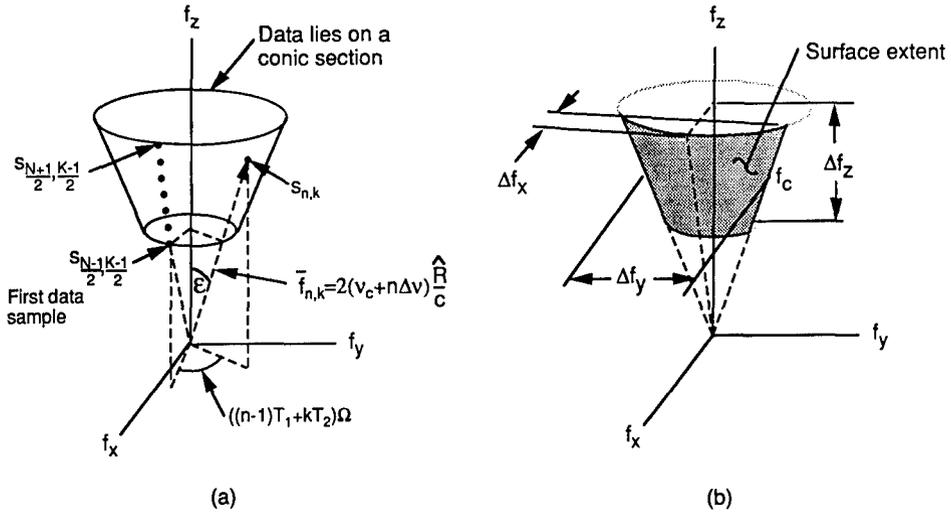
Fig. 1.57 Data in spatial frequency space: (a) data points on a conic section and (b) the extent of the data support.

$$\rho_z = \frac{1}{\Delta f_z} \approx \frac{c}{2B \cos\varepsilon} \,, \tag{1.233}$$

where B is the transmitted bandwidth given by

$$B = (N - 1)\Delta\nu \,. \tag{1.234}$$

Similarly, the resolution in the y direction is given by

$$\rho_y = \frac{1}{\Delta f_y} \approx \frac{c}{2\nu_c\Delta\theta} = \frac{\lambda_c}{2\Delta\theta} \,, \tag{1.235}$$

where

$$\Delta\theta = (NK - 1)T_1\Omega \tag{1.236}$$

is the rotation angle for coherent processing. In the x direction we have

$$\rho_x = \frac{1}{\Delta f_x} \approx \frac{4c}{\nu_c\Delta\theta^2} = \frac{4\lambda_c}{\Delta\theta^2} \,. \tag{1.237}$$

If we have a full 360-deg (or more) of rotation, then the best resolution obtainable is

$$\rho_x = \rho_y = \frac{c}{2(B + \nu_c) \sin\varepsilon} \,. \tag{1.238}$$

1.6.5.2 Stretch Processing.

A particularly attractive technique that has been used extensively in the microwave region and has also been applied at laser wavelengths is known as the *stretch technique*. In this technique the transmitted waveform is a linear FM (chirp) signal given by

$$U(t) = U_0 \exp(\pi i \gamma t^2) , \qquad (1.239)$$

where U_0 is a constant and γ is the chirp rate. The unique difference is that the reference wave now is also a chirp waveform, i.e.,

$$H^*(t) = H_0 \exp[-\pi i \gamma (t - 2\tau_0)^2] , \qquad (1.240)$$

but delayed by the nominal roundtrip time $2\tau_0$. Thus, the received signal for no Doppler becomes

$$S(t') = C_2 \int \sigma(\mathbf{r}) U(t' - 2\tau_{r0}) H^*(t') \exp(-2\pi i \nu_c 2\tau_{r0}) \, d\mathbf{r}$$

$$C_3 \int \sigma(\mathbf{r}) \exp(-4\pi i \gamma \tau_{r0}^2) \exp[-2\pi i 2\tau_{r0}(\nu_c - \gamma t')] d\mathbf{r} , \qquad (1.241)$$

where as before $t' = t - 2\tau_0$. Typically, the residual phase $4\pi i \gamma \tau_{r0}^2$ is small enough to be ignored. We can then write

$$S(t') = C_3 \int \sigma(\mathbf{r}) \exp(-2\pi i \mathbf{f} \cdot \mathbf{r}) \, d\mathbf{r} , \qquad (1.242)$$

where the spatial frequency term is now given by

$$\mathbf{f} = \frac{2\hat{R}_0}{c}(\nu_c - \gamma t') . \qquad (1.243)$$

The formal inversion gives

$$\hat{\sigma}(r) = C_3^{-1} \int S(t') \exp(2\pi i \mathbf{f} \cdot \mathbf{r}) \, d\mathbf{r} . \qquad (1.244)$$

In this case, we need only write the received signal $S(t')$ along the LOS in frequency space as described by $\mathbf{f}$. We can interpret $\mathbf{f}$ as being proportional to the instantaneous frequency of the chirp waveform. Thus, for a rotating target the data lie on straight radial lines on a cone in frequency space. Each received pulse adds another line of data on the cone. As before, the 3-D Fourier transform produces the image.

An advantage of this technique is that no correlation or 1-D Fourier transform is required on the received signal. Further, the range of the instantaneous detected frequency can be considerably less than the range of the FM sweep. Hence, much lower bandwidth receiver circuits can be used. However, we should note that if we allowed the reference waveform to step in frequency with the staircase waveform of Example 1.3, the same reduction in bandwidth can be accomplished. A disadvantage is the additional waveform generation and timing required for the reference waveform.

1.6.6   Operational Systems

Although various aspects of range-Doppler ladar imaging have been demonstrated in the laboratory[426] and analysis has been performed on down-looking (earth-sensing) range-Doppler imaging systems,[429,430] the only actual operating system has been the ground-based Lincoln Laboratory Firepond facility. The facility, illustrated in Fig. 1.58, includes argon ion, ruby, and Nd:YAG laser sources as well as the CO_2 laser range-Doppler imaging source. The range-Doppler imaging system is basically the MOPA (master oscillator power amplifier) arrangement schematically illustrated in Fig. 1.47. The system uses linear FM (chirp) waveforms generated by CdTe electro-optic modulators.[431] The $^{13}C^{16}O_2$ isotope was chosen for the laser in order to operate at a different wavelength than that associated with the absorption line of the common isotope $^{12}C^{16}O_2$ that prevails in the atmosphere. The absorption resonance effect (anomalous dispersion) can produce a significant distortion on a wideband chirp pulse operating at a resonance frequency.[432] The waveform is processed by the stretch technique. The output of the detector is low pass filtered and then digitized. The subsequent processing is then done digitally in I and Q channels to handle the complex data. Fast Fourier transform processing is used to isolate data in range bins and then each range bin is frequency analyzed to isolate the data for each Doppler bin. These data are then mapped into range-Doppler space to form the image.

A demonstration of the Lincoln Laboratory system's capability was presented in 1990 as part of the SDIO Firefly experiment. The system tracked a rocket launched from Wallops Island, Virginia, with the argon ion laser and performed range-Doppler imaging of a decoy deployment with the CO_2 laser system at ranges of 600 to 750 km.
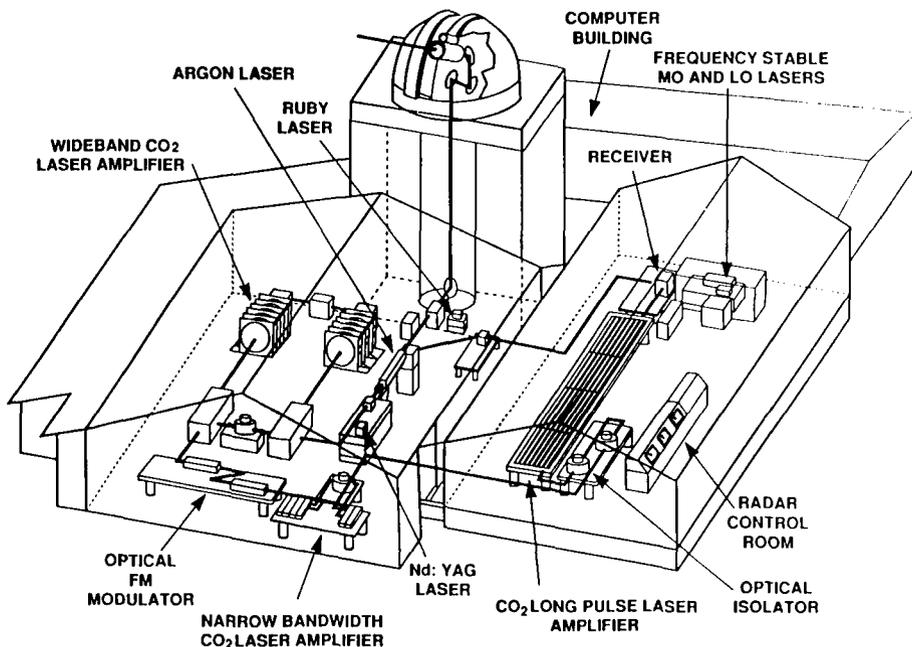
Fig. 1.58   The Lincoln Laboratory Firepond facility. (Reprinted by permission of W. Keicher, MIT Lincoln Laboratory.)

1.7 LASER GUIDE STAR ADAPTIVE OPTICS FOR COMPENSATED IMAGING

Written by Robert Q. Fugate

This section discusses the theory and practical aspects of artificial laser beacon—or laser guide star—adaptive optics for imaging through turbulence. We first discuss the effects of turbulence on wavefront propagation.

1.7.1 Background

1.7.1.1 Atmospheric Turbulence-Induced Wavefront Distortion.
The index of refraction of the atmosphere varies with temperature, pressure, and humidity. Furthermore, the atmosphere is in turbulent motion—the velocity of the air's motion measured at spatially separated points varies randomly. The atmosphere is an extended randomly inhomogeneous medium. An optical wavefront propagating through the atmosphere becomes distorted, limiting the angular resolution of large aperture telescopes to ~ 1 arcsec.

Perhaps the most important and influential works on the effects of wave propagation in turbulence are those of Tatarskii.[433,434] He found that the *structure function* of the index of refraction, $\mathcal{D}_n$, is the key quantity from which one can determine the effects on imaging system performance. The structure function is defined as

$$\mathcal{D}_n(\mathbf{r}_1, \mathbf{r}_2) \equiv \langle [n(\mathbf{r}_1) - n(\mathbf{r}_2)]^2 \rangle = C_n^2 r^{2/3} \ , \tag{1.245}$$

where $\mathbf{r}_1$ and $\mathbf{r}_2$ are two points in space separated a distance $r = |\mathbf{r}_1 - \mathbf{r}_2|$, $n(\mathbf{r})$ is the spatially varying index of refraction of the atmosphere, $\langle \cdot \rangle$ denotes an ensemble average, and C_n^2 is a constant of proportionality called the *index of refraction structure constant*—a measure of the strength of the fluctuations. The $r^{2/3}$ dependence of the index of refraction structure function follows from assuming the Kolmogorov power spectral density of fluctuations, $\Phi(\kappa)$,[435]

$$\Phi(\kappa) = 0.033 C_n^2 \kappa^{-11/3} \ , \tag{1.246}$$

where κ is the spatial wave number. The wave number κ is inversely related to the size of the turbulent eddies in the atmosphere.

The effects of this distribution of varying index of refraction on electromagnetic wave propagation can be determined from solutions of Maxwell's equations. Tatarskii introduced the Rytov transformation, which substitutes the natural logarithm of the field into the wave equation. The resulting solution is a field of the form

$$\mathbf{U}(x,y) = \sqrt{I_0} \exp[\chi(x,y) + i\phi(x,y)] \ , \tag{1.247}$$

where I_0 is the intensity of the plane wave incident on the turbulence, $\chi(x,y)$ is the logarithm of the amplitude fluctuations, and $\phi(x,y)$ is the phase fluctuation. Both χ and ϕ are Gaussian random variables given by the Rytov

solution (discussion of which is beyond the scope of this section) and are dependent on the randomly varying index of refraction of the medium. Experimental data support the theoretical solution that the logarithm of the amplitude obeys Gaussian statistics.[436]

Using the fact that ϕ is a Gaussian random variable, the *phase structure function* of a propagating electromagnetic wave, $\mathcal{D}_\phi$, is proportional to the 5/3 power of spatial separation[437] and is given by[e]

$$\mathcal{D}_\phi(\mathbf{r}) = \langle [\phi(\mathbf{r}') - \phi(\mathbf{r}' - \mathbf{r})]^2 \rangle$$

$$= \left\{ 2.91 \left(\frac{2\pi}{\lambda}\right)^2 \int C_n^2(\xi)\left(1 - \frac{\xi}{z_g}\right) d\xi \right\} r^{5/3} , \qquad (1.248)$$

where λ is the wavelength of propagation, $C_n^2(\xi)$ is the position-dependent structure constant, z_g is the distance to the source, and the integration is over the optical path from the source to the observation point. For plane waves the term ξ/z_g vanishes.

Fried[438] developed a convenient relationship between the phase structure function and a particular measure of optical resolution—the volume under the two-dimensional optical transfer function (which is also directly related to the Strehl ratio). As the aperture size increases, the resolution defined in this way asymptotically approaches a limiting value set by the strength of the turbulence along the path. Fried defined a quantity r_0 such that the limiting resolution obtained in the presence of atmospheric turbulence is the same as that obtained by a diffraction-limited lens of diameter r_0 in the absence of the atmosphere. The numerical value of r_0 is between 10 and 30 cm at 0.5 μm for vertical propagation at a site of good seeing. Its value can be determined from its definition[439]:

$$r_0 \cong \left[\frac{0.06\lambda^2}{\int C_n^2(\xi)\, d\xi} \right]^{3/5} , \qquad (1.249)$$

where C_n^2 is the index of refraction structure constant, and the integral is performed along the propagation path. With this definition, the phase structure function becomes

$$\mathcal{D}_\phi(r) = 6.88 \left(\frac{r}{r_0}\right)^{5/3} . \qquad (1.250)$$

Another very useful concept was introduced by Noll.[440] He extended Fried's work by developing a Zernike polynomial representation of wavefront distortion induced by atmospheric turbulence. The Zernike polynomials are orthog-

---

[e]In the more general case where amplitude effects are significant, the *wave* structure function, $\mathcal{D}(r) = \mathcal{D}_\chi(r) + \mathcal{D}_\phi(r)$ is used but has exactly the same value as given by Eq. (1.248).

onal over the unit circle and are often used to describe the classical aberrations of an optical system—defocus, astigmatism, coma, and spherical aberration. Noll represents the phase of the distorted wavefront as

$$\phi(\rho,\theta) = \sum_j a_j Z_j(\rho,\theta) , \tag{1.251}$$

where the Z_j are a modified set of the orthogonal Zernike polynomials defined over points in the aperture at radius ρ and azimuthal angle θ, and the a_j are coefficients having proper mean square values that accurately represent the Kolmogorov distribution of turbulence. The a_j are Gaussian random variables with zero mean whose second-order statistics are described by a covariance matrix. The covariance matrix can be computed from the Fourier representation of the Wiener spectrum of the phase fluctuations. The wavefront distortion is often given as a mean square value (in radians squared). The piston-removed wavefront distortion averaged over an aperture of diameter D for Kolmogorov turbulence (including the two components of tilt) expressed as a mean square value is

$$\langle\phi^2\rangle = 1.0299\left(\frac{D}{r_0}\right)^{5/3} . \tag{1.252}$$

Figure 1.59 is a plot of the mean square wavefront distortion in units of $(D/r_0)^{5/3}$ as a function of the number of Zernike terms removed from the wave-
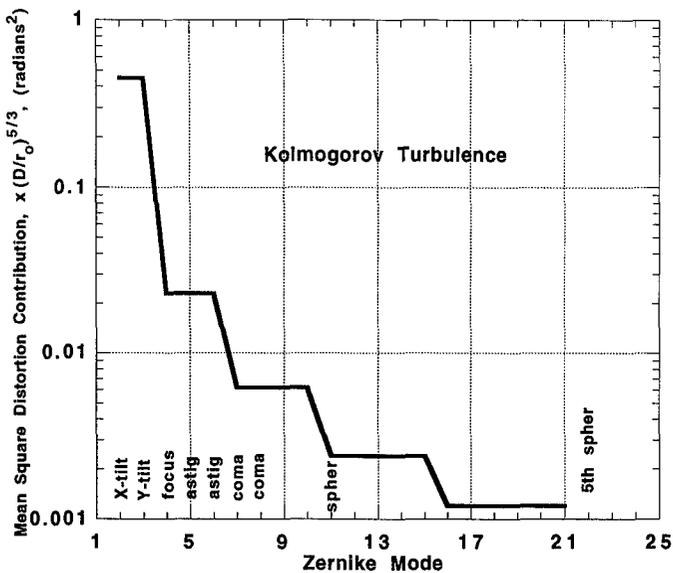
Fig. 1.59 Normalized Zernike mode contributions to the total atmospheric turbulence wavefront distortion. The sum of all modes yields a total mean square wavefront distortion of 1.03 $(D/r_0)^{5/3}$ rad^2.

front. Notice that tilt is a major contributor to the total wavefront distortion. These representations are useful in estimating the performance of an adaptive optics system if one can estimate how many Zernike terms the system will correct.

1.7.1.2 Anisoplanatism. A specific realization of atmospheric turbulence-induced wavefront distortion is highly dependent on a particular path through the turbulence—the distortions rapidly become uncorrelated outside a small region called the *isoplanatic angle*. Fried[441] defines the isoplanatic angle θ_0 as

$$\theta_0 = \left[2.91 \left(\frac{2\pi}{\lambda} \right)^2 \int d\xi \, C_n^2(\xi) \xi^{5/3} \right]^{-3/5} , \qquad (1.253)$$

where the integral is taken along the path of propagation. One interpretation of the isoplanatic angle is that it is the angle between two propagation directions for which the difference in mean square wavefront distortion is 1 rad^2. Fried shows that for large D/r_0, if an imaging system viewing along one path is corrected by an ideal adaptive optics system, the optical transfer function of the system viewing along a different path separated by an angle θ is reduced by a factor $\exp[-(\theta/\theta_0)^{5/3}]$. The numerical value of the isoplanatic angle for zenith viewing at a wavelength of 0.5 μm is of the order of a few arcseconds for most sites. The isoplanatic angle is strongly influenced by high-altitude turbulence (note the $\xi^{5/3}$ weighting in the above definition, where ξ is the altitude). The small value of the isoplanatic angle can have very significant consequences for adaptive optics by limiting the sky coverage possible when natural stars are used as reference beacons and by limiting the corrected field of view to only a few arcseconds.

1.7.1.3 Temporal Variations. The spatial variations of the wavefront distortion we have discussed thus far are ensemble averages over the turbulence statistics. Temporal variations in the wavefront distortion at a specific point in the aperture can be modeled to first order as due to spatial variations moving over the aperture at some speed V. The motion of this "frozen turbulence" past the aperture is caused by the wind or by an "artificial wind" if the aperture is being slewed at some angular rate. Greenwood[442] defined the variance of the uncorrected phase error due to finite servo frequency in terms of a characteristic frequency of the servo filter response function. The frequency at which the variance of the residual error is 1 rad^2 is usually referred to as the *Greenwood frequency* f_0, and for an RC-type servo response is defined by

$$f_0 = \left[0.102 \left(\frac{2\pi}{\lambda} \right)^2 \int_0^L C_n^2(\xi) \, V^{5/3}(\xi) \, d\xi \right]^{3/5} . \qquad (1.254)$$

The term $V(\xi)$ is the transverse wind velocity at positions ξ along the path and the integral is performed over the total path length L of propagation. The Greenwood frequency varies with the turbulence and wind profile but is typically of the order of a few hertz to a few tens of hertz for most astronomical sites. Tyler[443] shows that for large D/r_0, the mean square residual phase error

in an adaptive optics control system having a -3-dB closed-loop bandwidth of f_s Hz is

$$\sigma^2_{\text{servo}} = (f_0/f_s)^{5/3} \; . \tag{1.255}$$

In a practical system this means that the closed-loop servo bandwidth in an adaptive optics system should be several times the Greenwood frequency in order to "stay up with the turbulence" and keep the residual wavefront error to less than 1 rad^2. We now discuss adaptive optics and the need for artificial beacons.

1.7.1.4   Adaptive Optics and the Laser Guide Star Concept.

Adaptive optics is a technique for correcting wavefront distortions in real-time. (A detailed discussion of adaptive optics appears elsewhere in this Handbook.) The concept was introduced by the astronomer Horace W. Babcock[444] in 1953. The idea is to measure instantaneously the wavefront distortion and impress its conjugate on a mirror whose surface can be deformed to the desired shape. As the atmospherically distorted wavefront reflects off the surface of the deformed mirror, the aberrations are canceled and the blurred image is corrected.

The key processes in adaptive optics are sensing the wavefront distortion and adjusting a deformable mirror to correct the distortion. In a conventional configuration, a sensor measures the distortion imposed on an optical wavefront that has propagated through the turbulence from a light source, or beacon, located at the object being imaged. In many applications, the self-emission of the object or light reflected by the object can serve as the beacon. However, in those cases where the object is not bright enough for the wavefront sensor to make a measurement in a time short compared to the atmospheric time constant (the time required for a significant change in the wavefront's distortion pattern), an artificial beacon is useful. The artificial beacon is also useful in situations where there is no natural beacon at all. An example is when transmitting a corrected laser beam to a deep-space probe for high-bandwidth communication.

Artificial beacons can be generated by the light backscattered from a laser beam focused in the atmosphere. Two scattering mechanisms have been demonstrated: Rayleigh scattering from nitrogen and oxygen molecules (practical at backscatter ranges of 10 to 20 km), and resonant scattering from free sodium atoms in the mesosphere (85- to 100-km altitudes).

The artificial beacon has the advantage that it can be made bright enough for high-bandwidth sensing and placed in an arbitrary direction in the sky. This is important because the atmosphere is anisoplanatic—induced wavefront distortions are insufficiently correlated outside the isoplanatic angle.[441] In astronomy applications, this means a bright natural star cannot be used as an adaptive optics beacon unless it and the object of interest lie within the same isoplanatic patch. The isoplanatic angle is of the order of a few arcseconds at visible wavelengths and (assuming Kolmogorov turbulence) increases only as the 6/5 power of the wavelength.[441] Natural stars generally considered suitable for higher order adaptive optics beacons are separated 200 to 500 arcsec or more.[445] As a result, only a small fraction of the sky is accessible to adaptive

optics using natural guide stars. The artificial guide star can be used to fill in those regions where a natural guide star does not exist.

1.7.1.5 Historical Perspective.

The use of laser beacons for adaptive optics has been under development[446] by the U.S. Department of Defense (DoD) since 1982. In May 1991 the U.S. Air Force obtained permission to relax controls on certain information obtained under DoD-sponsored research. Subsequently, some of the analytical and experimental results of the DoD work have been published and presented.[f] The principal events in the early development of laser guide star adaptive optics are summarized here.

The concept of using focused lasers in the atmosphere to generate artificial beacons was formally proposed to the Defense Advanced Research Projects Agency by Julius Feinleib of Adaptive Optics Associates in the fall of 1981. The idea was expanded in scope by Prof. Will Happer of Princeton in the summer of 1982 by his suggestion of utilizing resonant scattering of mesospheric sodium.

In 1983, Fugate et al.[447,448] first demonstrated that an artificial beacon could be used for sensing atmospheric turbulence-induced wavefront distortion. Their experiment compared wavefront distortion measurements from a natural star with that from a laser guide star generated by Rayleigh scattering. Results of their experiment verified an important theoretical prediction by Fried and Belsher[449,450] that the wavefront measurement error incurred by using a beacon at a finite range (an effect called *focus anisoplanatism* discussed in detail later) instead of infinite range is proportional to the 5/3 power of the aperture diameter.

During 1984 and early 1985, Humphreys et al.[451] performed a similar experiment using a dye laser to generate a beacon in the mesospheric sodium layer. They used two widely separated subapertures and were first to demonstrate wavefront slope measurements using a sodium laser guide star. Their results confirmed quantitative predictions of how the focus anisoplanatism error is reduced by using a higher altitude beacon.

Foy and Labeyrie[452] published the laser guide star concept in 1985. In 1987, Thompson and Gardner[453] created a laser guide star in the mesospheric sodium layer at Mauna Kea using a pulsed dye laser. They imaged the guide star through a 2.2-m telescope and verified theoretical predictions of backscatter signal level.

Primmerman et al.[454] were first to demonstrate real-time correction of atmospheric turbulence-induced wavefront distortion using a Rayleigh backscatter beacon and deformable mirror. Their experiments (first results in the summer of 1988) were done on a 60-cm aperture at 0.5 μm and performed a full correction for each pulse of the dye laser beacon (operating at 2.5 pulses/s). Even though these experiments provided only a low duty cycle correction ($\sim$1%), they demonstrated high Strehl ratios ($\sim$0.4) and verified analytical predictions of expected performance. A follow-on experiment[455] demonstrated

---

[f]While this manuscript was in preparation, a topical issue on atmospheric compensation and adaptive optics was announced by the *Journal of the Optical Society of America A*. In addition to the references cited here, papers were in preparation reporting the results of the DoD-sponsored research to be submitted for that topical issue. Tentative publication date: October and November 1993.

a technique to combine data from two beacons—a technique that reduces the focus anisoplanatism error inherent in a single beacon.

In February 1989 Fugate et al.[456, 457] demonstrated closed-loop, high-bandwidth, 100% duty cycle correction using a laser beacon and deformable mirror. Their experiments were conducted on a 1.5-m aperture with image quality evaluated at 0.88 μm. They used a high-repetition-rate (5000 pulses/s) copper vapor laser to create a Rayleigh backscatter beacon at a 10-km range. They achieved a Strehl ratio of 0.20 and FWHM point spread functions of 0.18 arcsec (λ/D = 0.12 arcsec) by operating the system at the 65-Hz closed-loop bandwidth. In 1992 this group began evaluation of a second-generation high-bandwidth system, employing a new deformable mirror, wavefront sensor, tracker, optics, and upgraded copper vapor laser. Initial results showed considerably improved performance over the first system with Strehl ratios of 0.64 for natural star operation and 0.48 for laser guide star operation. FWHM point spread functions were 0.13 arcsec in both cases and the modulation transfer function showed nonzero response to the cutoff frequency. Closed-loop bandwidths of 130 Hz were demonstrated. This laser guide star adaptive optics system has been used to image extended astronomical objects.[457]

Sandler et al.[458,459] have demonstrated single pulse correction using excimer lasers on 0.5- and 1.0-m apertures. They have also demonstrated a technique to use ~20 laser beacons positioned over the aperture as a means of reducing focus anisoplanatism resulting from the use of one beacon.

Thompson[460] has reported excellent results in generating laser guide stars using an excimer laser operating at 0.351 μm. Short-wavelength excimer lasers are an excellent choice since Rayleigh scattering strength is proportional to λ^{-4} and since operation in the ultraviolet provides excellent isolation for scientific cameras and instruments in the visible and infrared bands.

Major projects in the United States to develop laser guide star adaptive optics for astronomy are under way at the University of Chicago[461] and Lawrence Livermore Laboratory.[462] European astronomers have been actively engaged in laser guide star adaptive optics since its introduction by Foy and Labeyrie. An experimental program called ATLAS[463] is being conducted at CERGA. Beckers[464] and Merkle[465] are responsible for plans to incorporate sodium laser guide star adaptive optics in the European Southern Observatory's Very Large Telescope—four 8-m telescopes being built for installation in Chile. Reliable, high-performance laser guide star adaptive optics systems for astronomy may be in routine operation in 5 to 10 years.

1.7.2 Laser Guide Star Theory

The major sources of error in a laser guide star adaptive optics system are (1) focus anisoplanatism, (2) lack of full aperture tilt information, (3) time delay between sensing the wavefront distortion and making a correction with the deformable mirror (latency and servo lag), (4) wavefront sensing and reconstruction errors, and (5) a spatial fitting error (due to the finite number of actuators). The last three sources of error are present in a conventional adaptive optics system. The normal trade-offs associated with subaperture size, spacing of deformable mirror actuators, wavefront sensing sample rate, and servo averaging and filtering are also required in a laser guide star system

and are discussed in more detail in the adapative optics chapter of this Handbook. Before analytical results are presented, we qualitatively summarize the effects of the first two of these error sources.

The artificial beacon has two significant limitations when applied to adaptive optics. The first is an effect called *focus anisoplanatism* caused by the finite altitude of the beacon, which limits the size of the aperture that can be corrected. The second is that the laser beacon provides no information on turbulence-induced full-aperture tilt.

1.7.2.1 Lack of Tilt Information.

Full-aperture tilt is the orientation of the best fit plane to the aberrated wavefront arriving at the telescope. If the wavefront distortion due to turbulence is decomposed into surfaces represented by the Zernike polynomials (as discussed previously),[440] the orthogonal components of full-aperture tilt are described by the X and Y Zernike tilt terms. When the laser beam propagates up through the atmosphere from the ground, its direction of travel as determined by the Zernike tilt varies randomly and the focused spot ends up laterally shifted from the optical axis of the telescope by an unknown amount. Furthermore, if the laser beam is sharing the aperture of the telescope, the backscattered light returns to the telescope down the same turbulent path, arriving at the aperture on axis. That is, because of reciprocity, the image of the laser guide star obtained using the full aperture always *appears* to be centered on the optical axis, providing no information about full-aperture tilt induced by the atmosphere.

The lack of tilt information is a serious limitation for applications requiring long exposures or very precise pointing (to within a fraction of λ/D) of laser beams at undetectable targets (such as deep-space probes). For astronomy, a potential solution is to use a natural guide star close to the object being imaged for full-aperture tilt correction and the laser guide star for higher order correction. Using a natural guide star for tilt-only correction provides more sky coverage than for higher order correction since (1) the full aperture of the telescope can be used to sense tilt (compared to only subapertures for higher order) and (2) the higher order wavefront distortion is being corrected by the adaptive optics. These two effects combine to produce a near diffraction-limited, more intense image of the guide star and hence allow fainter guide stars to be used.

1.7.2.2 Focus Anisoplanatism.

The other limitation in the use of artificial beacons is called *focus anisoplanatism*.[g] Focus anisoplanatism is a wavefront sensing error caused by the fact that the laser guide star is at a finite range from the telescope and a wavefront from this artificial star propagates through a slightly different atmospheric path than a wavefront from a natural guide star at infinity. This is illustrated schematically in Fig. 1.60. This effect introduces an error in estimating the actual wavefront distortion. The focus anisoplanatism error increases as the diameter of the telescope being corrected increases. The magnitude of the distortion is significant for telescope aperture sizes greater than 1 m and artificial beacon altitudes of less than 20 km. Quantitative details are discussed next.

---

[g]T. R. O'Meara, of Hughes Research Laboratory, Malibu, introduced the term *focus anisoplanatism*. The term *focal anisoplanatism* also appears in the literature.
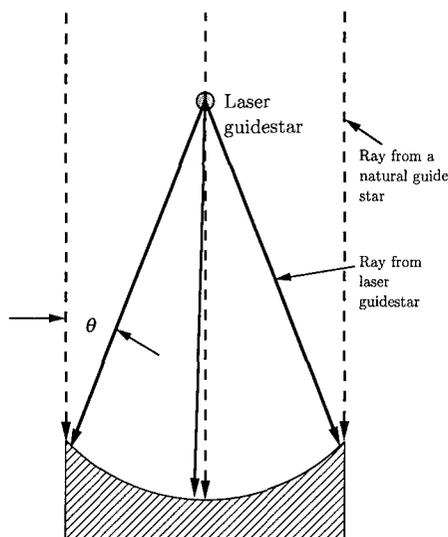
Fig. 1.60   The laser guide star concept. Schematic diagram showing rays from a laser guide star positioned over the center of the aperture to be compensated. The laser guide star can be created by Rayleigh scattering in the stratosphere (10- to 20-km altitude) or by resonant fluorescence scattering in atomic sodium in the mesosphere. The dotted lines show rays from a natural guide star and the solid lines show rays from the laser guide star. Note that at the edge of the aperture, the rays are traveling through different atmospheric paths separated by the angle θ. When this angle is comparable to the isoplanatic angle, a significant error is made in estimating the phase distortion using the laser guide star. This effect is known as focus anisoplanatism. Larger diameter apertures can be corrected by moving the laser beacon to higher altitudes or by using multiple beacons.

1.7.2.3   Analytical Estimate of Focus Anisoplanatism.

Useful figures of merit for quantitative evaluation of an imaging system's performance include the Strehl ratio, the point spread function, and the modulation transfer function.

First consider the effects of focus anisoplanatism on the Strehl ratio. A good starting point is the Maréchal approximation to the Strehl ratio (the Strehl ratio is discussed elsewhere in this Handbook):

$$SR = \exp(-\sigma_\phi^2) \ , \tag{1.256}$$

where SR is the Strehl ratio and σ_ϕ^2 is the phase variance (or mean square wavefront error) in radians squared. This approximation is valid only when the phase variance is less than a few tenths of a radian squared.

We can use this formula to estimate the Strehl ratio due to focus anisoplanatism if we have an estimate of the mean square error in the wavefront measurement due to the finite range of the artificial beacon. Consider the geometry of the artificial laser beacon adaptive optics system shown in Fig. 1.60. The problem is to evaluate the mean square difference between a wave-

front arriving at the telescope aperture from infinity, ϕ_∞, and one arriving from the laser beacon, ϕ_{LB}, averaged over the aperture. As discussed in Sec. 1.7.2.1, the laser guide star does not provide any information on full-aperture tilt. We must use a natural guide star to provide that information and a typical arrangement is to use a separate sensor and two-axis mirror for fast tracking (guiding). As a consequence, we do not want to include tilt in the calculation of the mean square wavefront difference. The mean square residual wavefront distortion due to focus anisoplanatism, $\sigma_{\phi FA}^2$, is defined by

$$\sigma_{\phi FA}^2 = \left\langle \frac{4}{\pi D^2} \int dr W(\mathbf{r}/D)\{[\phi_\infty(\mathbf{r}) - \phi_{LB}(\mathbf{r})] - \overline{\phi}_P - \phi_T(\mathbf{r})\}^2 \right\rangle . \quad (1.257)$$

The $\langle \cdot \rangle$ in this formula represents an ensemble average over the turbulence, $W(\mathbf{r}/D)$ is the aperture function given by

$$W(\mathbf{r}/D) = \begin{cases} 1 & \text{if } |\mathbf{r}/D| \leq 1 \\ 0 & \text{otherwise} \end{cases} ,$$

D is the diameter of the telescope, $\mathbf{r}$ is a point in the aperture, $\overline{\phi}_P$ is the piston component of the phase difference $\phi_\infty - \phi_{LB}$, and $\phi_T(\mathbf{r})$ is the wavefront tilt.

Fried and Belsher[450] evaluated the expression on the right side of Eq. (1.257) in a comprehensive analysis to show that $\sigma_{\phi FA}^2$ increases as the 5/3 power of the telescope diameter. They introduced a constant of proportionality d_0 such that

$$\sigma_{\phi FA}^2 = (D/d_0)^{5/3} , \quad (1.258)$$

where d_0 is the diameter of a physical aperture for which the mean square phase error due to focus anisoplanatism is 1 rad^2. For good performance (a Strehl of 0.73) we would like d_0 to be at least twice the diameter of the telescope we are trying to compensate. The value of d_0 depends on the imaging wavelength, zenith angle, the altitude of the artificial beacon, and the value of the index of refraction structure constant C_n^2 as a function of altitude. The largest contribution to the focus anisoplanatism error occurs at the edge of the telescope aperture, where the rays from the artificial beacon deviate most from a natural beacon (see Fig. 1.60).

Tyler[466] evaluated the plane wave covariance and spherical wave covariance and plane wave–spherical wave cross-covariance resulting from expansion of Eq. (1.257). The results permit evaluation of d_0 in terms of a single integration over altitude:

$$d_0 = \lambda^{6/5} \cos^{3/5}(\psi) \left[\int dh C_n^2(h) F_n(h/H) \right]^{-3/5} , \quad (1.259)$$

where

$$F_n(h/H) =$$

$$
\begin{cases}
16.71371210\left([1.032421640 - n0.8977579487]\left[1 + \left(1 - \dfrac{h}{H}\right)^{5/3}\right]\right. \\[2ex]
\quad - 2.168285442\left\{\dfrac{6}{11}{}_2F_1\left[-\dfrac{11}{6}, -\dfrac{5}{6}; 2; \left(1 - \dfrac{h}{H}\right)^2\right] - \dfrac{6}{11}\left(\dfrac{h}{H}\right)^{5/3}\right. \\[2ex]
\quad \left.\left. - n\dfrac{10}{11}\left(1 - \dfrac{h}{H}\right){}_2F_1\left[-\dfrac{11}{6}, \dfrac{1}{6}; 3; \left(1 - \dfrac{h}{H}\right)^2\right]\right\}\right) & \text{for } h < H \\[4ex]
16.71371210(1.032421640 - n0.8977579487) & \text{for } h > H
\end{cases}
$$

where $n = 0$ when only piston is removed and $n = 1$ when piston and tilt are removed; h is the altitude; H is the altitude of the laser backscatter beacon; ${}_2F_1(a,b;c;z)$ is the hypergeometric function; ψ is the zenith angle; and λ is the wavelength. Although these expressions look formidable, $F_n(h/H)$ is dependent only on the quantity h/H and is actually quite simple in form and can easily be evaluated on a desk-top computer or programmable scientific calculator.

The value of d_0 has been computed[466] for a variety of vertical turbulence profiles $C_n^2(h)$. Results for three turbulence profile models are shown in Fig. 1.61 for zenith viewing at a wavelength of 0.5 µm. Note that d_0 is proportional to $\lambda^{6/5}$ and the 3/5 power of the cosine of the zenith angle, so the results of Fig. 1.61 can easily be scaled to other wavelengths and zenith angles for these particular turbulence profiles. Using these values of d_0, we can easily calculate the Strehl ratio using Eqs. (1.258) and (1.259) for a given laser guide star altitude and aperture size.

The effects of focus anisoplanatism have also been determined by researchers at MIT Lincoln Laboratory using analytical techniques developed in recent years by Sasiela and Parenti.[467,468] These techniques use Mellin transform theory to solve complex turbulence calculations in terms of expansions involving moments of the turbulence profile having the form

$$\mu_n = \int dh\, C_n^2(h) h^n , \tag{1.260}$$

where μ_n is called the n'th turbulence moment. Using these techniques, Sasiela derives two expressions for the focus anisoplanatism error: one for contributions above the beacon at altitude H, σ_{upper}^2, and one for contributions below the beacon, σ_{lower}^2. These results are

$$\sigma_{upper}^2 \simeq 0.057 D^{5/3}\left(\frac{2\pi}{\lambda}\right)^2 \sec(\psi)\mu_0^{\uparrow}(H) \tag{1.261}$$

and

$$\sigma_{lower}^2 \simeq D^{5/3}\left(\frac{2\pi}{\lambda}\right)^2 \sec(\psi)\left[0.50\frac{\mu_{5/3}^{\downarrow}(H)}{H^{5/3}} - 0.45\frac{\mu_2^{\downarrow}(H)}{H^2} + ...\right] , \tag{1.262}$$
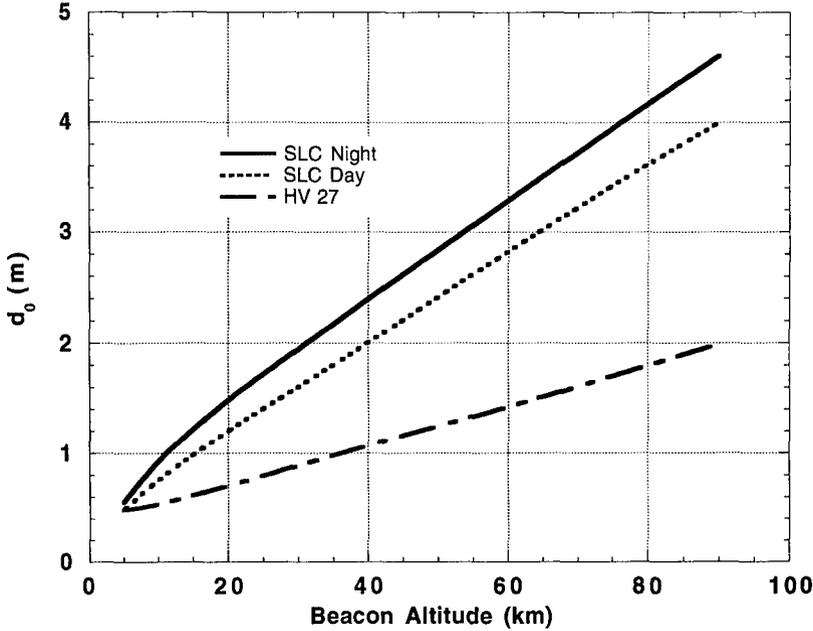
Fig. 1.61   Values of d_0 for three C_n^2 turbulence profiles. This graph shows the value of d_0 as a function of beacon altitude for the *(top to bottom)* SLC-Night, SLC-Day, and Hufnagel-Valley-27 turbulence profiles. If the physical aperture of the telescope is equal to d_0, the rms wavefront error due to focus anisoplanatism alone will be 1 rad, and the Strehl ratio will be approximately $1/e$.

where

$$\mu_n^{\downarrow}(H) = \int_0^H dh C_n^2(h) h^n \tag{1.263}$$

and

$$\mu_n^{\uparrow}(H) = \int_H^{\infty} dh C_n^2(h) h^n \ . \tag{1.264}$$

These results are in complete agreement with those presented in Eq. (1.259) for the tilt included case.

Welsh and Gardner[469] also present a detailed analysis of the performance of laser guide stars including the characteristics of the wavefront sensor and deformable mirror. They define the diameter of the largest telescope correctable by a single guide star as $D_{\text{IP}} = 2 z_g \theta_{\text{IP}}$, where z_g is the height of the guide star and θ_{IP} is the isoplanatic angle. They estimate that three sodium layer guide stars are needed for a 2-m-diameter telescope assuming the Hufnagel C_n^2 profile is scaled to give $r_0 = 20$ cm.

1.7.2.4 Multiple Beacons. We can think of focus anisoplanatism as arising from two effects. The first is, for a single laser beacon at altitude H, we sense no turbulence above H, and below H our ability to sense the strength of the turbulence is weighted by the factor $1 - h/H$. The second is that a single beacon over a large aperture senses turbulence along the incorrect path as the sampling point in the aperture moves out from under the laser beacon. If the laser guide star is positioned on the axis of the telescope, the largest discrepancy occurs at the edge of the aperture.

Two obvious ways to improve this situation are to (1) move the laser beacon to a higher altitude and (2) use multiple beacons. We can limit the region for sensing to that just under each beacon and then combine the information from each region to construct a phase measurement for the entire aperture. Unfortunately, the process of combining the data from each region introduces a new source of error. This "stitching error" arises from not knowing the relative positions of the laser beacons. That is, each beacon travels through a slightly different atmospheric path, which induces a slightly different overall tilt, causing random variations in the resulting beacon pattern. Welsh and Gardner[469] treat the multiple beacon case but ignore the problem of beacon position uncertainty. This makes their performance predictions optimistic for large numbers of beacons.

Sasiela[470] gives an expression for the stitching error in terms of the partial moments of turbulence, Eq. (1.260), as

$$
\sigma^2_{\text{stitching}} \simeq D^{5/3} \left(\frac{2\pi}{\lambda} \right)^2 \sec(\psi) \left[0.040 \frac{\mu_2^{\frac{1}{2}}(H)}{H^2} + ... \right] . \tag{1.265}
$$

This error is only weakly dependent on the number of beacons deployed. For a 4-m aperture operating at a 45-deg zenith angle at 0.55-μm wavelength, the stitching error variance is less than 1 rad^2 at all backscatter altitudes for a turbulence profile representative of a good seeing site. The beacon spacing is determined by that portion of the aperture (sometimes called the *section size*) for which the focus anisoplanatism error is of the order of 1 rad^2.

Fried[471] has studied the case of three beacons and analyzed the combined effects of beacon position uncertainty and focus anisoplanatism due to a single beacon. For a given point in the aperture, he combines data from the three beacons using linear interpolation. He finds that optimum performance is achieved with the beacons spaced at 120-deg intervals on a circle having a radius equal to 0.65 of the aperture's radius. For three beacons, focus anisoplanatism limits the maximum value of d_0 to $\sqrt{3}$ times the value of d_0 for a single beacon. He concludes that it is reasonable to expect that the value of d_0 could be increased by a factor $\sqrt{N}$ over the value for a single beacon by using N beacons and a suitable data interpolation scheme. However, these rules apply only for small values of N and Fried's analysis shows that for beacon altitudes below 20 km, little advantage is gained from the use of more than three beacons. Fried also concludes that a maximum of about nine beacons will produce as large a value of d_0 as allowed by focus anisoplanatism.[h]

---

[h]Additional information may appear in the previously mentioned special issue of the *Journal of the Optical Society of America.*

1.7.3   Laser Guide Star Generation

Laser guide stars are generated either by Rayleigh scattering from air molecules in the stratosphere (10 to 40 km) or by resonant fluorescence scattering from atomic sodium in the mesosphere (80 to 100 km). In most cases the laser guide star will be generated by a pulsed laser, allowing the wavefront sensor to range-gate out the strong local backscatter. It is feasible, however, in certain cases (e.g., a sodium beacon for applications in the near IR) to use a continuous laser. We limit the discussion here to pulsed lasers.

For a conventional closed-loop servo, the laser pulse rate should be high enough to sample at about 20 times the highest expected Greenwood frequency, in anticipation of a closed-loop bandwidth operation at approximately twice the Greenwood frequency. It is also possible to operate the system in a "go to" mode in which a calculated shape is commanded onto the deformable mirror with no feedback as to the mirror's actual shape. This mode of operation requires careful calibration of both the wavefront sensor and the deformable mirror. The "go to" mode relaxes the pulse rate requirement for the beacon laser but does not relax processing speed requirements (in order to minimize errors due to latency). It is also possible to operate the deformable mirror in a fast local loop using an external phase measuring sensor to control the mirror's figure. The laser pulse length should be less than one-tenth the range gate length for Rayleigh scattering and not shorter than several tens of microseconds for sodium layer scattering to avoid saturation.

1.7.3.1   Rayleigh Laser Guide Stars.
The number of photodetected electrons (pde) per subaperture for Rayleigh scattering is given by a lidar-type expression:

$$N_{\mathrm{pde}} = \eta_{\mathrm{QE}} T_t T_r T_{\mathrm{atm}}^2 \frac{A_{\mathrm{sub}} \beta_{\mathrm{BS}} \Delta l}{R^2} \frac{E_p \lambda}{hc} \ , \tag{1.266}$$

where

E_p = energy per pulse
T_t = laser transmitter optical transmission
T_r = optical transmission of the wavefront sensor
T_{atm} = one-way transmission of the atmosphere
A_{sub} = area of a subaperture in the wavefront sensor
β_{BS} = fraction of incident laser photons scattered in the backward
        propagation direction per meter of scattering volume
        $(\mathrm{sr}^{-1}\,\mathrm{m}^{-1})$
Δl = length of the scattering volume (the range gate)
η_{QE} = quantum efficiency of the wavefront sensor detection process
λ = wavelength
R = range to the midpoint of the scattering volume
h = Planck's constant
c = speed of light.

The volume backscattering function β_{BS} is given approximately by[472]

$$\beta_{\mathrm{BS}}(h) = \frac{\pi^2 (n_0^2 - 1)^2}{N_0 \lambda^4} \frac{T_0}{P_0} \frac{P(h)}{T(h)} \ , \tag{1.267}$$

where n_0 is the index of refraction of air at sea level, N_0 is the number density of air molecules at sea level, and P_0, T_0, $P(h)$, and $T(h)$ are the atmospheric pressure and temperature at sea level and altitude h, respectively. Using the sea-level parameters from the U.S. Standard Atmosphere,[473] we can reduce Eq. (1.267) to[i]

$$\beta_{BS} = 2.8 \times 10^{-32} \frac{P(h)}{T(h)} \, \text{sr}^{-1} \, \text{m}^{-1} \ . \tag{1.268}$$

At an altitude of 10 km, β_{BS} is approximately $5.1 \times 10^{-7} \, \text{sr}^{-1} \, \text{m}^{-1}$ so a 1-km-long scattering length returns only 0.05% of the incident photons per steradian.

There are at least two criteria for limiting the length of the scattering volume. Gardner, Welsh, and Thompson[474] require that the $1/e^2$ intensity points be no larger than the angular size of a natural guide star and derive the following expression for the maximum length of the range gate

$$\Delta l = 4.88 \frac{\lambda H^2}{D_t r_0} \ , \tag{1.269}$$

where D_t is the diameter of the laser transmitter optics. This equation does not take into account any increase in the size of the focused spot due to imperfect laser beam quality.

The second criterion requires that subapertures at the edge of the telescope aperture not be able to resolve the projected length of the guide star. Increased laser guide star image size in the wavefront sensor increases the error in wavefront slope measurement by each subaperture. A simple geometrical analysis leads to the result

$$\Delta l = 2 \frac{\lambda H^2}{D r_0} \ , \tag{1.270}$$

where D is the aperture diameter of the wavefront sensing (and imaging) telescope. In both cases note that the maximum range-gated length of the laser guide star is proportional to the square of the average beacon height.

1.7.3.2   Mesospheric Sodium Guide Stars.

It is desirable to generate the laser guide star as far from the telescope as possible in order to minimize the effects of focus anisoplanatism. It is feasible to generate laser guide stars by resonance fluorescence scattering of atomic sodium in the mesosphere 85 to 100 km above the earth's surface. The sodium layer is very tenuous, having a density of only a few thousand atoms per cubic centimeter (all the sodium in the earth's atmosphere condensed into a solid would fit into a telephone booth).

The D_2 transition at 0.589 μm is the most useful for generation of laser guide stars. The ground state of this transition has two energy levels split 1.77 GHz by the interaction of the electron and nuclear magnetic dipole moments. Furthermore, the natural absorption linewidth of 10 MHz is inhomogeneously Doppler broadened to 1.07 GHz, making the total absorption linewidth ~3

---

[i]Gardner, Welsh, and Thompson[474] list the coefficient in this equation as 3.6×10^{-32}.

GHz. Other subtle effects such as optical pumping[475–477] can significantly affect the backscattered signal level but are not discussed here.

The analysis of the signal returned from resonant scattering in mesospheric sodium is a complex subject. In general, the excitation laser emits a train of pulses. If the pulses are short compared to the spontaneous emission time (16 ns), it is necessary to treat the subject using quantum mechanics and the Bloch equations[478,479] (which are beyond the scope of this section). For pulses longer than 16 ns, a rate equation approach is adequate. Rate equations describe the time rate of change of the population of the excited state in terms of transitions stimulated by incident laser energy and by transitions caused by thermal relaxation. Solutions of the rate equation[480] are often expressed in terms of a saturation intensity I_{sat} such that the fraction f of the atoms in the excited state is

$$f = \frac{1}{2}\left(\frac{I/I_{\text{sat}}}{1 + I/I_{\text{sat}}}\right) , \tag{1.271}$$

where I_{sat} is approximately 64 W/m^2 for sodium. When the incident laser intensity is equal to I_{sat}, one-quarter of the atoms are in the excited state. As the intensity increases, the absorption decreases since fewer atoms are in the ground state. In the limit $I \gg I_{\text{sat}}$, the maximum excited state population of one-half is achieved. At these intensities, stimulated emission becomes the dominant deexcitation mechanism, causing the fluorescence to saturate.

Welsh and Gardner,[481] Gardner,[482] and Gardner, Welsh, and Thompson[474] have treated laser excitation of mesospheric sodium extensively, especially in relation to generation of laser guide stars for adaptive optics. Their analysis uses a rate equation, two level atom approach and accounts for laser pulse shape, spatial intensity, and linewidth. Their cross-section computations account for saturation in terms of the laser beam parameters. The level of saturation is characterized by a saturation time τ_s, which depends on pulse energy, pulse length, and laser beam size. In terms of a lidar equation, their analyses can be slightly extended to predict the number of photodetection events in a subaperture of area A_{sub} as

$$N_{\text{pde}} = \eta_{\text{QE}} T_t T_r T_{\text{atm}}^2 E_p \frac{\lambda}{hc} \frac{A_{\text{sub}}}{4\pi h_s^2} \frac{C_s \sigma_{\text{eff}}}{[1 + (\tau_n/\tau_s)]} . \tag{1.272}$$

In this equation, h_s is the height of the sodium layer above the site, C_s is the sodium column density, σ_{eff} is the effective cross section of the sodium atom's interaction to the laser beam, the factor $(1 + \tau_n/\tau_s)$ is due to saturation, and τ_n is the natural radiative lifetime of sodium (16 ns). Note that τ_s should be long compared to τ_n to reduce saturation effects and losses. The saturation time τ_s is given by

$$\tau_s = \frac{2\Delta t \pi \Delta \rho_{\text{rms}}^2 hc}{E_p \lambda T_{\text{atm}} \sigma_{\text{eff}}} . \tag{1.273}$$

In this equation, $\Delta \rho_{\text{rms}}$ is the root-mean-square radius of the Gaussian laser beam profile in the mesosphere and Δt is the laser pulse width (assumed

rectangular in time). The rms radius of the beam is (for a beam profile limited by uncorrected turbulence)

$$\Delta\rho_{\mathrm{rms}} = 1.22\frac{\lambda h_s}{r_0} \ . \tag{1.274}$$

The nature of the sodium layer places some rather difficult to achieve requirements on the characteristics of the laser used for excitation. The laser linewidth and line shape should be matched to the absorption profile, and in general circular polarization should be used to maximize the effects of optical pumping and prevent optical transparency. Jeys[483,484] has been most successful in developing a laser source for sodium layer excitation by sum frequency mixing two strong lines of the Nd:YAG laser. By an incredible coincidence in nature, the sum frequency of 1.06-μm light and 1.319-μm light generates 0.589-μm light. These lasers are configured to generate a 100-MHz train of mode-locked pulses each 0.4 to 0.7 ns wide in a 60- to 100-μs-long macro pulse. A train of mode-locked pulses is not the ideal format to minimize saturation but is essential for generating the high peak powers needed for efficient sum frequency conversion in the nonlinear crystal and for creating a spectrum of frequencies to access the inhomogeneously broadened absorption spectrum present in the mesosphere.

1.7.4   System Considerations

A laser guide star adaptive optics system can be integrated into a large ground-based telescope in several configurations. The hardware (in particular, the laser) will vary considerably depending on the wavelength region of interest for the scientific instruments. An infrared system for instance may employ a very modest (10- to 20-W) sodium wavelength laser (or even a laser for creating Rayleigh beacons) and tens of actuators, while a visible wavelength system may need a several hundred watt laser and several hundred actuators. Giant telescopes of the 8-m class may need several sodium beacons to reduce the effects of focus anisoplanatism at wavelengths shorter than 1 μm.

A major design choice is where to locate the laser beam: (1) on the same mount as the telescope or (2) on a separate mount. If the laser beam is propagated from the telescope mount, it may (1) share the telescope aperture or (2) be sent out a separate aperture. If a separate aperture is used, it may be (1) offset from the telescope axis or (2) coincident with it (e.g., mounted behind the secondary). Some of these choices will be influenced by the size, wavelength, and beam quality of the laser. It would be practical to mount a compact, water-cooled, laser-diode-pumped, solid state sodium frequency laser on an 8- or even 4-m telescope gimbal. However, a 250-W copper vapor or excimer laser system would best be located off mount and may require sharing nearly the full aperture of the telescope to reduce the beam divergence geometrically because of poor beam quality.

Laser beams transmitted from separate mounts create elongated guide stars viewed by a bistatic telescope. This elongation is significant even at the altitude of the sodium layer and creates additional system complexity if means are devised to compensate for this effect.[485] However, transmitting the laser from a separate mount tens of meters away from the telescope eliminates optical

component and near-field atmospheric backscatter that may interfere with (or even overwhelm) the scientific instruments. In general, light at the laser frequency will be filtered out of the scientific camera (either by use of a special filter or automatically by virtue of existing spectral filters in the setup). If a laser mounting configuration is chosen that puts near-field scatter into the field of the scientific instruments, much greater extinction will be needed in these filters, potentially compromising the performance of the instruments.

The adaptive optical components such as the fast steering mirror for full-aperture tilt correction, the deformable mirror, the wavefront sensor, and relay optics may be mounted at the Cassegrain, nasmyth, or coudé focus of modern large telescopes. If these components experience large temperature variations during an observation session, special installations may be required to preserve overall system performance. A means must also be provided for wavefront reference sources in order to calibrate out optical imperfections in the system.

References

References for Section 1.2

1. Fizeau, *Compt. Rend.* **66,** 932 (1868).
2. M. Born and E. Wolf, *Principles of Optics,* 6th ed., p. 508, Pergamon, London (1980).
3. A. A. Michelson, *Nature* **45,** 160 (1891).
4. A. A. Michelson and F. G. Pease, *Ap. J.* **53,** 249 (1921).
5. R. Hanbury Brown, *The Intensity Interferometer,* Taylor and Francis, London (1974).
6. A. Labeyrie, *Astron. Astrophys.* **6,** 85 (1970).
7. D. Y. Gezari, A. Labeyrie, and R. V. Stachnik, *Ap. J.* **173,** L 1 (1972).
8. J. C. Dainty, in *Symp. Recent Advances in Observational Astronomy,* H. L. Johnson, C. Allen, Eds., p. 95 (1981). A bibliography on stellar interferometry is maintained by the author and is available on request.
9. A. Papoulis, *Probability, Random Variables and Stochastic Processes,* p. 338, 447, McGraw-Hill, New York (1965).
10. J. W. Goodman, *An Introduction to Fourier Optics,* Chap. 6, McGraw-Hill, New York (1968).
11. J. W. Goodman in Ref. 10, Eq. (6-26). Note that the form used in Eq. (1.11) assumes that $H(\xi,\eta)$ is the pupil function, and not its mirror image.
12. R. E. Hufnagel and N. R. Stanley, *J. Opt. Soc. Am.* **54,** 52 (1964).
13. D. L. Fried, *J. Opt. Soc. Am.* **56,** 1372 (1966).
14. F. Roddier, in *Scientific Importance of High Angular Resolution at Infrared and Optical Wavelengths, Proc. ESO Conf.,* p. 5 (1981).
15. F. Roddier, *Progress in Optics,* Vol. 19, p. 281, North-Holland, Amsterdam (1980).
16. R. L. Fante, *Proc. IEEE* **63,** 1669 (1975) and **68,** 1424 (1980).
17. J. C. Dainty and R. J. Scaddan, *M. N. R. A. S.* **170,** 519 (1975).
18. C. Roddier, *J. Opt. Soc. Am.* **66,** 478 (1976).
19. D. S. Brown and R. J. Scaddan, *Observatory* **99,** 125 (1979).
20. D. Korff, *J. Opt. Soc. Am.* **63,** 971 (1973).
21. J. C. Dainty, *Opt. Commun.* **7,** 129 (1973).
22. D. L. Fried, *Opt. Acta* **26,** 597 (1979).
23. C. Aime, S. Kadiri, G. Ricort, C. Roddier, and J. Vernin, *Opt. Acta* **26,** 575 (1979).
24. A. Chelli, P. Lena, C. Roddier, and F. Sibille, *Opt. Acta* **26,** 583 (1979).

25. J. C. Dainty, *M. N. R. A. S.* **169,** 631 (1974).
26. F. Roddier, G. Ricort, and C. Roddier, *Opt. Commun.* **24,** 281 (1978).
27. D. P. Karo and A. M. Schneiderman, *J. Opt. Soc. Am.* **67,** 1277 (1977).
28. K. A. O'Donnell, Thesis, University of Rochester (1983).
29. A. W. Lohmann and G. P. Weigelt, *Optik* **53,** 167 (1979).
30. J. C. Dainty, D. R. Hennings, and K. A. O'Donnell, *J. Opt. Soc. Am.* **71,** 490 (1981).
31. K. A. O'Donnell, B. J. Brames, and J. C. Dainty, *Opt. Commun.* **41,** 79 (1982).
32. R. J. Scaddan and J. G. Walker, *Appl. Opt.* **17,** 3779 (1978).
33. G. Parry, J. G. Walker, and R. J. Scaddan, *Opt. Acta* **26,** 563 (1979).
34. C. Roddier and F. Roddier, *J. Opt. Soc. Am.* **65,** 664 (1975).
35. F. Roddier, J. M. Gilli, and G. Lund, *J. Opt.* **13,** 263 (1982).
36. D. P. Karo and A. M. Schneiderman, *J. Opt. Soc. Am.* **68,** 480 (1978).
37. B. L. Morgan, D. R. Beddoes, R. J. Scaddan, and J. C. Dainty, *M. N. R. A. S.* **183,** 701 (1978).
38. G. M. Morris, *Appl. Opt.* **20,** 2017 (1981); C. G. Wynne, *Opt. Commun.* **28,** 21 (1979).
39. D. Korff, G. Dryden, and R. P. Leavitt, *J. Opt. Soc. Am.* **65,** 1321 (1975).
40. J. H. Shapiro, *J. Opt. Soc. Am.* **66,** 469 (1976); J. H. Shapiro, in *Laser Beam Propagation in the Atmosphere,* J. W. Strohbehn, Ed., *Topics in Appl. Phys.,* Vol. 25, Chap. 3, Springer Verlag, New York (1978).
41. F. Roddier, J. M. Gilli, and J. Vernin, *J. Opt.* **13,** 63 (1982).
42. P. Nisenson and R. V. Stachnik, *J. Opt. Soc. Am.* **68,** 169 (1978).
43. A. M. Schneiderman and D. P. Karo, *J. Opt. Soc. Am.* **68,** 338 (1978).
44. G. P. Weigelt, *Opt. Acta* **26,** 1351 (1979).
45. R. H. Bates and B. L. Davey in *Proceedings NATO Advanced Study Institute,* D. M. Alloin, J.-M. Mariotti, Eds., Kluwer Academic, Norwell, MA (1989).
46. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes,* p. 417, Cambridge University Press, Cambridge (1986).
47. J. C. Christou, D. W. McCarthy, Jr., and M. L. Cobb, *Ap. J.* **94,** 516 (1992).
48. R. J. Keyes, Ed., *Optical and Infrared Detectors,* 2nd ed., *Topics in Appl. Phys.,* Vol. 19, Springer, New York (1980).
49. J. G. Walker, *Opt. Acta* **28,** 885 (1981).
50. M. E. Barnett and G. Parry, *Opt. Commun.* **21,** 60 (1977).
51. J. C. Dainty, *M. N. R. A. S.* **183,** 223 (1978).
52. J. G. Walker, in *High Angular Resolution Stellar Interferometry, Proc. IAU Colloquium No. 50* (1979).
53. J. G. Walker, *Opt. Commun.* **29,** 273 (1979).
54. J. C. Dainty, Ed., *Laser Speckle and Related Phenomena, Topics in Appl. Phys.,* Vol. 9, Chap. 7, Springer, New York (1975).
55. A. H. Greenaway and J. C. Dainty, *Opt. Commun.* **35,** 307 (1980).
56. F. Roddier, in *Imaging in Astronomy, AAS/SAO/OSA/SPIE Topical Meeting,* Reprints ThC6 (1975).
57. J. W. Goodman and J. F. Belsher, *Proc. SPIE* **75,** 141 (1976).
58. J. W. Goodman and J. F. Belsher, RADC-TR-76-50 (Mar. 1976), RADC-TR-76-382 (Dec. 1976), RADC-TR-77-175 (May 1977). All ARPA Order No. 2646 (Rome Air Development Center, Griffin AFB, NY 13441).
59. M. G. Miller, *J. Opt. Soc. Am.* **67,** 1176 (1977).
60. J. C. Dainty and A. H. Greenaway, *J. Opt. Soc. Am.* **69,** 786 (1979).
61. J. C. Dainty and A. H. Greenaway, in *High Angular Resolution Stellar Interferometry, Proc. IAU Colloquium No. 50* (1979).
62. K. A. O'Donnell and J. C. Dainty, *J. Opt. Soc. Am.* **70,** 1354 (1980).
63. A. D. Code, *Astronomical Techniques,* pp. 50–87, University of Chicago Press (1962).
64. E. K. Hege, E. N. Hubbard, P. A. Strittmatter, and S. P. Worden, *Ap. J.* **248,** L 1 (1981).

65. R. L. Fante, *J. Opt. Soc. Am.* **69**, 1394 (1979).

66. L. N. Mertz, *Appl. Opt.* **18**, 611 (1979).

67. H. A. Ferwerda, in *Inverse Source Problems,* H. P. Baltes, Ed., *Topics in Curr. Phys.,* Vol. 9, Chap. 2, Springer, New York (1978).

68. L. S. Taylor, *IEEE Trans AP-29,* 386 (1981).

69. A. Labeyrie, *Progress in Optics,* Vol. 14, p. 47, North-Holland, Amsterdam (1976).

70. S. P. Worden, *Vistas in Astron.* **20**, 301 (1977).

71. A. Labeyrie, *Ann. Rev. Astron. Astrophys.* **16**, 77 (1978).

72. A. H. Greenaway, *J. Optics* (Paris) **10**, 308 (1979).

73. R. H. T. Bates, *Phys. Rep.* **90**, 203 (1982).

74. Lord Rayleigh, *Phil. Mag.* **34**, 407 (1892).

75. E. Wolf, *Proc. Phys. Soc.* **80**, 1269 (1962).

76. C. Roddier and F. Roddier, in *Image Formation from Coherence Functions in Astronomy,* C. van Schooneveld, Ed., p. 175, Reidel, Dordrecht (1979).

77. K. Itoh and Y. Ohtsuka, *Appl. Opt.* **20**, 4239 (1981).

78. A. H. Greenaway, *Opt. Commun.* **42**, 157 (1982).

79. A. M. J. Huiser, *Opt. Commun.* **42**, 226 (1982).

80. R. H. T. Bates, *Mon. Not. R. Astr. Soc.* **142**, 413 (1969).

81. R. E. Burge, M. A. Fiddy, A. H. Greenaway, and G. Ross, *Proc. R. Soc. Lond. Ser. A* **350**, 191 (1976).

82. G. Ross, M. A. Fiddy, M. Nieto-Vesperinas, and M. Wheeler, *Proc. R. Soc. Lond. Ser. A* **360**, 25 (1976).

83. M. A. Fiddy and G. Ross, *Opt. Acta* **26**, 1139 (1979).

84. Yu. M. Bruck and L. G. Sodin, *Opt. Commun.* **30**, 304 (1979).

85. P. J. Napier and R. H. T. Bates, *Astron. Astrophys. Suppl.* **15**, 427 (1974).

86. M. H. Hayes and J. H. McClellan, *Proc. IEEE* **70**, 197 (1982).

87. A. M. J. Huiser and P. van Toorn, *Opt. Lett.* **5**, 499 (1980).

88. P. van Toorn, A. H. Greenaway, and A. M. J. Huiser, *Opt. Acta* **7**, 767 (1984).

89. M. A. Fiddy, B. J. Brames, and J. C. Dainty, *Opt. Lett.* **8**, 96 (1983).

90. J. Fienup, *J. Opt. Soc. Am.* **73**, 1421 (1983).

91. R. H. T. Bates, P. T. Gough, and P. J. Napier, *Astron. Astrophys.* **22**, 319 (1973).

92. P. T. Gough and R. H. T. Bates, *Opt. Acta* **21**, 243 (1974).

93. G. P. Weigelt, *Optik* **43**, 111 (1975).

94. G. P. Weigelt, *Appl. Opt.* **17**, 2660 (1978).

95. G. P. Weigelt, *Proc. SPIE* **243**, 103 (1980).

96. C. Y. C. Liu and A. W. Lohmann, *Opt. Commun.* **8**, 372 (1973).

97. J. E. Baldwin and P. J. Warner, *M. N. R. A. S.* **182**, 411 (1978).

98. J. Christou, *Opt. Commun.* **37**, 331 (1981).

99. J. R. Fienup, T. R. Crimmins, and W. Holsztynski, *J. Opt. Soc. Am.* **72**, 610 (1982).

100. G. P. Weigelt, *Opt. Commun.* **21**, 55 (1977).

101. G. P. Weigelt and B. Wirnitzer, *Opt. Lett.* **8**, 389 (1983).

102. J. R. Fienup, *Opt. Lett.* **3**, 27 (1978).

103. J. R. Fienup, *Opt. Eng.* **18**, 529 (1979).

104. J. R. Fienup, in *Transformations in Optical Signal Processing,* SPIE, Bellingham, WA (1982).

105. J. R. Fienup, RADC-TR-81-63 (May 1981) (Rome Air Development Center, Griffiss AFB, NY 13441).

106. J. R. Fienup, *Appl. Opt.* **21**, 2758 (1982).

107. R. W. Gerchberg and W. O. Saxton, *Optik* **35**, 237 (1972).

108. R. H. T. Bates, *Optik* **61**, 247 (1982).

109. K. L. Garden and R. H. T. Bates, *Optik* **62**, 131 (1982).

110. W. R. Fright and R. H. T. Bates, *Optik* **62,** 219 (1982).

111. R. H. T. Bates and W. R. Fright, *J. Opt. Soc. Am.* **73,** 358 (1983).

112. R. H. T. Bates, *Comp. Graph & Image Proc.* **25,** 205 (1984).

113. R. H. T. Bates and W. R. Fright, *Advances in Computer Vision and Image Processing,* Vol. 1, T. S. Huang, Ed., Chap. 5, JAI Press (1983).

114. S. F. Gull and G. J. Daniell, *Nature* **272,** 686 (1978).

115. J. C. Dainty, M. A. Fiddy, and A. H. Greenaway, in *Image Formation from Coherence Functions in Astronomy,* C. van Schooneveld, Ed., p. 95, Reidel, Dordrecht (1979).

116. J. G. Walker, *Opt. Acta* **28,** 735 (1981).

117. J. W. Wood et al., *Opt. Lett.* **8,** 54 (1981).

118. J. G. Walker, *Opt. Acta* **28,** 1017 (1981).

119. J. G. Walker, *Appl. Opt.* **21,** 3132 (1982).

120. C. R. Lynds, S. P. Worden, and J. W. Harvey, *Ap. J.* **207,** 174 (1976).

121. S. P. Worden, C. R. Lynds, and J. W. Harvey, *J. Opt. Soc. Am.* **66,** 1243 (1976).

122. M. J. McDonnell and R. H. T. Bates, *Ap. J.* **208,** 443 (1976).

123. R. H. T. Bates and F. M. Cady, *Opt. Commun.* **32,** 365 (1980).

124. F. M. Cady and R. H. T. Bates, *Opt. Lett.* **5,** 438 (1980).

125. J. A. Hogbom, *Astron. Astrophys. Suppl.* **15,** 417 (1974).

126. B. R. Hunt, W. R. Fright, and R. H. T. Bates, *J. Opt. Soc. Am.* **73,** 456 (1983).

127. R. H. T. Bates and M. J. McDonnell, *Image Restoration and Reconstruction,* Clarendon Press, Oxford, UK (1986).

128. E. K. Hege, in *Proceedings NATO Advanced Study Institute,* D. M. Alloin, J.-M. Mariotti, Eds., p. 293, Kluwer Academic, Norwell, MA (1989).

129. B. L. McGlamery, NASA TR SP-256, p. 167 (1971).

130. W. J. Cocke, Steward Obs. Reprint No. 264 (1980).

131. B. T. O'Conner and T. S. Huang, *Comp. Graph. Image Proc.* **15,** 25 (1981).

132. J. M. Tribolet, *IEEE Trans. ASSP* **25,** 170 (1977).

133. H. W. Swan, Ph.D. Thesis, Stanford University (1982).

134. K. T. Knox and B. J. Thompson, *Ap. J.* **193,** L 45 (1974).

135. K. T. Knox, *J. Opt. Soc. Am.* **66,** 1236 (1976).

136. P. Nisenson and D. C. Ehn, *Proc. SPIE* **75,** 83 (1976).

137. M. P. Rimmer, *Appl. Opt.* **13,** 623 (1974).

138. J. W. Sherman, *Proc. SPIE* **74,** 249 (1976).

139. R. L. Frost, C. K. Rushforth, and B. S. Baxter, *Appl. Opt.* **18,** 2056 (1979).

140. B. R. Hunt, *J. Opt. Soc. Am.* **69,** 393 (1969).

141. P. Nisenson, R. Stachnik, C. Papaliolios, and P. Horowitz, *Proc. SPIE* **243,** 88 (1980).

142. R. Stachnik, P. Nisenson, and C. Papaliolios, *Solar Instrumentation—What's Next?,* Sunspot, New Mexico (1980).

143. P. Nisenson, R. Stachnik, and C. Papaliolios, in *Optical and Infrared Telescopes for the 1990s,* A. Hewitt, Ed., p. 401, Kitt Peak National Observatory, Tucson, AZ (1980).

144. G. J. M. Aitken and D. L. Desaulniers, *Opt. Commun.* **28,** 26 (1979).

145. J. W. Sherman, *Proc. SPIE* **243,** 51 (1980).

146. B. J. Brames and J. C. Dainty, *J. Opt. Soc. Am.* **71,** 1542 (1981).

147. P. Nisenson and C. Papaliolios, *Opt. Commun.* **45,** 311 (1983).

148. G. Weigelt, in *Proceedings NATO Advanced Study Institute,* D. M. Alloin, J.-M. Mariotti, Eds., p. 191, Kluwer Academic, Norwell, MA (1989).

149. A. Glindemann, R. G. Lane, and J. C. Dainty, in *Proceedings High Resolution Imaging by Interferometry II,* Garching, Germany (1991).

150. G. R. Ayers, M. J. Northcott, and J. C. Dainty, *J. Opt. Soc. Am.* **5,** 963 (1988).

151. T. Nakajima, *J. Opt. Soc. Am.* **5,** 1477 (1988).

152. A. Chelli, in *Proceedings NATO Advanced Study Institute,* D. M. Alloin, J.-M. Mariotti, Eds.,

p. 201, Kluwer Academic, Norwell, MA (1989).

153. A. Labeyrie, "Auxiliary instrumentation for large telescopes," p. 389, ESO/CERN (1972).

154. D. R. Beddoes, J. C. Dainty, B. L. Morgan, and R. J. Scaddan, *J. Opt. Soc. Am.* **66**, 1247 (1976).

155. A. M. Schneiderman and D. P. Karo, *Opt. Eng.* **16**, 72 (1977).

156. J. B. Breckinridge, H. A. McAlister, and W. G. Robinson, *Appl. Opt.* **18**, 1034 (1979).

157. G. Hubbard, K. Hege, M. A. Read, P. A. Strittmatter, N. J. Woolf, and S. P. Worden, *Astron. J.* **84**, 1437 (1979).

158. G. W. Simon, *Astron. J.* **71**, 190 (1966).

159. E. P. Wallner and W. B. Wetherell, *J. Opt. Soc. Am.* **69**, 1413 (1979) (abstract only).

160. P. R. Vokac, *Proc. SPIE* **119**, 223 (1977).

161. A. Blazit, *Image Processing Techniques in Astronomy,* Reidel, Dordrecht (1975).

162. J. Meaburn, B. L. Morgan, H. Vine, A. Pedlor, and R. Spencer, *Nature* **296**, 331 (1982).

163. T. W. Cole, *Proc. ASA* **4**, 19 (1980).

164. D. F. Barbe, Ed., *Charge-Coupled Devices, Topics in Appl. Phys.,* Vol. 38, Springer, New York (1980).

165. C. Papaliolios and L. Mertz, *Proc. SPIE* **331**, 360 (1982).

166. L. N. Mertz, T. D. Tarbell, and A. Title, *Appl. Opt.* **21**, 628 (1982).

167. C. Papaliolios, P. Nisenson, and S. Ebstein, *Appl. Opt.* **24**, 285 (1985).

168. J. G. Timothy and J. S. Morgan, *Proc. SPIE* **627**, 654–659 (1986).

169. M. Clampin, J. Crocker, F. Paresce, and M. Rafal, *Rev. Sci. Instrum.* **59**, 1053 (1988).

170. P. Nisenson, in *Proceedings NATO Advanced Study Institute,* D. M. Alloin, J.-M. Mariotti, Eds., p. 157, Kluwer Academic, Norwell, MA (1989).

171. M. J. Selby, R. Wade, and C. S. Magro, *M. N. R. A S.* **187**, 553 (1979).

172. F. Sibille, A. Chelli, and P. Lena, *Astron. Astrophys.* **79**, 315 (1979).

173. R. R. Howell, D. W. McCarthy, and F. J. Low, *Ap. J.* **251**, L 21 (1981).

174. C. Perrier, *ESO Messenger* **29**, 26 (1981).

175. J. M. Mariotti, A. Chelli, R. Foy, P. Lena, F. Sibille, and G. Tchountonov, *Astron. Astrophys.* **120**, 237 (1983).

176. C. Aime, S. Kadiri, and G. Ricort, *Opt. Commun.* **35**, 169 (1980).

177. C. Aime, J. Demarcq, F. Martin, and G. Ricort, *Opt. Eng.* **22**, 224 (1983).

178. J. M. Beckers, J. C. Christou, R. G. Probst, S. T. Ridgway, and O. von der Luhe, in *High Resolution Imaging by Interferometry,* F. Merkle, Ed., Garching, Germany (1988).

179. D. Gezari, F. Roddier, and C. Roddier, NASA Publication 1245 (1990).

180. D. W. McCarthy, F. J. Low, S. G. Kleinmann, and F. C. Gillett, *Astrophys. J.* **257**, L 7 (1982).

181. S. H. Ebstein, N. P. Carleton, and C. Papaliolios, "Speckle imaging of NGC 1068 and NGC 4151 in the [O III] 5007 A line and nearby continuum," *Ap. J.* **336**, 103–111 (1989).

182. A. Chelli, C. Perrier, I. Cruz-Gonzales, and L. Carrasco, "High angular resolution imaging from the ground using interferometric techniques," in *Proc. 1st ESO/NOAO Workshop,* p. 253 (1987).

183. S. P. Worden, M. K. Stein, G. D. Schmidt, and J. R. P. Angel, *Icarus* **32**, 450 (1977); G. L. Welter, and S. P. Worden, *J. Opt. Soc. Am.* **68**, 1271 (1978).

184. S. P. Worden and M. K. Stein, *A. J.* **84**, 140 (1979).

185. S. P. Worden, *Asteroids,* p. 119, University of Arizona Press (1979).

186. G. Baier and G. Weigelt, "Speckle interferometry observations of the asteroids Juno and Amphitrtrite," *Astron. Astrophys.* **121**, 137–141 (1983).

187. J. D. Drummond, W. J. Cocke, E. K. Hege, P. A. Strittmatter, and J. V. Lambert, "Speckle interferometry of asteroids I. 433 Eros," *Icarus* **61**, 132–151 (1985).

188. J. D. Drummond, E. K. Hege, W. J. Cocke, J. D. Freeman, J. C. Christou, and R. P. Binzel, "Speckle interferometry of asteroids II. 532 Herculina," *Icarus* **61**, 232–240 (1985).

189. J. D. Drummond and E. K. Hege, "Speckle interferometry of asteroids III. 511 Davida and its photometry," *Icarus* **67**, 251–263 (1986).

190. J. D. Drummond, E. K. Hege, and A. Eckart, "Knox-Thompson images of 4 Vesta," *Proc. SPIE* **828,** 27–31 (1987).

191. P. Nisenson, J. Apt, R. Goody, and P. Horowitz, *A. J.* **86,** 1690 (1981).

192. D. Bonneau and R. Foy, "First direct measurements of the diameters of the large satellites of Uranus and Neptune," *Astron. Astrophys.* **161,** L12–L13 (1986).

193. S. J. Arnold, A. Boksenburg, and W. L. W. Sargent, *Ap. J.* **234,** L 159 (1979).

194. D. Bonneau and R. Foy, *Astron. Astrophys.* **92,** L 1 (1980); see also G. Baier, N. Hetterich, and G. Weigelt, *ESO Messenger* **30,** 23 (1982).

195. G. Baier and G. Weigelt, "Speckle interferometric observations of Pluto and its moon Charon on seven different nights," *Astron. Astrophys.* **174,** 295–298 (1987).

196. R. R. Howell and M. T. McGinn, "Infrared speckle interferometry of Io: an eruption in the Loki region," *Science* **230,** 63–65 (1985).

197. J. W. Harvey and J. B. Breckinridge, *Ap. J.* **182,** L 137 (1973).

198. J. W. Harvey and M. Schwarzschild, *Ap. J.* **196,** 221 (1975).

199. R. V. Stachnik, P. Nisenson, D. C. Ehn, R. H. Hudgin, and V. E. Shirf, *Nature* **266,** 149 (1977).

200. C. Aime, "Some solar observations using speckle interferometry techniques," in *Proc. JOSO Conf. Future Solar Optical Observations, Needs and Constraints* (1978).

201. C. Aime, G. Ricort, and J. W. Harvey, "One-dimensional speckle interferometry of the solar granulation," *Ap. J.* **221,** 362–367 (1978).

202. C. Aime, "Observations of the sun with interferometry and speckle-interferometry techniques," in *Proc. High Angular Resolution Stellar Interferometry,* IAU Collection 50, paper 30 (1978).

203. C. Aime, "Application des techniques de speckle interfometrie a l'etude de la granulation solair," *J. Optics* (Paris) **10**(6), 318–319 (1979).

204. O. von der Luhe and R. B. Dunn, "Solar granulation power spectra from speckle interferometry," *Astron. Astrophys.* **177,** 265–276 (1987).

205. P. Druesne, J. Borgnino, F. Martin, G. Ricort, and C. Aime, "Speckle-interferometric study of the solar granulation from centre to limb," *Astron. Astrophys.* **217,** 229–236 (1989).

206. C. U. Keller and O. von der Luhe, *Astron. Astrophys.* **261,** 321 (1992).

207. C. U. Keller, *Nature* **359,** 307 (1992).

208. H. A. McAlister and W. I. Hartkopf, "Catalog of interferometric measurements of binary stars I," in *Proc. Center for High Angular Resolution Astronomy,* Vol. 1, Atlanta, Georgia (1984).

209. H. A. McAlister, W. I. Hartkopf, D. J. Hutter, M. M. Shara, and O. G. Franz, "ICCD speckle observations of binary stars. I. a survey for duplicity among the bright stars," *A. J.* **93,** 183–194 (1987).

210. H. A. McAlister, W. I. Hartkopf, D. J. Hutter, and O. G. Franz, "ICCD speckle observations of binary stars. II. measurements during 1982–1985 from the Kitt Peak 4 m telescope," *A. J.* **93,** 688–723 (1987).

211. P. K. Lu, P. Demarque, W. van Altena, H. A. McAlister, and W. I. Hartkopf, "ICCD speckle observations of binary stars. III. a survey for duplicity among high velocity stars," *A. J.* **94,** 1318–1326 (1987).

212. H. A. McAlister, W. I. Hartkopf, J. R. Sowell, and O. G. Franz, "ICCD speckle observations of binary stars. IV. measurements during 1986 from the Kitt Peak 4 m telescope," *A. J.* (1987).

213. H. A. McAlister, W. I. Hartkopf, J. R. Sowell, E. G. Dombrowski, and O. G. Franz, "ICCD speckle observations of binary stars. V. measurements during 1986–1988 from the Kitt Peak 4 m telescope," *A. J.* **97,** 510–531 (1989).

214. H. A. McAlister, "Masses and luminosities for the giant spectroscopic/speckle interferometric binaries Phi Cygni and Gamma Persei," *A. J.* **87,** 563 (1982).

215. H. A. McAlister and W. J. Hartkopf, "Speckle interferometry of the spectroscopic binary 94 Aquarii A," *P. A. S. P.* **95,** 778 (1983).

216. J. Tomkin, H. A. McAlister, W. I. Hartkopf, and F. C. Fekel, "The orbit of the speckle and double-lined spectroscopic binary Chi Draconis," *A. J.* **93,** 1236 (1987).

217. B. L. Morgan, D. R. Beddoes, R. J. Scaddan, and J. C. Dainty, *M. N. R. A. S.* **183,** 701 (1978).

218. B. L. Morgan, G. K. Beckmann, and R. J. Scaddan, *M. N. R. A. S.* **192,** 143 (1980).

219. B. L. Morgan, G. K. Beckmann, R. J. Scaddan, and H. A. Vine, *M. N. R. A. S.* **198,** 817 (1982).

220. A. Labeyrie, D. Bonneau, R. V. Stachnik, and D. Y. Gezari, *Ap. J.* **194,** L 147 (1974).

221. A. Blazit, D. Bonneau, L. Koechlin, and A. Labeyrie, *Ap. J.* **214,** L 79 (1977).

222. D. Bonneau, A. Blazit, R. Foy, and A. Labeyrie, *Astron. Astrophys. Suppl. Ser.* **42,** 185 (1980).

223. D. Bonneau and R. Foy, *Astron. Astrophys.* **86,** 295 (1980).

224. G. Weigelt, *Astron. Astrophys.* **68,** L 5 (1978).

225. H. A. McAlister, *Modern Astrometry, IAU Colloquium No. 48,* p. 325, University Observatory, Vienna (1978).

226. H. A. McAlister, *Ap. J.* **223,** 526 (1978).

227. H. A. McAlister, *Ap. J.* **212,** 459 (1977).

228. H. A. McAlister, W. I. Hartkopf, W. G. Bagnuolo, Jr., J. R. Sowell, O. G. Franz, and D. S. Evans, "Binary star orbits from speckle interferometry. I. the Hyades binary Finsen 342 (70 Tauri)," *A. J.* **96,** 1431–1438 (1988).

229. W. I. Hartkopf and H. A. McAlister, "Binary star orbits from speckle interferometry. II. combined visual/speckle orbits of 28 close systems," *A. J.* **98,** 1014–1039 (1989).

230. J. M. Beckers, *Current Techniques in Double and Multiple Star Research, IAU Colloquium No. 62,* Flagstaff, AZ (1981).

231. J. M. Beckers, *Opt. Acta* **29,** 361 (1982).

232. R. G. Petrov, "Submilliarcsecond imaging of rotating stars using differential speckle interferometry," in *Proc. 1st ESO/NOAO Workshop High Angular Resolution Imaging from the Ground Using Interferometric Techniques,* pp. 89–92 (1987).

233. H. M. Dyck, T. Simon, and B. Zuckerman, "Discovery of an infrared companion to T Tauri," *Ap. J.* **255,** L103–L106 (1982).

234. D. W. McCarthy, M. L. Cobb, and R. G. Probst, "Gliese 866: a new, low-mass binary in the solar neighborhood," *A. J.* **93**(6), 1535–1538 (1987).

235. D. W. McCarthy and T. J. Henry, "Direct infrared observations of the very low mass object Gliese 623B," *Ap. J.* **319,** L93–L98 (1987).

236. D. W. McCarthy, "Imaging of low mass binary companions and circumstellar disks," in *Proc. 1st ESO/NOAO Workshop High Angular Resolution Imaging from the Ground Using Interferometric Techniques,* pp. 237–240 (1987).

237. A. Chelli, H. Zinneker, L. Carrasco, I. Cruz-Gonzales, and C. Perrier, "Infrared companion to T Tauri stars," *Astron. Astrophys.* **207,** 46–54 (1988).

238. H. Zinnecker, A. Chelli, L. Carrasco, I. Cruz-Gonzales, and C. Perrier, "GSS31: another T Tauri star with an infrared companion," *Proc. Symp. Circumstellar Matter,* IAU Symposium 122 (in press).

239. A. M. Ghez, G. X. Neugebauer, P. W. Gorham, C. A. Haniff, S. R. Kulkarni, and K. Matthews, "Infrared nonredundant mask imaging at Palomar," *A. J.* **102,** 2066 (1991).

240. D. Bonneau and A. Labeyrie, *Ap. J.* **181,** L 1 (1973).

241. M. S. Wilkerson and S. P. Worden, *A. J.* **82,** 642 (1977).

242. C. Roddier and F. Roddier, *Ap. J.* **270,** L 23 (1983).

243. L. Goldberg, E. K. Hege, E. N. Hubbard, P. A. Strittmatter, and W. J. Cocke, "Speckle interferometry of Alpha Ori: preliminary results," in *Proc. 2nd Cambridge Workshop Stars, Stellar Systems and the Sun,* p. 131 (1982).

244. E. K. Hege, J. C. Hebden, and J. C. Christou, "The H Alpha envelope of Alpha Orionis," in *Proc. 4th Cambridge Workshop Cool Stars, Stellar Systems and the Sun* (1985).

245. A. Y. S. Cheng, E. K. Hege, E. N. Hubbard, L. Goldberg, P. A. Strittmatter, and W. J. Cocke, "Diameter and limb-darkening measures for Alpha Orionis," *Ap. J.* **309,** 737–744 (1986).

246. F. Roddier and C. Roddier, "An image reconstruction of Alpha Orionis," *Ap. J.* **295,** L21–L23 (1985).

247. J. C. Hebden, J. C. Christou, Y. S.Cheng, E. K. Hege, P. A. Strittmatter, J. M. Beckers, and H. P. Murphy, "Two dimensional images of Alpha Orionis," *Ap. J.* **309,** 745–754 (1986).

248. M. Karovska, P. Nisenson, and R. W. Noyes, "Resolution of the halo binary mu CAS at optical wavelengths," *Ap. J.* **308,** 260–269 (1986).

249. J. C. Hebden, A. Eckart, and E. K. Hege, "The H Alpha chromosphere of Alpha Orionis," *Ap. J.* **314,** 690–698 (1987).

250. C. Roddier and F. Roddier, "The Alpha Ori envelope and its evolution," in *Proc. 1st ESO/NOAO Workshop High Angular Resolution Imaging from the Ground Using Interferometric Techniques,* p. 231 (1987).

251. A. Labeyrie, L. Koechlin, D. Bonneau, A. Blazit, and R. Foy, *Ap. J.* **218,** L 75 (1977).

252. J. Christou and S. P. Worden, *A. J.* **85,** 302 (1980).

253. M. Karovska, P. Nisenson, and C. Papaliolios, *Ap. J.* **374,** L51 (1991).

254. J. Meaburn, J. R. Walsh, J. C. Hebden, B. L. Morgan, and H. Vine, "Speckle observations of Eta Carinae," *M. N. R. A. S.* **204,** 41P–46P (1983).

255. A. Chelli, C. Perrier, and Y. Biraud, "One dimensional high resolution image reconstruction of Eta Carinae at 4.6 microns with speckle data, " *Astron. Astrophys.* **117,** 199–204 (1983).

256. G. Weigelt and J. Ebersberger, "Eta Carinae resolved by speckle interferometry," *Astron. Astrophys.* **163,** L5–L6 (1986).

257. K. G. Hofmann and G. Weigelt, "Speckle masking observation of Eta Carinae," *Astron. Astrophys.* **203,** L21–L22 (1988).

258. G.Weigelt and R. Baier, "R136a in the 30 Doradus nebula resolved by holographic speckle interferometry," *Astron. Astrophys.* **150,** L18–L20 (1985).

259. M. Karovska, L. Koechlin, P. Nisenson, C. Papaliolios, and C. Standley, *Ap. J.* **340,** 435 (1989).

260. P. E. J. Nulsen, P. R. Wood, P. R. Gillingham, M. S. Bessell, M. A. Dopita, and C. McCowage, *Ap. J.* **358,** 266 (1990).

261. P. R. Wood, P. E. J. Nulsen, P. R. Gillingham, M. S. Bessell, M. A. Dopita, and C. McCowage, *Ap. J.* **339,** 1073 (1990).

262. M. Karovska, P. Nisenson, C. Standley, and S. R. Heathcote, *Ap. J.* **367,** L15 (1991).

263. C. Papliolios, M. Karovska, L. Koechlin, P. Nisenson, C. Standley, and S. Heathcote, *Nature* **338,** 565 (1989).

264. P. Nisenson, C. Papaliolios, M. Karovska, and R. Noyes, *Ap. J.* **320,** L15 (1987).

265. A. Blazit, D. Bonneau, M. Josse, L. Koechlin, A. Labeyrie, and J. L. Onéto, *Astrophys. J.* **217,** L 55 (1977).

266. A. Chelli, P. Lena, and F. Sibille, *Nature* **278,** 143 (1979).

267. H. M. Dyck and R. R. Howell, *A. J.* **87,** 400 (1982).

268. D. W. McCarthy, *Ap. J.* **257,** L 93 (1982).

269. D. W. McCarthy, "The search for substellar companions to nearby stars: infrared imaging from the ground and from space," *Proc. Conf. Astrophysics of Brown Dwarfs* (1986).

270. D. W. McCarthy, Jr., J. C. Christou, and T. J. Henry, "Near-infrared imaging of low mass objects as close companions to nearby stars," *Proc. ESO Conf. No. 29,* p. 533 (1988).

271. R. Foy, A. Chelli, F. Sibille, and P. Lena, *Astron. Astrophys.* **79,** L 5 (1979).

272. H. M. Dyck, S. Beckwith, and B. Zuckerman, *Ap. J.* **271,** L 79 (1983).

273. D. A. Allen, J. R. Barton, and P. T. Wallace, *M. N. R. A. S.* **196,** 797 (1981).

274. H. M. Dyck, T. Simon, and R. D. Wolstencroft, "The infrared dust shell around the WC9 star Ve 2-45," *Ap. J.* **277,** 675–677 (1984).

275. S. T. Ridgway, "Infrared interferometric studies of circumstellar dust shells," in *Proc. 1st ESO/NOAO Workshop High Angular Resolution Imaging from the Ground Using Interferometric Techniques,* pp. 245–246 (1987).

276. H. M. Dyck, T. Simon, and B. Zuckerman, *Ap. J.* **255,** L103 (1982).

277. A. Chelli, C. Perrier, and Y. G. Biraud, *Astron. Astrophys.* **117,** 199 (1983).

References for Section 1.3

278. (a). J. B. Breckinridge, "Coherence interferometer and astronomical applications," *Appl. Opt.* **11**(12), 2996–2998 (1972); (b) J. B. Breckinridge, "A spatial structure interferometer," *Pub. Astr. Soc. Pac.* **85,** 524–525 (1973).

279. C. E. KenKnight, "Autocorrelation methods to obtain diffraction limited resolution with large telescopes," *Ap. J.* **176,** L 43–L 45 (1972).

280. F. Roddier, "Interferometric imaging in optical astronomy," *Physics Reports* **170**(2), 97–166 (1988).

281. W. T. Rhodes and J. W. Goodman, "Interferometric technique for recording and restoring images degraded by unknown aberrations, *J. Opt. Soc. Am.* **63,** 647–657 (1973).

282. F. D. Russel and J. W. Goodman, "Non-redundant arrays and post-detection processing for aberration compensation in incoherent imaging, *J. Opt. Soc. Am.* **61,** 182–191 (1971).

283. D. F. Buscher, C. A. Haniff, J. E. Baldwin, and J. P. Warner, "Detection of a bright feature on the surface of Betelgeuse," *M. N. R. A. S.* **245,** 7p–11p (1990).

284. P. W. Gorham, A. M. Ghez, S. R. Kulkarni,T. Nakajima, G. Neugebauer, J. B. Oke, and T. A. Prince, "Diffraction-limited imaging III. 30 mas closure phase imaging of six binary stars with the Hale 5 m telescope, *A. J.* **98**(5), 1783–1799 (1989).

285. C. A. Haniff, C. D. Mackay, D. J. Titterington, D. Silvia, J. E. Baldwin, and P. J. Warner, "The first images from optical aperture synthesis," *Nature* **328,** 694–696 (1987).

286. T. Nakajima, S. R. Kulkarni, and P. W. Gorham, "Diffraction-limited imaging II. Optical aperture-synthesis imaging of two binary stars," *A. J.* **97**(5), 1510 (1989).

287. A. C. S. Readhead, T. S. Nakajima, T. J. Pearson, G. Neugebauer, J. B. Oke, and W. L. W. Sargent, "Diffraction-limited imaging with ground-based optical telescopes," *A. J.* **95**(4), 1278–1296 (1988).

288. F. Roddier, "Speckle interferometry through small multiple apertures: Michelson stellar interferometry and aperture synthesis," *Opt. Commun.* **10,** 103–105 (1974).

289. C. Roddier and F. Roddier, "Pupil-plane interferometry," in *Diffraction-limited Imaging with Very Large Telescopes,* D. M. Alloin and J. M. Mariotti, Eds., Kluwer, Norwell, MA, pp. 221–236 (1989).

290. F. Roddier and C. Roddier, "Image reconstruction from rotational shear interferograms: laboratory and astronomical results," *Proc. SPIE* **828,** 108–113 (1987).

291. F. Roddier and C. Roddier, "Image reconstruction from rotational shearing interferograms," *Proc. SPIE* **879,** 4–9 (1988).

292. F. Roddier, C. Roddier, and S. Van Peursem, "Diffraction-limited imaging through aberrated optics using pupi-plane and/or image plane information, *Proc. SPIE* **1059,** 173–179 (1989).

293. M. Cagnet, "Measures interférometriques des diamètres apparents stellaires," *Opt. Commun.* **8,** 430–434 (1973).

294. E. Ribak, "A shearing, modulating interferometer," *Proc. ESO Conf. Scientific Importance of High Angular Resolution at Infrared and Optical Wavelengths,* M. H. Ulrich and K. Kjar, Eds., pp. 171–173 (1981).

295. E. Ribak and S. G. Lipson, "Complex spatial coherence function: its measurement by means of phase modulated shearing interferometer," *Appl. Opt.* **20,** 1102–1106 (1981).

296. E. Ribak, E. Leibowitz, and K. Hege, "Real-time fringe contrast measurement in stellar interferometry," *Proc. SPIE* **564,** 201–208 (1985).

297. E. Ribak and E. Leibowitz, "Shearing stellar interferometer. 1: Digital data analysis scheme," *Appl. Opt.* **24,** 3088–3093 (1985).

298. E. Ribak, E. Leibowitz, and E. K. Hege, "Shearing stellar interferometer. 2. Opto-electronic phase-locked system," *Appl. Opt.* **24,** 3094–3100 (1985).

299. W. M. Sinton, "An achromatic stellar interferometer," *A. J.* **59,** 369–375 (1954).

300. A. A. Tokovinin, "A phase grating stellar interferometer," *Sov. Astron. Lett.* **5,** 229–231 (1979).

301. W. C. Wickes and R. H. Dicke, "Achromatic double star interferometry," *A. J.* **79,** 1433–1444 (1974).

302. L. N. Mertz, *Transformations in Optics,* John Wiley & Sons, New York (1970).

303. D. G. Currie, S. L. Knapp, and K. M. Liewer, "Four stellar diameter measurements by a new technique, amplitude interferometry," *Ap. J.* **187,** 131–134 (1974).

304. D. G. Currie, "Binary star observations with the multi-aperture amplitude interferometer, *Proc. IAU Colloquium No. 62 Current Techniques in Double and Multiple Star Research,* R. S. Harrington, O. G. Franz, Eds., *Lowell Obs. Bull.* **167,** pp. 202–216 (1983).

305. D. G. Currie, The University of Maryland program on multi-aperture amplitude interfer-

ometry, *Proc. of the 1st ESO/NOAO Workshop on High-Resolution Imaging from the Ground Using Interferometric Techniques,* J. W. Goad, Ed., Oracle, Arizona, Jan. 12–15, 1987 NOAO Publ., 21–24 (1987).

306. K. Itoh and Y. Ohtsuka, "Spatial coherence measurements through turbulent atmosphere using a computer aided interferometer," *Opt. Commun.* **36**(4), 250–254 (1981).

307. K. Itoh and Y. Ohtsuka, "Interferometric image reconstruction through the turbulent atmosphere," *Appl. Opt.* **20**(24), 4239–4244 (1981).

308. Y. Ohtsuka, "Proposal for the determination of the complex degree of spatial coherence," *Opt. Lett.* **1**(4), 133–134 (1977).

309. S. Lowenthal, J. Serres, and C. Froehly, "Enregistrements d'hologrammes en lumière spatialement incohérente," *C. R. Acad. Sc. (Paris)* **268 B,** 841–844 (1969).

310. O. Bryngdahl and A. Lohmann, "Variable magnification in incoherent holography," *Appl. Opt.* **9**(1), 231–232 (1970).

311. J. D. Armitage and A. Lohmann, "Rotary shearing interferometry," *Opt. Acta* **12,** 185 (1965).

312. M. V. R. K. Murty, "Interference between wavefronts rotated or reversed with respect to each other and its relation to spatial coherence," *J. Opt. Soc. Am.* **54**(10), 1187–1190 (1964).

313. M. V. R. K. Murty and E. C. Hagerott, "Rotational-shearing interferometry," *Appl. Opt.* **5**(4), 615–619 (1966).

314. J. B. Breckinridge, "Two-dimensional white light coherence interferometer," *Appl. Opt.* **13**(12), 2760–2762 (1974).

315. J. B. Breckinridge, "Obtaining information through the atmosphere at the diffraction limit of a large aperture," *J. Opt. Soc. Am.* **65**(7), 755–759 (1975).

316. J. B. Breckinridge, "A white-light amplitude interferometer with 180° rotational shear," *Opt. Eng.* **17**(2), 156–159 (1978).

317. K. Itoh and Y. Ohtsuka, "Interferometric imaging of a thermally luminous two-dimensional object," *Opt. Commun.* **48**(2), 75–79 (1983).

318. C. Roddier, "Measurements of the atmospheric attenuation of the spectral components of astronomical images," *J. Opt. Soc. Am.* **66,** 478–482 (1976).

319. C. Roddier and F. Roddier, "Imaging with a coherence interferometer in optical astronomy," *Proc. IAU colloq.* **49,** 175–179 (1979).

320. C. Roddier, F. Roddier, and J. Demarcq, "Compact rotational shearing interferometer for astronomical applications," *Opt. Eng.* **28**(1), 66–70 (1989).

321. C. Roddier, F. Roddier, F. Martin, A. Baranne, and R. Brun, "Twin-image holography with spectrally broad light," *J. Optics (Paris)* **11**(3), 149–152 (1980).

322. P. Ghez, J.-M. Mariotti, J.-L. Monin, C. Perrier, C. Dainty, and J. J. Fuensalida, "Diffraction-limited observations of astronomical sources with a rotation shearing interferometer in the near infrared," *Proc. SPIE* **1351,** 689–697 (1990).

323. J.-M. Mariotti, A. Zadrozny, I. Vauglin, and J.-L. Monin, "A rotation shearing interferometer for the near infrared," *Workshop on Ground-based astronomical observations with infrared array detectors,* Hilo, Hawaii, March 24–26, 1987, C. G. WynnWilliams and E. E. Becklin, Eds., University of Hawaii Publ., 468–472 (1987).

324. J.-M. Mariotti, J.-L. Monin, A. Zadrozny, and C. Perrier, "A near infrared rotation shearing interferometer," *Proc. ESO-NOAO Conf. High-Resolution Imaging by Interferometry,* F. Merkle, Ed., Garching, Germany, March 15–18, 1988, ESO Publ., 281–292 (1988).

325. J.-M. Mariotti, J.-L. Monin, P. Ghez, C. Perrier, and A. Zadrozny, "Pupil plane interferometry in the near infrared: I. Methodology of observation and first results," *Astron. Astrophys.* **255,** 462–476 (1992).

326. F. Roddier, C. Roddier, and J. Demarcq, "A rotation shearing interferometer with phase-compensated roof-prisms," *J. Optics (Paris)* **9**(3), 145–149 (1978).

327. J. W. Goodman, "Analogy between holography and interferometric image formation," *J. Opt. Soc. Am.* **60**(4), 506–509 (1970).

328. G. Cochran, "New methods of making Fresnel transforms with incoherent light," *J. Opt. Soc. Am.* **56**(11), 1513–1517 (1966).

329. P. J. Peters, "Incoherent holograms with mercury light source," *Appl. Phys. Let.* **8**(8), 209–210 (1966).

330. G. W. Stroke and R. C. Reistrick III, "Holography with spatially noncoherent light," *Appl.*

Phys. Let. **7**(9), 229–230 (1965).

331. S. Lowenthal, J. Serres, and H. Arsenault, "Resolution and film-grain noise in Fourier transform holograms recorded with coherent or spatially incoherent light," *Opt. Commun.* **1**, 438–442 (1970).

332. E. Ribak, C. Roddier, F. Roddier, and J. Breckinridge, "Signal-to-noise limitations in white light holography," *Appl. Opt.* **27**(5), 1183–1186 (1988).

333. J. W. Goodman, "Comparative sensitivities of stellar speckle interferometry and amplitude interferometry, summary," *J. Opt. Soc. Am.* **66**(10), 1077 (1976).

334. J. W. Goodman and J. F. Belsher, "Photon limitations in imaging and image restoration, Report TR-76-382, ARPA order no. 2646 (1976).

335. F. Roddier, "The effects of atmospheric turbulence in optical astronomy," *Progress in Optics* **19**(5), 281–376 (1981).

336. A. H. Greenaway, "On long-baseline amplitude interferometers in astronomical applications," *Opt. Acta* **25**, 181 (1978).

337. J. J. Burke and J. B. Breckinridge, "Passive imaging through the turbulent atmosphere: fundamental limits on the spatial frequency resolution of a rotational shearing interferometer," *J. Opt. Soc. Am.* **68**, 67–77 (1978).

338. A. Chelli and J.-M. Mariotti, "Visibility and phase analysis for image and pupil-plane interferometry at optical wavelengths," *Astron. Astrophys.* **157**, 372–382 (1986).

339. J. C. Dainty and A. H. Greenaway, "The signal-to-noise ratio in speckle interferometry," *Proc. IAU Colloquium* **50**, Paper 23 (1979).

340. K. Itoh and Y. Ohtsuka, "Photon noise limitations in wavefront folding interferometry," *J. Opt. Soc. Am.* **73**(4), 479–485 (1983).

341. F. Roddier, "How to achieve diffraction-limited resolution with large space telescopes," in *Proc. COSPAR Symposium, Adv. Space Res.* **2**(4), 3–9 (1983).

342. F. Roddier, "Pupil plane versus image plane in Michelson stellar interferometry," *J. Opt. Soc. Am. A* **3**(12), 2160–2166 (1986).

343. L. N. Mertz, "Prism configurations for wavefront sensing," *Appl. Opt.* **29**(25), 3573–3574 (1990).

344. F. Roddier and C. Roddier, "Phase closure with rotational shear interferometers," *ESO–NOAO Workshop on High Resolution Imaging from the Ground Using Interferometric Techniques,* Oracle, Arizona, Jan. 12–15, 1987, J. Goad, Ed., NOAO, 79–82 (1987).

345. A. H. Greenaway and C. Roddier, "A 180-degree rotation shearing interferometer with increased optical efficiency," *Opt. Commun.* **32**, 48–50 (1980).

346. C. Roddier and F. Roddier, "Interferogram analysis using Fourier transform techniques," *Appl. Opt.* **26**(9), 1668–1673 (1987).

347. F. Roddier and C. Roddier, "Imaging with a multi-mirror telescope," *Proc. ESO/CERN Conf. Optical Telescopes of the Future,* Geneva, Switzerland, Dec. 12–15, 1977, F. Pacini, W. Ritcher, and R. N. Wilson, Eds., pp. 359–370 (1978).

348. F. Roddier, "Rotation-shearing interferometry," *Proc. IAU Coll. 50 High Angular Resolution Stellar Interferometry,* Univ. of Maryland, Aug. 30–Sept. 1, 1978, J. Davis and W. J. Tango, Eds., Sydney Univ., pp. 32-1 to 32-15 (1979).

349. K. Itoh, "Analysis of the phase unwrapping algorithm," *Appl. Opt.* **21**(14), 2470 (1982).

350. K. Itoh and Y. Ohtsuka, "Phase estimation based on the maximum likelihood criterion," *Appl. Opt.* **22**(19), 3054–3057 (1983).

351. L. N. Mertz, "Speckle imaging, photon by photon," *Appl. Opt.* **18**(5), 611–614 (1979).

352. J.-L. Monin, J.-M. Mariotti, P. Ghez, C. Perrier, and L. Desbat, "Pupil plane interferometry in the near infrared: II. Phase recovery and image reconstruction," *Astron. Astrophys.* **260**, 510–516 (1992).

353. J. Herrmann, "Least squares wavefront errors of minimum norm," *J. Opt. Soc. Am.* **70**, 28–35 (1980).

354. R. H. Hudgin, "Wavefront reconstruction for compensated imaging," *J. Opt. Soc. Am.* **67**, 375–378 (1977).

355. H. Takajo and T. Takahashi, "Least-squares phase estimation from the phase difference," *J. Opt. Soc. Am. A* **5**(3), 416–425 (1988).

356. R. Ribak, "Phase closure with a rotational shear interferometer," *Appl. Opt.* **26**, 197–199 (1987).

357. E. Ribak, "Phase relations with a rotational shear interferometer," *ESO-NOAO Workshop on High Resolution Imaging from the Ground Using Interferometric Techniques,* Oracle, Arizona, Jan. 12–15, 1987, J. Goad, Ed., NOAO, pp. 215–218 (1987).

358. E. Ribak, "Phase relations and imaging in pupil plane interferometry," *Proc. ESO-NOAO Conf. High-Resolution Imaging by Interferometry,* Garching, Germany, March 15–18, 1988, F. Merkle, Ed., ESO, 271–280 (1988).

359. E. Ribak, "Astronomical imaging by pupil plane interferometry," *Proc. SPIE* **1038,** 418–426 (1988).

360. K.-H. Hofmann and G. Weigelt, "High angular resolution shearing spectroscopy and triple shearing interferometry," *Appl. Opt.* **25,** 4280–4287 (1986).

361. K. H. Hofmann and G. Weigelt, "Triple shearing interferometry and shearing spectroscopy," *Proc. ESO-NOAO Conf. High-Resolution Imaging by Interferometry,* F. Merkle, Ed., Garching, Germany, March 15–18, 1988, ESO Publ., 281–292 (1988).

362. F. Roddier and C. Roddier, "Phase closure with rotational shear interferometers," *Opt. Commun.* **60**(6), 350–352 (1986).

363. C. Roddier and F. Roddier, "Phase recovery from pupil-plane interferograms," *Proc. ESO-NOAO Conf. High-Resolution Imaging by Interferometry,* F. Merkle, Ed., Garching, Germany, March 15–18, 1988, ESO Publ., 257–265 (1988).

364. C. Roddier and F. Roddier, "Diffraction-limited imaging of unknown objects through fixed unknown aberrations using interferometry," *J. Opt. Soc. Am.* **7**(10), 1824–1833 (1990).

365. C. Roddier and F. Roddier, "Self-calibration in rotational shearing interferometry," *Proc. SPIE* **1351,** 698–705 (1990).

366. A. A. Michelson and F. G. Pease, "Measurement of the diameter of α Orionis with the interferometer," *Astrophys. J.* **53,** 249–259 (1921).

367. F. Roddier and C. Roddier, "High angular resolution with rotation shearing interferometer: preliminary results and future potentials," *Proc. IAU Colloquium* **67,** 207–211 (1982).

368. C. Roddier and F. Roddier, "High angular resolution observations of Alpha Orionis with a rotation shearing interferometer," *Astrophys. J.* **270,** L23–L26 (1983).

369. F. Roddier and C. Roddier, "An image reconstruction of Alpha Orionis," *Astrophys. J.* **295,** L21–L23 (1985).

370. F. Roddier, C. Roddier, R. Petrov, F. Martin, G. Ricort, and C. Aime, "New observations of Alpha Orionis with a rotation shearing interferometer," *Astrophys. J.* **305,** L77–L80 (1986); F. Roddier, C. Roddier, M. Karovska, "High angular resolution interferometric observations of Betelgeuse in the visible," *Proc. UCLA Workshop Mass Loss from Red Giants,* M. Morris and B. Zuckerman, Eds., p. 63, Reidel (1984).

371. J. L. Monin, I. Vauglin, F. Sibille, and L. Audaire, "A new infrared camera for the 2–5 μm range," *Astron. Astrophys.* **172,** 368–374 (1987).

372. J.-L. Monin, P. Ghez, C. Perrier, J.-M. Mariotti, and L. Desbat, "Phase recovery and image reconstruction in pupil plane interferometry," *ESO Conf. High-resolution imaging by interferometry II,* J. Beckers and F. Merkle, Eds., Garching, Germany, October 14–18, 1991, ESO Publ., 863–872 (1992).

373. K. T. Knox and B. J. Thompson, "Recovery of images from atmospherically degraded short-exposure photographs," *Astrophys. J.* **193,** L45–L48 (1974).

374. P. Ghez, J.-L.Monin, C. Perrier, and J.-M. Mariotti, "Diffraction-limited imaging of circumstellar shells by rotation-shearing interferometry," *Proc. ESO Conf. on High-resolution imaging by interferometry II,* J. Beckers and F. Merkle, Eds., Garching, Germany, October 14–18, 1991, ESO Publ., 521–529 (1992).

375. J.-M. Mariotti, "Pupil plane beam-recombination for a non-monolithic array," *Proc. SPIE* **1237,** 647–653 (1990).

376. P. Bely and F. Roddier, "Concepts for a large telescope in space with interferometric imaging," *Proc. AIAA 24th Aerospace Sciences Meeting,* January 6–9, 1986, Reno, Nevada.

377. F. Roddier and J. B. Breckinridge, "Interferometric image reconstruction using the L. D. R. in a light bucket mode," in *Proc. Workshop on Optical Interferometry in Space,* Baltimore, Maryland, R. D. Reasenberg, Ed., *Bull. Am. Ast. Soc.* **16**(3), 832–837 (1984).

378. K. Itoh and Y. Ohtsuka, "Fourier-transform spectral imaging: retrieval of source information from three-dimensional spatial coherence," *J. Opt. Soc. Am. A* **3**(1), 94–100 (1986).

379. K. Itoh and Y. Ohtsuka, "Holographic spectral imaging," *J. Opt. Soc. Am. A* **3**(8), 1239–1242 (1986).

380. K. Itoh, T. Inoue, T. Yoshida, and Y. Ichioka, "Interferometric supermultispectral imaging," *Appl. Opt.* **29**(11), 1625–1630 (1990).

381. K. Itoh, T. Inoue, and Y. Ichioka, "Interferometric spectral imaging and optical three-dimensional Fourier transformation," *Jap. J. Appl. Phys.* **29**(8), L1561–L1564 (1990).

382. J.-M. Mariotti and S. T. Ridgway, "Double Fourier spatio-spectral interferometry: combining high spectral and high spatial resolution in the near infrared," *Astron. Astrophys.* **195**, 350–363 (1988).

383. S. T. Ridgway, J.-M. Mariotti, "A Method for Multispectral Infrared Interferometry," *Proc. 1st ESO/NOAO Workshop High Angular Resolution Imaging from the Ground Using Interferometric Techniques,* J. Goad, Ed., NOAO, 93–96 (1987).

References for Section 1.4

384. S. A. Hovanassian, "Introduction to Synthetic Array and Imaging Radars, Artech House, Dedham, MA (1980).

385. W. M. Brown and J. L. Porcello, "An introduction to synthetic aperture radar," *IEEE Spectrum* **6**, 52 (1969).

386. J. L. Walker, "Range-Doppler imaging of rotating objects," *IEEE Trans, Aerospace Electron. Syst.* **AES-16**, 23–52 (1980).

387. M. Born and E. Wolf, "Interference and diffraction with partially coherent light," Chap. 10 in *Principles of Optics,* 5th ed., Pergamon Press, New York (1975).

388. J. B. Breckinridge, "A white light amplitude interferometer with 180° rotational shear," *Opt. Eng.* **17**, 156–159 (1978).

389. K. Itoh and Y. Ohtsuka, "Interferometric imaging of a thermally luminous two-dimensional object," *Opt. Comm.* **48**, 75–79 (1983).

390. C. Roddier, F. Roddier, and J. Demarcq, "Compact rotational shearing interferometer for astronomical applications," *Opt. Eng.* **28**, 66–70 (1989).

391. E. B. Fomalont, "Earth-rotation aperture synthesis," *Proc. IEEE* **61**, 1211–1218 (1973).

392. A. M. Tai, "Passive synthetic aperture imaging using an achromatic grating interferometer," *Appl. Opt.* **25**, 3179–3190 (1986).

References for Section 1.5

393. J. W. Goodman, *Introduction to Fourier Optics,* McGraw-Hill, San Francisco (1968).

394. J. W. Goodman, "Wavefront-reconstruction imaging of coherently illuminated objects," Chap. 4 in *Restoration of Atmospherically Degraded Images,* Woods Hole Summer Study, National Academy of Sciences/NRC, Washington, D.C. (1966).

395. J. W. Goodman, "Analysis of wavefront reconstruction imaging through random media," Appendix 29 in *Restoration of Atmospherically Degraded Images,* Vol. 2, Woods Hole Summer Study, National Academy of Sciences/NRC, Washington, D.C. (1966).

396. J. W. Goodman, "Holographic arrays," Appendix 22 in *Restoration of Atmospherically Degraded Images,* Woods Hole Summer Study, National Academy of Sciences/NRC, Washington, D.C. (1966).

397. J. W. Goodman, "Statistical properties of laser speckle patterns," Chap. 2 in *Laser Speckle and Related Phenomena,* 2nd ed., J. C. Dainty, Ed., Springer-Verlag, New York (1984).

398. M. Born and E. Wolf, *Progress in Optics,* MacMillan, New York (1964).

399. E. N. Leith and J. Upatnieks, "Reconstructed wavefronts and communication theory," *J. Opt. Soc. Am.* **52**, 1123–1130 (1962).

400. P. S. Idell, J. D. Gonglewski, D. G. Voeltz, and J. Knopp, "Image synthesis from nonimaged laser-speckle patterns: experimental verification," *Opt. Lett.* **14**, 154–156 (1989).

401. T. R. Crimmins, J. R. Fienup, and B. J. Thelen, "Improved bounds on object support from autocorrelation support and application to phase retrieval," *J. Opt. Soc. Am. A* **7**, 3–13 (1990).

402. J. R. Fienup, T. R. Crimmins, and W. Holsztynski, "Reconstruction of the support of an object from the support of its autocorrelation," *J. Opt. Soc. Am.* **72**, 610–624 (1982).

403. J. C. Dainty and J. R. Fienup, "Phase retrieval and image reconstruction for astronomy,"

Chap. 7 in *Image Recovery: Theory and Application,* H. Stark, Ed., pp. 231–275, Academic Press, New York (1987).

404. J. R. Fienup, "Reconstruction of a complex-valued object from the modulus of its Fourier transform using a support constraint," *J. Opt. Soc. Am. A* **4,** 118–123 (1987).

405. R. G. Paxman, J. R. Fienup, and J. T. Clinthorne, "Effect of tapered illumination and Fourier intensity errors on phase retrieval," in *Digital Image Recovery and Synthesis, Proc. SPIE* **828,** 184–189 (1987).

406. J. R. Fienup and A. M. Kowalczyk, "Phase retrieval for a complex-valued object by using a low-resolution image," *J. Opt. Soc. Am. A* **7,** 450–458 (1990).

407. R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik* **35,** 237–246 (1972).

408. J. R. Fienup, "Phase retrieval algorithms: a comparison," *Appl. Opt.* **21,** 2758–2769 (1982).

409. P. S. Idell and J. D. Gonglewski, "Image synthesis from wave-front sensor measurements of a coherent diffraction field," *Opt. Lett.* **15,** 1309–1311 (1990).

410. L. I. Goldfischer, "Autocorrelation function and power spectral density of laser-produced speckle patterns," *J. Opt. Soc. Am.* **55,** 247–253 (1965).

411. M. Elbaum, et al., "Laser correlography: transmission of high-resolution object signatures through the turbulent atmosphere," Riverside Research Institute Report T-1/306-3-11 (Oct. 31, 1974). (Available from NTIS, Report AD-A013-424.)

412. P. S. Idell, J. R. Fienup, and R. S. Goodman, "Image synthesis from nonimaged laser speckle patterns," *Opt. Lett.* **12,** 858–860 (1987).

413. J. R. Fienup and P. S. Idell, "Imaging correlography with sparse arrays of detectors," *Opt. Eng.* **27,** 778–784 (1988).

414. J. W. Goodman, *Statistical Optics,* John Wiley & Sons, New York (1985).

415. R. Hanbury Brown and R. Q. Twiss, "Correlation between photons in two coherent beams of light," *Nature* **177,** 27–29 (1956).

416. C. W. Helstrom, "Image restoration by the method of least squares," *J. Opt. Soc. Am.* **57,** 297–303 (1967).

417. D. G. Voelz, J. D. Gonglewski, and P. S. Idell, "Image synthesis from nonimaged laser-speckle patterns: comparison of theory, computer simulation, and laboratory results," *Appl. Opt.* **30,** 3333–3344 (1991).

418. P. J. Napier, R. T. Thompson, and R. D. Ekers, "The very large array: design and performance of a modern synthesis radio telescope," *Proc. IEEE* **71,** 1295–1320 (1983).

419. K. O'Donnell, "Time-varying speckle phenomena in astronomical imaging and in laser scattering," PhD Thesis, University of Rochester, Rochester, NY (1983).

References for Section 1.6

420. D. A. Ausherman, A. Kozma, J. L. Walker, H. M. Jones, and E. C. Pozzio, "Developments in radar imaging," *IEEE Transactions on Aerospace and Electronic Engineering* **AES-20**(4) (July 1984).

421. J. C. Curlander and R. N. McDonough, *Synthetic Aperture Radar Systems and Signal Processing,* John Wiley & Sons, New York (1991).

422. A. W. Rihaczek, *Principles of High-Resolution Radar,* Peninsula Publishing, Los Altos, CA (1985).

423. C. E. Cook and N. Bernfield, *Radar Signals,* Academic Press, New York (1967).

424. G. W. Deley, "Waveform design," Chapter 3 in *Radar Handbook,* 1st ed., M. I. Skolnik, Ed., McGraw-Hill, New York (1970).

425. C. C. Aleksoff and C. R. Christensen, "Holographic Doppler imaging of rotating objects," *Applied Optics* **14**(1) (January 1975).

426. C. C. Aleksoff et al., "Synthetic aperture imaging with a pulsed CO_2 TEA Laser," *Proceedings SPIE Laser Radar II* **783,** 29–40 (May 1987).

427. A. L. Kackelmyer, "Range-Doppler imaging: Waveforms and receiver design," *Proceedings SPIE Laser Radar III* **999,** 138–161 (September 1988).

428. A. L. Kackelmyer, "Coherent laser radar performance in the presence of random and systematic errors," *Proceedings SPIE Laser Radar III* **999,** 176–191 (September 1988).

429. D. Park and J. H. Shapiro, "Performance analysis of optical synthetic aperture radars," *Proceedings SPIE Laser Radar II* **999**, 100–116 (1988).

430. T. G. Kyle, "High resolution laser imaging system," *Applied Optics* **28**, 2651–2655 (July 1, 1989).

431. R. S. Eng et al., "Conversion efficiency, spectral and mode parities of a single sideband electro-optic modulator," *Proceedings SPIE Laser Radar III* **999**, 40–55 (September 1988).

432. A. L. Kachelmyer, R. E. Knowlden, and W. E. Keicher, "Effect of atmospheric distortion of carbon dioxide laser radar waveforms," *Proceedings SPIE Laser Radar II* **783**, 101–108 (May 1987).

References for Section 1.7

433. V. I. Tatarskii, *Wave Propagation in a Turbulent Medium,* McGraw-Hill, New York (1961).

434. V. I. Tatarskii, *The Effects of the Turbulent Atmosphere on Wave Propagation,* Report TT-68-50464, National Science Foundation (1968).

435. A. Kolmogorov, in *Turbulence, Classical Papers on Statistical Theory,* S. K. Friedlander and L. Topper, Eds., Wiley-Interscience, New York (1961).

436. V. I. Tatarskii, A. S. Gurvich, M. A. Kallistratova, and L. V. Terenteva, "On the influence of meteorological conditions on the intensity of twinkling of light in the layer of atmosphere near the earth," *Astronom. Zh* **35**, 123 (1958).

437. R. E. Hufnagel and N. R. Stanley, "Modulation transfer function associated with image transmission through turbulent media," *J. Opt. Soc. Am.* **54**, 52–61 (1964).

438. D. L. Fried, "Optical resolution through a randomly inhomogeneous medium for very long and very short exposures," *J. Opt. Soc. Am.* **56**(10), 1372–1379 (1966).

439. D. L. Fried, "Statistics of geometric representation of wavefront distortion," *J. Opt. Soc. Am.* **55**(11), 1427–1437 (1965).

440. R. J. Noll, "Zernike polynomials and atmospheric turbulence," *J. Opt. Soc. Am.* **66**(3), 207–211 (1976).

441. D. L. Fried, "Anisoplanatism in adaptive optics," *J. Opt. Soc. Am.,* **72**(1), 52–61 (1982).

442. D. P. Greenwood, "Bandwidth specification for adaptive optics systems," *J. Opt. Soc. Am.* **67**, 390–392 (1977).

443. G. A. Tyler, "Turbulence-induced adaptive-optics performance degradation: evaluation in the time domain," *J. Opt. Soc. Am.* **A1**, 251–264 (1984).

444. H. W. Babcock, "The possibility of compensating astronomical seeing," *Publ. Astron. Soc. Pac.* **65**, 229–237 (1953).

445. W. L. Wolfe and G. J. Zissis, Eds., *The Infrared Handbook,* Environmental Research Institute of Michigan, Ann Arbor, MI (1985).

446. G. P. Collins, "Making stars to see stars: DOD adaptive optics work is declassified," *Physics Today* **45**(2) 17–21 (1992).

447. R. Q. Fugate et al., "Experimental demonstration of real time atmospheric compensation with adaptive optics employing laser guide stars," *Bull. Am. Ast. Soc.* **23**(2), 898 (1991).

448. R. Q. Fugate et al., "Measurement of atmospheric wavefront distortion using scattered light from a laser guide-star," *Nature* **353**, 144–146 (1991).

449. J. F. Belsher and D. L. Fried, "Expected antenna gain when correcting tilt-free wavefronts," Report No. TR-576, Optical Sciences Co., Placentia, CA (1984).

450. D. L. Fried and J. F. Belsher, "Analysis of fundamental limits to artificial guide-star (AGS) adaptive-optics system (AOS) performance for astronomical imaging," accepted by *J. Opt. Soc. Am. A.*

451. R. A. Humphreys, C. A. Primmerman, L. C. Bradley, and J. Hermann, "Atmospheric-turbulence measurements using a synthetic beacon in the mesospheric sodium layer," *Opt. Lett.* **16**, 1367–1369 (1991).

452. R. Foy and A. Labeyrie, "Feasibility of adaptive telescope with laser probe," *Astron. Astrophys.* **152**, L29–L31 (1985).

453. L. A. Thompson and C. S. Gardner, "Experiments on laser guide stars at Mauna Kea observatory for adaptive imaging in astronomy," *Nature* **328**, 229–231 (1987).

454. C. A. Primmerman, D. V. Murphy, D. A. Page, B. G. Zollars, and H. T. Barclay, "Compen-

sation of atmospheric optical distortion using a synthetic beacon," *Nature* **353,** 141–143 (1991).

455. D. V. Murphy, C. A. Primmerman, B. G. Zollars, and H. T. Barclay, "Experimental demonstration of atmospheric compensation using multiple synthetic beacons," *Opt. Lett.* **16,** 1797–1799 (1991).

456. R. Q. Fugate et al., "Two generations of laser guide star adaptive optics experiments at the Starfire Optical Range," accepted by *J. Opt. Soc. Am. A.*

457. R. Q. Fugate, "High bandwidth laser guide star adaptive optics at the Starfire Optical Range," *Proceedings Progress in Large Telescopes '92,* Garching (1992).

458. D. G. Sandler, "A multiple spot laser beacon for high-order wavefront control: theory and experiment," *Laser Guide Star Adaptive Optics Workshop Proceedings,* R. Q. Fugate, Ed., March 10–12, 1992, Albuquerque, NM, Phillips Laboratory, pp.164–195 (1992).

459. D. G. Sandler, "A new 500-channel adaptive optic system for laser guide star experiments," *Laser Guide Star Adaptive Optics Workshop Proceedings,* R. Q. Fugate, Ed., March 10–12, 1992, Albuquerque, NM, Phillips Laboratory, pp. 686–717 (1992).

460. L. A. Thompson, "Experimental demonstration of a Rayleigh guide star at 351 nm," *Laser Guide Star Adaptive Optics Workshop Proceedings,* R. Q. Fugate, Ed., March 10–12, 1992, Albuquerque, NM, Phillips Laboratory, pp. 491–493 (1992).

461. E. Kibblewhite, W. Wild, B. Carter, M. Chun, F. Shi, and M. Smutko, "A description of the Chicago Adaptive Optics System (ChAOS)," *Laser Guide Star Adaptive Optics Workshop Proceedings,* R. Q. Fugate, Ed., March 10–12, 1992, Albuquerque, NM, Phillips Laboratory, pp. 522–534 (1992).

462. C. Max et al., "Feasibility experiment for sodium-layer laser guide stars at Lawrence Livermore National Laboratory," *Laser Guide Star Adaptive Optics Workshop Proceedings,* R. Q. Fugate, Ed., March 10–12, 1992, Albuquerque, NM, Phillips Laboratory, pp. 535–541 (1992).

463. M. Sechaud, N. Hubin, L. Brixon, R. Jolin, R. Foy, and M. Tallon, "Laser backscatter reference star for atmospheric wavefront disturbances," *Very Large Telescopes and Their Instrumentation, Proceedings, Vol. II,* M. H. Ulrich, Ed., European Southern Observatory, pp. 705–714, Garching, Germany (1988).

464. J. M. Beckers, "Increasing the size of the isoplanatic patch with multiconjugate adaptive optics," *Very Large Telescopes and Their Instrumentation, Proceedings, Vol. II,* M. H. Ulrich, Ed., European Southern Observatory, pp. 693–703, Garching, Germany (1988).

465. F. Merkle, "Adaptive optics developments at ESO," *Very Large Telescopes and Their Instrumentation, Proceedings, Vol. II* M. H. Ulrich, Ed., European Southern Observatory, pp. 639–656 (1988).

466. G. A. Tyler, "Rapid evaluation of d_0," accepted by *J. Opt. Soc. Am. A.*

467. R. J. Sasiela, "A unified approach to electromagnetic wave propagation in turbulence and the evaluation of multiparameter integrals," Technical Report 807, MIT Lincoln Laboratory (1988).

468. R. R. Parenti and R. J. Sasiela, "Laser guide-star systems for astronomical applications," submitted to *Proc. IEEE.*

469. B. M. Welsh and C. S. Gardner, "Effects of turbulence-induced anisoplanatism on the imaging performance of adaptive-astronomical telescopes using laser guide stars," *J. Opt. Soc. Am.* **A8**(1), 69–80 (1991).

470. R. J. Sasiela, "Wavefront correction using one or more synthetic beacons," submitted to *Proc. IEEE.*

471. D. L. Fried, "Analysis of focus anisoplanatism: the fundamental limit in performance of artificial guide star adaptive optics systems (AGS/AOS)," *Laser Guide Star Adaptive Optics Workshop Proceedings,* R. Q. Fugate, Ed., March 10–12, 1992, Albuquerque, NM, Phillips Laboratory, pp. 37–80 (1992).

472. R. Penndorf, "Tables of the refractive index for standard air and the Rayleigh scattering coefficient for the spectral region between 0.2 and 20.0 μ and their application to atmospheric optics," *J. Opt. Soc. Am.* **47**(2), 176–182 (1957).

473. National Oceanic and Atmospheric Administration, National Aeronautics and Space Administration, and the U.S. Air Force, *U.S. Standard Atmosphere, 1976,* Washington, D.C. (1976).

474. C. S. Gardner, B. M. Welsh, and L. A. Thompson, "Design and performance analysis of adaptive optical telescopes using laser guide stars," *Proc. IEEE* **703**(11), 1721–1743 (1990).

475. T. H. Jeys, R. M. Heinrichs, K. F. Wall, J. Korn, T. C. Hotaling, and E. Kibblewhite, "Ob-

servation of optical pumping of mesospheric sodium," Post deadline paper 41, Optical Society of America Meeting, San Jose, CA (October 1991).

476. T. H. Jeys, R. M. Heinrichs, K. F. Wall, J. Korn, T. C. Hotaling, and E. Kibblewhite, "Optical pumping of mesospheric sodium," *Laser Guide Star Adaptive Optics Workshop Proceedings,* R. Q. Fugate, Ed., March 10–12, 1992, Albuquerque, NM, Phillips Laboratory, pp. 238–254 (1992).

477. M. P. Jelonek, R. Q. Fugate, W. J. Lange, A. C. Slavin, R. E. Ruane, and R. A. Cleis, "Characterization of artificial guide stars generated in the mesospheric sodium layer with a sum-frequency laser," accepted by *J. Opt. Soc. Am. A.*

478. P. W. Milonni and L. E. Thode, "Theory of mesospheric sodium fluorescence excited by pulse trains," *Appl. Opt.* **31**(6), 785–800 (1992).

479. L. C. Bradley, "Pulse-train excitation of sodium for use as a synthetic beacon," submitted to *J. Opt. Soc. Am. B.*

480. A. E. Siegman, *An Introduction to Lasers and Masers,* McGraw-Hill, New York (1971).

481. B. M. Welsh and C. S. Gardner, "Nonlinear resonant absorption effects on the design of resonance fluorescence lidars and laser guide stars," *Appl. Opt.* **28**(19), 4141–4153 (1989).

482. C. S. Gardner, "Sodium resonance fluorescence lidar applications in atmospheric science and astronomy," *Proc. IEEE* **77**(3), 408–418 (1989).

483. T. H. Jeys, A. A. Brailove, and A. Mooradian, "Sum frequency generation of sodium resonance radiation," *Appl. Opt.* **28,** 2588 (1989).

484. T. H. Jeys, "Development of a mesospheric sodium laser beacon for atmospheric adaptive optics," *The Lincoln Laboratory Journal* 4(2), 133–149 (1991).

485. J. M. Beckers, "Removing perspective elongation effects in laser guide stars and their use in the ESO very large telescope," *Laser Guide Star Adaptive Optics Workshop Proceedings,* R. Q. Fugate, Ed., March 10–12, 1992, Albuquerque, NM, Phillips Laboratory, pp. 629–638 (1992).

Adaptive Optics

Robert K. Tyson
W. J. Schafer Associates
Chelmsford, Massachusetts

Peter B. Ulrich
National Aeronautics and Space Administration Headquarters
Washington, D.C.

CONTENTS

2.1   INTRODUCTION

Adaptive optics is an engineering synthesis of design techniques used to build optical systems that improve performance by dynamically and autonomously correcting for aberrations. In so doing, a property of the system such as image resolution or on-target power density is optimized. Three elements of the system accomplish this: a measuring device that senses the aberrations that the system must compensate for, a processor to convert this information into a usable signal, and a wavefront altering system to undo the damage from the aberrations.

The principles that are used to extract information, process it, and apply the appropriate correction in a controlled manner make up the content of this chapter.

The animal visual system is an example of adaptive optics. The eye is capable of adapting to various conditions to improve its imaging quality. The active focus "system" of the eye-brain combination is a perfect example. The brain interprets an image, determines the correction necessary, and, either voluntarily or involuntarily, applies that correction through biomechanical movement of the lens of the eye. This is both *closed-loop* and *phase-only* correction. The iris can open or close in response to light levels, which demonstrates adaptive optics in an intensity control mode. And the muscles around the eye can force a "squint" that, as an aperture stop, is an effective spatial filter and phase-controlling mechanism.

The realization of adaptive optics by hardware systems can be classified into four generic types, as shown[1] in Fig. 2.1. This divides the systems into either beam propagation or imaging systems and into received or outgoing wave systems. The transmitted wavefront systems use a laser to irradiate a target and sense the return from a glint on the target to interrogate the
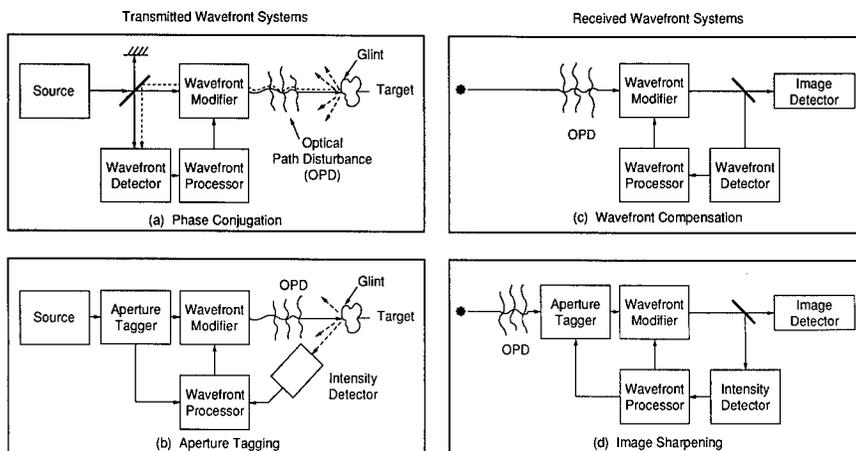
Fig. 2.1   Four generic types of adaptive optics systems.

intervening atmosphere for phase aberration information. The phase conjugation system can also employ an independent source, or beacon, to do this task. In the *phase conjugation*[2] approach, the return signal is combined with the outgoing signal on the wavefront modifier, a deformable mirror for example, and sent to the wavefront sensor, which determines the differences between the two waves. The processor then extracts the beacon phase and applies the conjugate to the outgoing wave, which then arrives at the target with the atmospheric distortion ideally exactly compensated.

The *aperture tagging*[3] system applies spatial wavefront modifications at distinct temporal frequencies. These are deconvolved from the return signal and are modified on a trial basis so as to drive a hill-climbing servo mechanism to maximize the power density in the target glint.

The two systems on the right of Fig. 2.1 are appropriate for imaging systems such as astronomical observatories or ground-based satellite reconnaissance systems. The *wavefront compensation*[4] system is similar in spirit to the phase conjugate approach for the active system. The object is the source of return signal, which is passed from the telescope primary to the wavefront corrector, which, in the case of a deformable mirror, is commanded to produce a null in the combined incoming and processed wavefronts. A portion of the collected signal is passed to the wavefront sensor and processor, which produces a local map of the wavefront deviation from the ideal. The conjugate of this map is applied to the mirror so that the main incoming signal will have the atmospheric distortion effectively removed before it is viewed by the imaging system, e.g., the eyeball of the astronomer.

The final system portrayed in Fig. 2.1 is the *image sharpening* method.[5] Here the image is perturbed on a trial basis and the effect is measured by means of a criterion such as flux in a specified area at the image plane. Hill-climbing techniques are then used to alter the trials to maximize this selected criterion.

The purpose of this chapter is to review a variety of methods and implementations of those methods to achieve adaptive correction. Detailed derivations and hardware descriptions are not given. These details are available in the bibliography presented at the end of this chapter. Rather, our approach is to give sufficient detail, together with examples, to allow the reader with knowledge of classical optics to perform scoping designs of adaptive optics systems and to estimate performance envelopes.

We have chosen to use the *Strehl ratio* as a figure of merit or performance criterion for analyzing the impact of aberrations with and without correction by adaptive optics. Roughly speaking, the Strehl ratio is a number between 0 and 1 that quantifies the performance of a system in the presence of aberrations as compared with the same system when free from aberrations. The main input required to calculate the Strehl ratio is the mean-square deviation of the system wavefront from ideal, generally spherical, curvature. Thus, a great part of this chapter is analyses designed to extract from particular systems and aberration classes the mean-square wavefront deformations resulting from a wide variety of causes.

The chapter is structured as follows: The next section is a brief historical survey. The following section is devoted to a derivation of the Strehl ratio expression and some examples of its utility. Then follows a discussion of various

aberrations in beam propagation and imaging, together with useful engineering formulas to be used in subsequent sections. The next few sections present an overview of various tasks that are performed by all adaptive optics systems: signal sampling, wavefront sensing of the sample, reconstruction of phase from the sensor measurements, and correction of the outgoing signal or image. The final section provides a summary of Strehl ratio formulas building on the material presented in earlier sections.

2.2   HISTORY

Part of the lore of this subject includes[6] the purported use of adaptive optics by the Greeks in defense of the city of Syracuse during the siege by the Roman fleet in 215 B.C. It is said that no less a scientific genius than Archimedes devised a method for reflecting and focusing sunlight to start fires on the Roman vessels. Because the sources of this story all lived centuries after the alleged event, a lively debate has developed surrounding the feasibility of such a feat.[7] One suggestion has the Greek army coordinating reflections from their highly polished copper shields.

It is not our intent to enter into this debate here. We simply note that if, in fact, the Greeks were able to start fires at long distances, they could not have done it with a single mirror. This is the relevant point to the subject of adaptive optics. Archimedes would necessarily have had to devise a method of accurately pointing a number of independent reflectors to multiply the effect. Because it was not until the latter part of the twentieth century A.D. that analogous techniques were realized in practice, such as the array of 63 flat mirrors comprising the French Pyrenees solar concentrator system, Archimedes achievement is truly fantastic. In any case, the method, if used, did not prevent the Romans from ultimately sacking the city and killing Archimedes.

These methods, in both ancient and modern implementations, are examples of incoherent addition of light beams. Adaptive optics in its most useful realizations deals with coherent combining of beams or images and the controlling of phases of wavefronts.

A pioneer in the field of adaptive optics often cited as the first to apply this technology is Horace W. Babcock of the Mount Wilson and Palomar observatories. His first contribution was the design and installation of a star guider for the 100-in. Mt. Wilson telescope.[8] This device replaced a human observer whose job it was to keep the image of a star on the slits of a spectrometer. Guiding corrections are required every few seconds because of random fluctuations in the image position resulting from atmospheric turbulence and imperfections in the telescope drive. The guider works at the focus of the telescope normally used for observing the star image reflected from the back side of the spectrometer slit. A rotating knife edge chops the image at the focal plane to create a varying signal. The amplitude of the signal is a measure of the axial deviation of the image and the phase indicates the direction the telescope must be moved to return the signal to the axis. Corrections as small as a fraction of an arc second are possible, whereas a large dynamic range relieves the necessity for demanding specifications in the construction of the drive mechanism.

In 1953, Babcock[9] proposed using a deformable optical element, driven by a wavefront sensor, to compensate for atmospheric distortions that affected telescope images. This appears to be the earliest reference to use of adaptive optics as we define the field today. Babcock proposed bringing the light of a star into focus at F on the axis of the objective of the telescope (see Fig. 2.2). A lens images the objective on a mirror (off-axis parabola) that is coated with a thin layer of oil. The oil layer causes light rays to be advanced or delayed according to its thickness and bent according to the local curvature of the layer. The thickness of the oil is controlled by the application of a standard cathode ray tube electron beam in a raster scan. The light from the oil-coated mirror is brought to focus in the plane of a rotating knife-edge, K. Then the light diverges to form a Schlieren image on an image orthicon tube. As the Schlieren image is scanned by the orthicon, the resulting signal is integrated and used to modulate the intensity of the electron beam, which is scanned synchronously across the mirror with the oil layer. The corrected star image is picked off with, for example, a partially transmitting pellicle P in front of the knife edge, thence to the focal plane S for observation. A later proposal from Babcock[10] replaced the oil layer with a flexible reflecting film and charge-
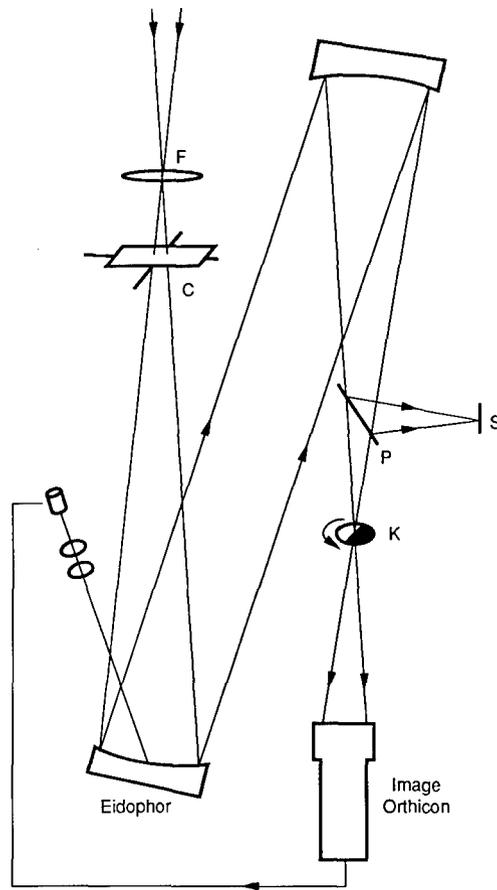
Fig. 2.2   Adaptive optics system proposed by Babcock.

collecting electrodes to locally bend the film. Neither of these methods was successfully deployed at the 200-in. telescope.

The particular approach to adaptive control depends on the mission of the system under development. An early use of adaptive optics was for maximizing the power density on target of a laser beam in a turbulent atmosphere.[2] Here glints (specular reflections) from the target were used as sources of phase shifts that were applied to the outgoing beam to maximize the power density in the vicinity of the glint. Another approach to the same problem was that of dithering (at different temporal frequencies) discrete sections of the outgoing laser with phase modulations of ± 30 deg. These signals, which could be demodulated by suitable filtering, acted as trial perturbations that produced a positive or negative effect on some figure of merit, such as power density. A hill-climbing servo maximized the intensity at a glint on the target. A schematic of a multidither experiment using an 18-element dither system is shown[3] in Fig. 2.3. This system demonstrated convergence in 2 ms.

Application to imaging problems such as occur in astronomical telescopes produce rather different requirements and implementations. For example, in some cases the star or planet itself is the source of the radiation for correction information. Because these sources emit a broad wavelength spectrum, one must compensate optical path length rather than phase differences. The first successful practical system was developed by Itek[11] for use at the Air Force Maui Optical Measurements Station (AMOS) in Mt. Haleakala in Maui, Hawaii. A schematic of the system is shown in Fig. 2.4. This was a real-time phase-only compensation system utilizing an ac white light shearing interferometer wavefront slope sensor, a resistor/op-amp analog network for phase reconstruction from slopes, and a 21-element monolithic piezoelectric deformable mirror for correction. This system was the predecessor to the compensated imaging system (CIS) successfully operated at AMOS for the past decade.

Another approach for astronomy is the image-sharpening scheme of Muller and Buffington.[5] In this method, the sharpness of the image is defined to be the integral of the square of the intensity over the image plane. They show that this function is uniquely and absolutely maximized for any undistorted
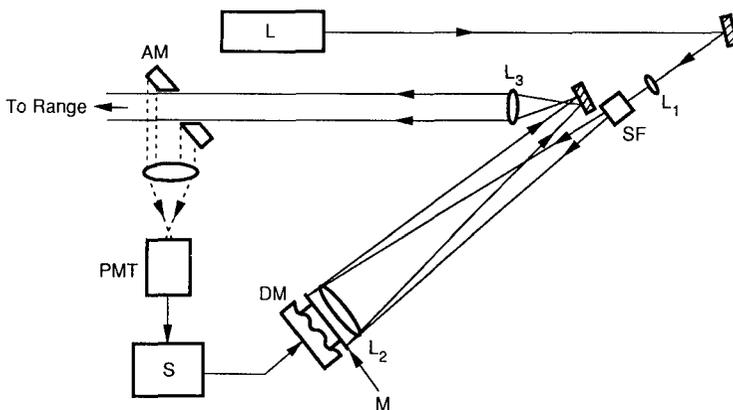
Fig. 2.3   Schematic of an 18-element multidither system.

object lying within the isoplanatic patch of the viewing system. This technique is implemented as shown in Fig. 2.5 both at the Leuschner Observatory and at the Lick Observatory. The six phase shifters are piezoelectric stacks, which are altered iteratively to maximize the sharpness as measured by photomultiplier PM1. Typically, three iterations were performed, taking about 4 ms each, to achieve satisfactory sharpness response to the six phase shifters.
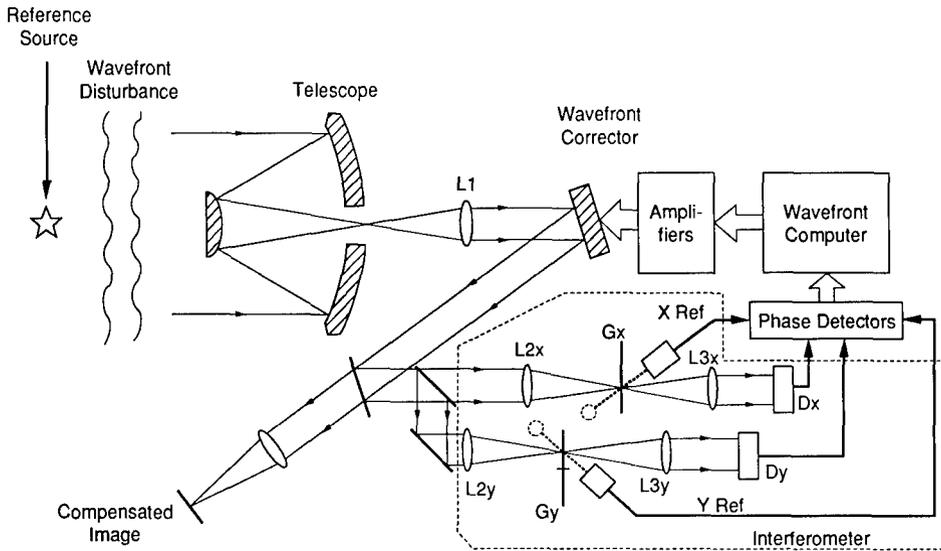
Fig. 2.4   Block diagram of a real-time atmospheric compensation (RTAC) system.
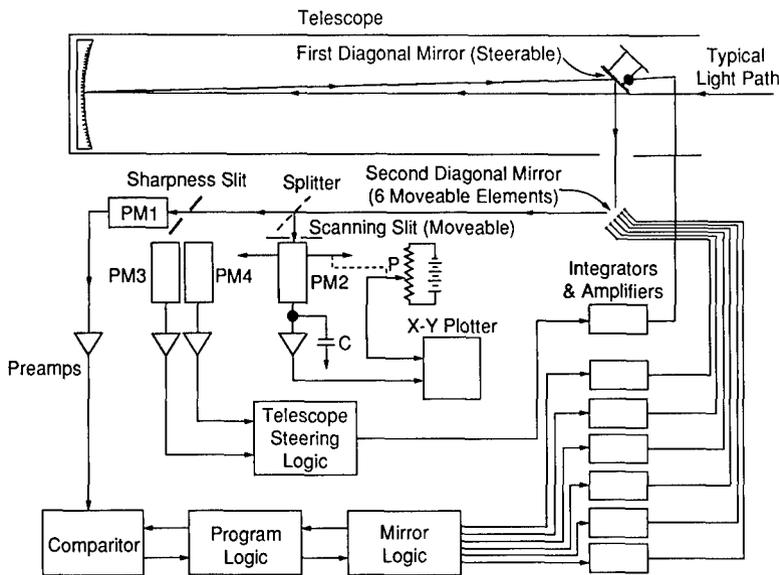
Fig. 2.5   Block diagram of an image sharpening system.

2.3   STREHL RATIO

We use the Strehl ratio (SR) as a figure of merit for performance of an optical system.[12] We begin with an analysis that relates mean-square deviation of the aberrated phase front from ideal to the ability of the optical system to maximize intensity or image quality. In this section, we derive a relation between the phase aberration of a wavefront and the intensity reduction that ensues. The ratio of this reduced on-axis intensity to the intensity that would be obtained without aberration is the SR.

The intensity in the neighborhood of an image point can be calculated when we know the waveform in the pupil plane that departs from a spherical surface centered on the image point in question.[12] Referring to Fig. 2.6, if $\Phi_P(\rho,\theta)$ is the local departure of the wavefront from spherical, then the normalized intensity is given by

$$
\begin{aligned}
i(P) &= \frac{1}{\pi^2}\left|\int_0^1\int_0^{2\pi}\exp(ik\Phi_P)\rho\,d\rho\,d\theta\right|^2 \\
&= \frac{1}{\pi^2}\left|\int_0^1\int_0^{2\pi}\left[1 + ik\Phi_P + \frac{1}{2}(ik\Phi_P)^2 + \cdots\right]\rho\,d\rho\,d\theta\right|^2,
\end{aligned}
\tag{2.1}
$$

where $k = 2\pi/\lambda$. Assuming that the aberrations are small, we can neglect higher powers of $k\Phi_P$ and write

$$
\begin{aligned}
i(P) &= \left|1 + ik\langle\Phi_P\rangle - \frac{1}{2}k^2\langle\Phi_P^2\rangle\right| \\
&= 1 - \left(\frac{2\pi}{\lambda}\right)^2[\langle\Phi_P^2\rangle - (\langle\Phi_P\rangle)^2],
\end{aligned}
\tag{2.2}
$$

where $\langle\Phi_P^n\rangle$ is the average of the n'th power of Φ_P,
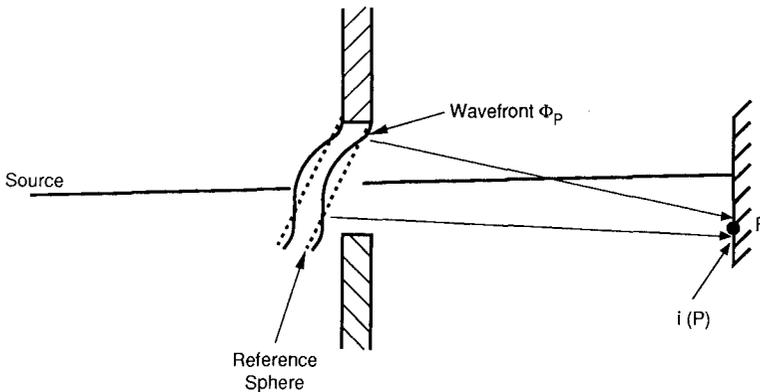
Fig. 2.6   Wave propagation geometry for the Strehl ratio derivation.

$$\langle \Phi_P^n \rangle \; = \; \frac{1}{\pi} \int_0^1 \int_0^{2\pi} \Phi_P^n \, \rho \, d\rho \, d\theta \; . \tag{2.3}$$

The quantity in the right-hand brackets of Eq. (2.2) is the mean-square deformation of the wavefront, $(\Delta\Phi)^2$, so that Eq. (2.2) can be written as

$$i(P) \; = \; 1 \; - \; \left(\frac{2\pi}{\lambda}\right)^2 (\Delta\Phi_P)^2 \; . \tag{2.4}$$

This is the Strehl ratio expression we have been seeking. Because $\Delta\Phi$ was assumed small in carrying through this derivation, we can write Eq. (2.4) as

$$i(P) \; = \; \exp\left[-\left(\frac{2\pi}{\lambda}\right)^2 (\Delta\Phi_P)^2 \right] \; . \tag{2.5}$$

This extremely valuable formula, which forms the basis of our estimates of various aberrations and their correctability by adaptive optics, tells us that the intensity in the focus or the far field is reduced from its ideal, diffraction-limited value by an amount proportional to the mean-square deformation of the wavefront.

2.4 SOURCES OF ABERRATION

Poor image or beam quality drive one to consider adaptive optics for improvements of system performance. In this section, we review the sources of aberration with which an adaptive optics system must deal. They include linear effects resulting from turbulence, optical manufacturing, and misalignments as well as errors that result from nonlinear thermal effects and fluid properties. The minimization of these effects is always a consideration in developing any optical system from the ground up. The correction of the effects of these disturbances is the realm of adaptive optics.

2.4.1 Atmospheric Turbulence

The atmosphere is a dynamic, viscous fluid in which macroscopic forces lead to microscopic fluctuations of fluid velocity, temperature, and density. Because the index of refraction is directly related to the air density, it fluctuates in time and from place to place. Thus, a light signal passing through the atmosphere has its properties randomly altered. These small changes in the index of refraction act like small lenses in the atmosphere. They focus and redirect waves and eventually, through interference, cause intensity variations. The most common effect of turbulence is manifested in the twinkling and quivering of stars. Twinkling is the random intensity variation of the light from a star because of the random interference between waves from the same star passing through slightly different atmospheric paths. The average position of the star also shows a random quiver because the average angle of arrival of light from the star is affected by the changing index of refraction along its path through the atmosphere. A third effect of turbulence, known long ago by astronomers, is the apparent spreading of the star image due to turbulence.

These three primary effects of turbulence are examined here because they are subject to correction by adaptive optics. Other atmospheric effects, such as aerosol scattering and molecular absorption, are not subject to control and simply must be accepted as unrecoverable losses in overall system specification and design.

Description of Atmospheric Turbulence. Because the physical phenomena relating to turbulence are random, we use a statistical description. A key quantity that provides a measure of the effect of the random medium is the *structure function,* which is defined as the ensemble average of the square of the difference between measurements made of a fluid quantity, for example, velocity, at two different points in the atmosphere. Consider the result of many such measurements of velocity with sensor probes separated by a distance $|\mathbf{r}|$. We write the ensemble average of the data taken as

$$D_v = \langle [v_r(\mathbf{r}_1 + \mathbf{r}) - v_r(\mathbf{r}_1)]^2 \rangle \ . \tag{2.6}$$

If the separation r is small (within the inertial subrange of turbulence), the structure function takes on a ⅔ law dependence[13] on r:

$$D_v = C_v^2 r^{2/3} \ . \tag{2.7}$$

The constant C_v^2 is the velocity structure constant, which is a measure of the energy in the turbulence. This form of the structure function is valid when the value of r is above the smallest eddy size l_0 and below the largest eddy size L_0. The small eddy, called the *inner scale* or the *microscale,* is the size below which viscous effects are important. The large eddy, called the *outer scale,* is the size above which isotropic behavior is violated. Near the ground, l_0 is a few millimeters and L_0 is a few meters. At higher altitudes, the outer scale can exceed hundreds of meters.

Similarly, if one performs the experiment with temperature probes and then relates the temperature to the density, one arrives at an equivalent statistical descriptor of the index of refraction structure. This quantity D_n is germane when we deal with optical propagation problems.

$$D_n(r) = C_n^2 r^{2/3} \ , \quad l_0 \ll r \ll L_0 \ , \tag{2.8}$$

where C_n^2 is the refractive index structure constant, a measure of the strength of turbulence.

We now turn to descriptors of the turbulent index in the spatial frequency domain. This is important because the distance between wavefront sensor samples and corrector mirror actuators depends critically on the spatial frequency content of the incoming wavefront.

The atmosphere can be assumed to be made up of a mean index of refraction $\langle n(\mathbf{r}) \rangle$ and a fluctuating index part $n_1(\mathbf{r})$. The covariance of the refractive index field B_n is a measure of the correlation of the field as probes are separated, and is defined to be

$$B_n = \langle n_1(\mathbf{r} + \mathbf{r}_1) n_1(\mathbf{r}_1) \rangle \ . \tag{2.9}$$

The power spectral density (PSD) is the Fourier transform of the covariance,

$$\Phi_n(\mathbf{K}) = \frac{1}{(2\pi)^3} \int d^3\mathbf{r} B_n(\mathbf{r}) \exp(-i\mathbf{K} \cdot \mathbf{r}) \; , \tag{2.10}$$

where $\mathbf{K}$ is the three-dimensional spatial wave number. Using Kolmogorov's inertial subrange expression, Eq. (2.8), and carrying out the ensemble averaging, one has for the PSD,

$$\Phi_n(K) = \frac{5}{18\pi} C_n^2 K^{-3} \int_{l_0}^{L_0} dr \, \sin(Kr) r^{-1/3} \; . \tag{2.11}$$

In the limits $l_0 \to 0$ and $L_0 \to \infty$, the integral reduces to the *Kolmogorov spectrum*,

$$\Phi_n(K) = 0.033 C_n^2 K^{-11/3} \; . \tag{2.12}$$

Tatarskii[14] proposed an *ad hoc* spectrum for finite inner scales.

$$\Phi_n(K) = 0.033 C_n^2 K^{-11/3} \exp\left[\frac{-K^2}{(5.92/l_0)^2}\right] \; . \tag{2.13}$$

A form of the spectrum for cases with finite outer scales is the *Von Karman* spectrum[15]:

$$\Phi_n(K) = 2.54 \times 10^{-4} \delta^2 L_0^3 (1 + K^2/K_0^2)^{11/6} \; , \tag{2.14}$$

where $K_0 = 2\pi/L_0$ and the variance of the refractivity fluctuations is

$$\delta^2 = \frac{C_n^2}{1.9 K_0^{2/3}} \; . \tag{2.15}$$

Each of these spectra is valid only for regions in the inertial subrange

$$\frac{2\pi}{L_0} \ll K \ll \frac{2\pi}{l_0} \; . \tag{2.16}$$

Refractive Index Structure Constant.   The refractive index structure constant C_n^2 is a measure of the strength of turbulence. In general, it varies along any path in the atmosphere, with the seasons as well as daily and hourly. It varies with geographic location and it varies with altitude. It can be perturbed very easily by artificial means, such as aircraft, or the mere presence of a structure, such as a building or a telescope dome that interrupt the laminar flow of air and create non-Kolomogorov turbulence. For the purposes of specifying adaptive optics systems, however, one usually resorts to models of the dependence of C_n^2 on path length or altitude above the astronomical site. These models are curve fits to data taken, for example, by balloon-borne instruments that measure temperature or wind fluctuations.[16] Some of these models include standard

deviations as well as mean values of C_n^2. Two popular models, which represent sites of interest, are the so-called SLC-Day/Night models and the Hufnagel-Valley boundary (HVB) model. The former is applicable to mountain top astronomical sites and was created from data taken at the observatory on Mt. Haleakala, Maui, Hawaii. The SLC stands for Submarine Laser Communication, the funding program. Two curve fits are used to distinguish data taken at day and at night. These fits are as follows:

SLC-Daytime Model. Height H is given in meters above ground (3.1 km), C_n^2 in m$^{-2/3}$:

$$
\begin{aligned}
&H < 18.5 && 1.70 \times 10^{-14} \ , \\
&18.5 < H < 240 && 3.13 \times 10^{-13}(H/1000)^{-1} \ , \\
&240 < H < 880 && 1.30 \times 10^{-15} \ , \\
&880 < H < 7000 && 8.87 \times 10^{-7}(H/1000)^{-3} \ , \\
&7000 < H < 20{,}500 && 2.00 \times 10^{-16}(H/1000)^{-1/2} \ , \\
&H > 20{,}500 && 0 \ .
\end{aligned}
\tag{2.17}
$$

SLC = Nighttime Model. Height H is given in meters above ground (3.1 km), C_n^2 in m$^{-2/3}$:

$$
\begin{aligned}
&H < 18.5 && 8.4 \times 10^{-15} \ , \\
&18.5 < H < 110 && 2.87 \times 10^{-12}(H/1000)^{-2} \ , \\
&110 < H < 1500 && 2.50 \times 10^{-16} \ , \\
&1500 < H < 7000 && 8.87 \times 10^{-7}(H/1000)^{-3} \ , \\
&7000 < H < 20{,}500 && 2.00 \times 10^{-16}(H/1000)^{-1/2} \ , \\
&H > 20{,}500 && 0 \ .
\end{aligned}
\tag{2.18}
$$

These distributions of turbulence are shown in Fig. 2.7.

The HVB model is based on measurements made primarily by Bufton[17] at the Goddard Spaceflight Center in Maryland. These data were fit by Hufnagel[16] and later altered by Valley,[18] who added an upper air jet-stream contribution, and by Ulrich,[19] who added a boundary layer or sea-level contribution. This model is as follows:

$$
C_n^2 = A\left[2.2 \times 10^{-23} z^{10} e^{-z} \left(\frac{V}{V_0}\right)^2 \right.
$$

$$
\left. + 10^{-16} \exp(-z/1.5) + 1.7 \times 10^{-14} \exp(-z/0.1) \right] \ ,
\tag{2.19}
$$

with the height z in kilometers and C_n^2 in units of m$^{-2/3}$. The distribution is shown in Fig. 2.8. Because this expression contains two adjustable constants, V/V_0 and A, one can use Eq. (2.19) to represent a turbulence distribution with
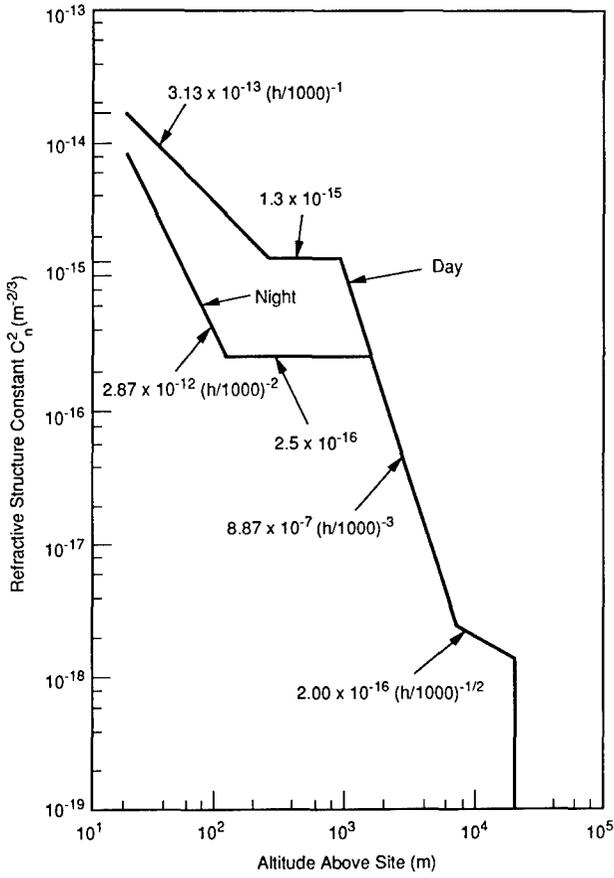
Fig. 2.7   Curve fits to mean value of turbulence structure constant above a mountain site.
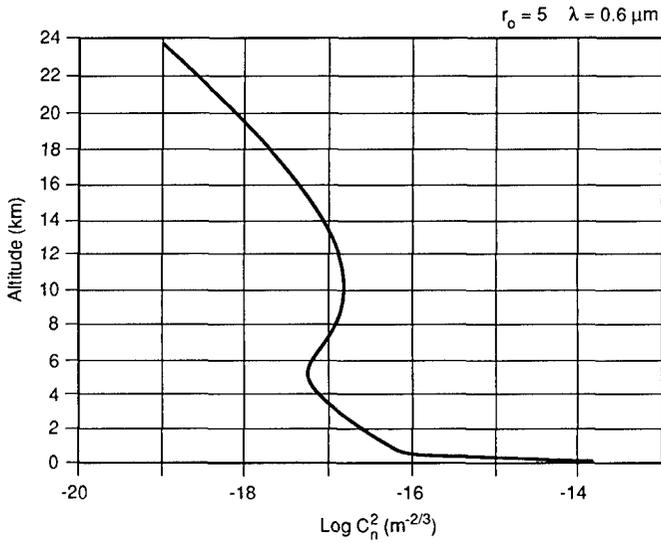
Fig. 2.8   Structure constant for the HVB model.

specified values[19] of transverse coherence length r_0 and isoplanatic angle θ_0. For instance, a 5-cm coherence length and 7-μrad isoplanatic angle is generated by setting $A = 1.03$ and $V/V_0 = 1.8$. This results in the Hufnagel-Valley "5/7" model.

2.4.2 Propagation Effects

Intensity Variation. An important quantity affecting the design of an adaptive optics system is the amount of intensity fluctuations expected to be incident on the wavefront sensor of a phase-only system. Intensity variations are expressed as the fluctuation of the log of the amplitude (*log-amplitude fluctuations*). The Kolmogorov spectrum suggests that log-amplitude fluctuations are produced by eddies on the order of $\sqrt{\lambda L}$ for light beams of wavelength λ at a distance L. Wander is produced by eddies that are larger than the beam size. Beam spread is produced by eddies that are smaller than the beam size. Assuming the Von Karman spectrum Eq. (2.14), and inner scale $l_0 \ll \sqrt{\lambda L}$, the log amplitude variance σ_χ^2 is given by[14]

$$\sigma_\chi^2 = 0.307 k^{7/6} L^{11/6} C_n^2 \tag{2.20}$$

for a plane wave, and by

$$\sigma_\chi^2 = 0.124 k^{7/6} L^{11/6} C_n^2 \tag{2.21}$$

for a spherical wave.

These perturbative formulas break down for long paths and strong turbulence. For $\sigma_\chi^2 > 0.3$ it has been found experimentally[20] and explained theoretically[21] that the scintillations saturate. In this case, the expression for the variance of log amplitude becomes

$$\sigma_\chi^2 = [0.5454 \sigma_{\chi_0}^2/(1 + 0.2327 \sigma_{\chi_0}^2)]^2 \ , \tag{2.22}$$

where $\sigma_{\chi_0}^2$ is the expression given in Eq. (2.20) or Eq. (2.21) for log-amplitude variance in the absence of scintillation.

Phase Variation. Phase variations arise because the refractive index is a random function of space and time. The random index variations produce random velocity variation in the propagating wave. The phase difference $\delta\phi$ between two segments of atmosphere separated by ρ is $\delta\phi \sim k\rho\Delta n(\rho)$, where the difference in index is Δn. By averaging over many segments and a path length L, the *phase structure function* can be determined:

$$D_\phi = 1.46 k^2 C_n^2 L \rho^{5/3} \ , \quad l_0 \ll \rho \ll L_0 \ . \tag{2.23}$$

When the averaging must be done over long paths and many eddies, the wave can lose coherence. If the wavefront is aberrated before it encounters an eddy, the plane wave approximations used thus far are no longer valid.

Beam Spread.   Lutomirski and Yura[22] used the extended Huygens-Fresnel principle to show that the angular spread θ of a collimated uniform beam of radius a in turbulence is

$$\theta^2 \approx \frac{1}{k^2 a^2} + \frac{1}{k^2 \rho_0^2} \ , \tag{2.24}$$

where $\rho_0 = (1.46 k^2 C_n^2 L)^{-3/5}$. A Gaussian beam, described by

$$I = \frac{w_0^2}{w_b^2} \exp\left(-\frac{2\rho^2}{w_b^2} \right) \ , \tag{2.25}$$

spreads from turbulence. For short propagation distances ($L \ll \pi w_0^2/\lambda$) the beam waist w_b increases from its initial size w_0 according to

$$w_b^2 = w_0^2 + 2.86 C_n^2 k^{1/3} L^{8/3} w_0^{1/3} \ . \tag{2.26}$$

For long distances ($L \gg \pi w_0^2/\lambda$), it increases according to

$$w_b^2 = \frac{4z^2}{k^2 w_0^2} + 3.58 C_n^2 L^3 w_0^{-1/3} \ . \tag{2.27}$$

From these formulas, we can write down the Strehl ratio as

$$\mathrm{SR}_{\mathrm{BS}} = \frac{w_0^2}{w_b^2} \ , \tag{2.28}$$

where w_b^2 is given by Eqs. (2.26) or (2.27).

Angle of Arrival.   When two beams, or two parts of the same beam, propagate a distance L through turbulence, they arrive at the same point through a slightly different path. The variance of their angle of arrival difference $\langle \alpha^2 \rangle$ is given by the expression[23]

$$\langle \alpha^2 \rangle = \frac{D_\phi}{k^2 \rho^2} \ , \tag{2.29}$$

where D_ϕ is the phase structure function [Eq. (2.23)]. For an aperture diameter D, the angle of arrival variance becomes

$$\langle \alpha^2 \rangle = 1.46 C_n^2 L D^{-1/3} \ , \tag{2.30}$$

and

$$\mathrm{SR} = \exp(-\langle \alpha^2 \rangle) \ . \tag{2.31}$$

Coherence Length.   In a study of an optical heterodyne communications receiver, Fried[24] found that the maximum diameter of a collector that is allowed

before atmospheric distortion seriously limits performance is r_0, where the *coherence length* or the *seeing cell size*[25] is

$$r_0 = \left[0.423k^2 \sec(\theta_Z) \int_0^L C_n^2(h) \, dh \right]^{-3/5} , \qquad (2.32)$$

L is the path length, θ_Z is the zenith angle, and C_n^2 can vary with altitude h (see Fig. 2.9). For plane waves and Kolmogorov turbulence, the phase structure function in terms of this parameter is[26]

$$D_\phi = 6.88 \left(\frac{r}{r_0} \right)^{5/3} . \qquad (2.33)$$

The radial variable r is normal to the propagation direction. For spherical waves the coherence length is slightly modified

$$(r_0)_{\text{sph}} = \left[0.423k^2 \sec(\beta) \int_0^L C_n^2(h) \left(\frac{h}{L} \right)^{5/3} dh \right]^{-3/5} . \qquad (2.34)$$

For constant C_n^2 (i.e., for a horizontal path), the coherence length reduces to (see Fig. 2.10):

$$r_0 = 1.68(C_n^2 L k^2)^{-3/5} . \qquad (2.35)$$
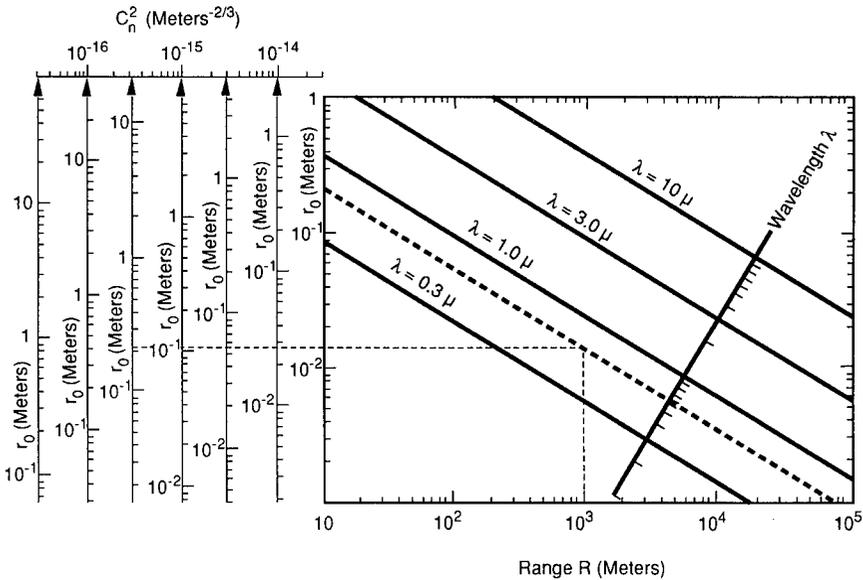
Fig. 2.9 Coherence length r_0 as a function of the path length L for various wavelengths.
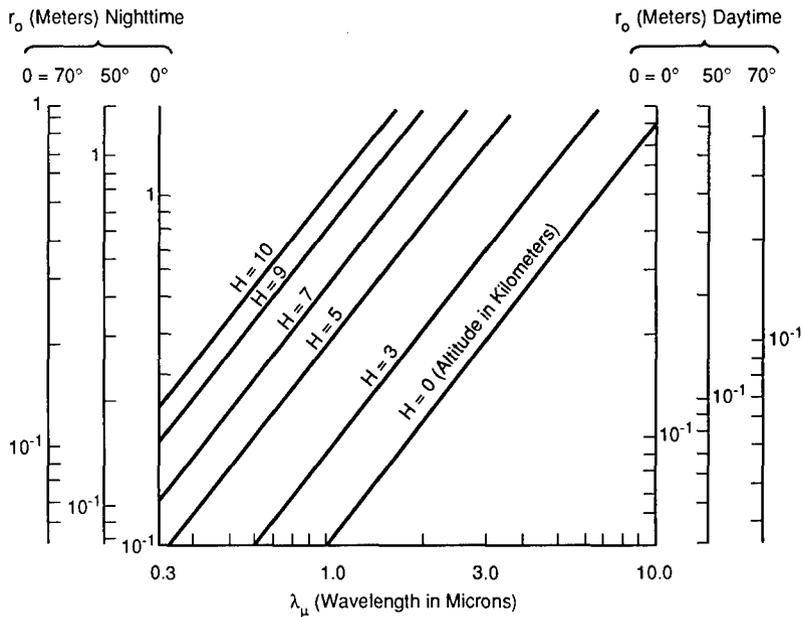
Fig. 2.10 Coherence length r_0 as a function of wavelength of various altitudes.

Fried's coherence length has been used to describe various atmospheric turbulence phenomena. Because it incorporates the turbulence strength C_n^2, the wavelength, and the propagation path into one parameter, it is widely used for scaling between laboratory and field experiments. For instance, the spatial frequency spectrum of turbulence,[27]

$$\Phi(K) = (0.023/r_0^{5/3})K^{-11/3} \ , \tag{2.36}$$

can be used to describe all the aberration modes of atmospheric turbulence in terms of Zernike and other polynomials. Also, the spectrum can be integrated to give an overall wavefront variance

$$\sigma^2 = \int \Phi(K) \, d^2K \ . \tag{2.37}$$

The variance of the wavefront tilt over an aperture of diameter D is given by

$$\sigma_{\text{tilt}}^2 = 0.896(D/r_0)^{5/3} \ , \tag{2.38}$$

or in terms of the angle measured from the aperture,[28]

$$\sigma_{\text{ang tilt}}^2 = 0.364(D/r_0)^{5/3}\left(\frac{\lambda}{D}\right)^2 \ , \tag{2.39}$$

$$SR_{tilt} = \exp(-\sigma_{tilt}^2) \ . \tag{2.40}$$

The turbulence MTF is

$$M_{turb} = 1, z \ll (0.4k^2C_n^2L_0^{5/3})^{-1} \ ,$$

$$M_{turb} = \exp[-(r/r_0)^{5/3}] \ , \quad z \gg (0.4k^2C_n^2L_0^{5/3})^{-1} \ . \tag{2.41}$$

Time Dependence of Turbulence. Turbulence is a random process that has temporal statistics as well as spatial statistics. From an adaptive optics point of view, these statistics can be important. If turbulence moves very slowly, then an adaptive optics system can view the disturbance as static. Correction would only be required in the spatial domain. If the turbulence was very fast, an adaptive optics system would see only the time-averaged disturbance. Turbulence is driven by winds that alter the atmosphere through which the beam or image is passing in times like a fraction of a second. This dictates that control systems operating at a few hundred hertz will generally be capable of keeping up with the temporal evolution of the aberrations. Such systems are routinely designed and developed. A detailed discussion of the effect of the adaptive optics system servo lag is given in Sec. 2.7.

2.4.3 Thermal Blooming

Thermal blooming occurs when laser light is absorbed by the gases and particles that make up the atmosphere. Because energy is conserved, two things happen: (1) the laser loses energy, i.e., its intensity diminishes, and (2) the atmosphere gains energy, i.e., it heats up. When the atmosphere heats up, its density decreases. Thus, wherever the laser goes, it reduces the density of the air as shown in Fig. 2.11. This means that the laser light that passes through the air at a later time interacts with a different atmosphere than the light that passed by earlier. Whereas the early light encountered an essentially uniform atmosphere, the later light sees an atmosphere with a dip in the density in the region through which the earlier light passed (see Fig. 2.12). In effect, it is passing through a concave lens, as shown in Fig. 2.13.

The shape of the lens created by the laser depends on the shape of the irradiance of the beam. For example, a Gaussian laser beam imposes a Gaussian density depression in the atmosphere. The laser beam forms an extended
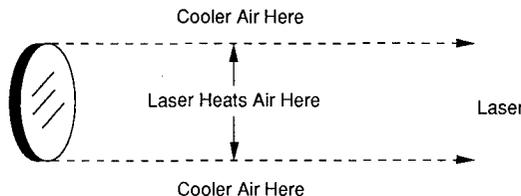
Fig. 2.11 Laser is absorbed by the atmosphere and heats it.
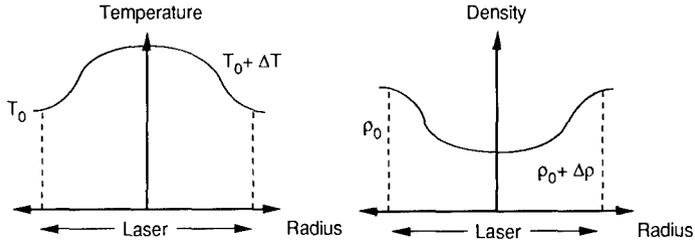
Fig. 2.12 Changes induced in the atmosphere by laser beam heating.
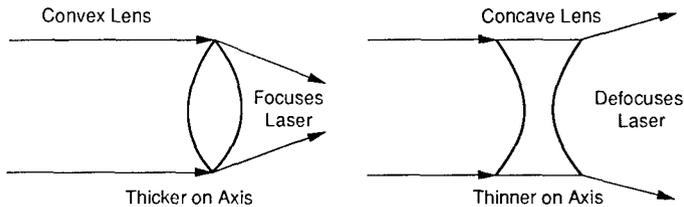
Fig. 2.13 Laser beam reduces the density of the atmosphere through which it passes relative to the density outside the beam. Hence, it creates a negative lens leading to beam divergence.

negative lens that deflects the succeeding rays out of the beam, away from the axis, causing the beam to appear to grow in radius or "bloom." Because the cause of this effect is heating of the atmosphere, this phenomenon is termed *thermal blooming*.

When the air moves relative to the beam, either because of wind or beam slewing, the air that is heating up by laser absorption remains in the beam for only a finite time. The blooming effect is limited by the amount of energy that can be absorbed in the time it takes for an air parcel to cross the beam (see Fig. 2.14). In addition, the light rays no longer bloom symmetrically outward equally in all directions because the addition of a wind-plus-slew has destroyed the circular symmetry. The blooming pattern, which, without wind or slew, is an ever growing symmetrical circle of light, becomes unsymmetrical and distorted. The rays bend into the wind, bringing cooler air into the beam. This causes the center of the beam to shift upwind. Meanwhile, the beam continues to bloom outward in the direction perpendicular to the wind. The net result is that the beam forms a crescent shape centered in an upwind position. An example from a wave optics simulation is shown in Fig. 2.15.

This distortion of the irradiance distribution of the laser beam is called *whole beam blooming* to distinguish it from the phenomenon of small-scale instability to be discussed later. An important effect of whole beam blooming is the ultimate decrease in power density that occurs as the power of the device is raised. Initially, the irradiance increases with power, but, at some point, the nonlinear heating produces an areal growth rate that exceeds the incremental effect of power increase, and the irradiance begins to drop. The curve of irradiance versus power has come to be called a *power optimization curve*.
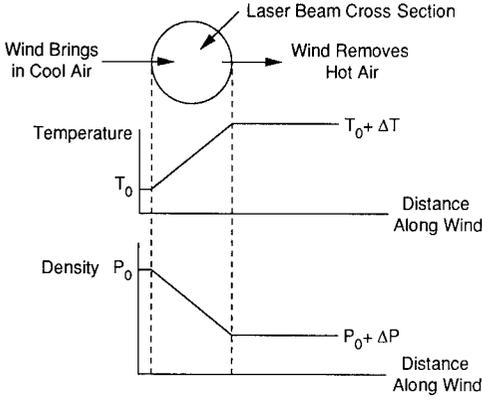
Fig. 2.14 Amount of temperature rise or density drop is limited by the residence time of the air as it passes through the laser beam because of a combination of natural wind and beam slewing.
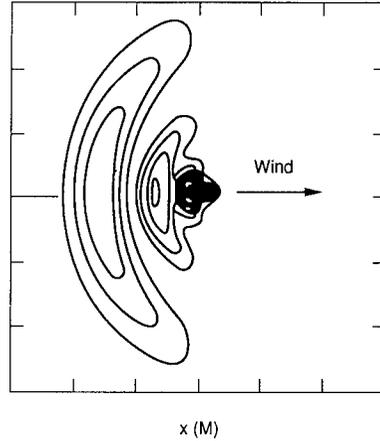
Fig. 2.15 Wave optics computer model result of an irradiance pattern showing the characteristic crescent shape resulting from bending of rays into the wind and into cooler regions transverse to the wind.

An example is given in Fig. 2.16. Gebhardt[29] derived a convenient analytical fit to the power optimization curve. The fit is expressed in terms of a *distortion number N,* which is a dimensionless scaling number proportional to the amount of distortion of the beam compared to vacuum propagation:

$$N = N_0 f(\alpha z) q(a_0/a) s(\omega z/v_0) \; , \tag{2.42}$$

where ω is the slew rate, a is the undistorted beam radius at focus, and N_0 is the distortion number for a collimated beam of radius a_0. Assuming a uniform wind of velocity v_0, in the weak attenuation limit,[30] $\alpha z \ll 1$,

$$N_0 = \frac{-(\partial n/\partial T)\alpha_{\text{abs}} P z^2}{\pi n_0 \rho_0 c_P v_0 a_0^3} \; , \tag{2.43}$$

where

$$n_0 \quad = \text{refractive index}$$

$$\frac{\partial n}{\partial T} = \text{index change with temperature}$$

$$\alpha_{\text{abs}} = \text{absorption coefficient}$$
$$P \quad = \text{power}$$
$$z \quad = \text{range}$$
$$\rho_0 \quad = \text{density}$$
$$c_P \quad = \text{specific heat at constant pressure.}$$

The index of refraction of air is given by[31,32]
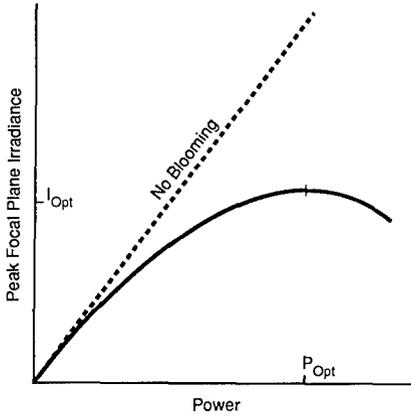
Fig. 2.16 Power optimization curve. The maximum irradiance I_{opt} occurs at the optimum power P_{opt}.
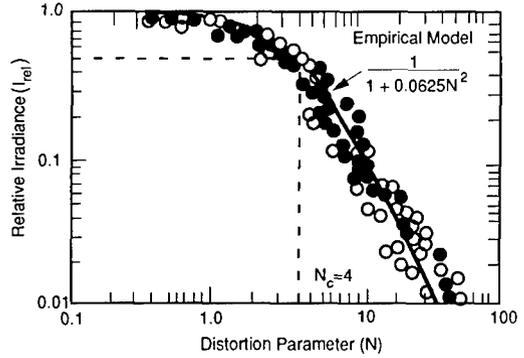
Fig. 2.17 Log-log plot of I_{REL} versus N.

$$n_0(T) = 1 + \left(8342.1 + \frac{2{,}406{,}030}{130 - \lambda^{-2}} + \frac{15{,}996}{38.9 - \lambda^{-2}}\right)$$
$$\times \left(\frac{1.0549 \times 10^{-8}}{1 + 0.00366T}\right) , \tag{2.44}$$

with the wavelength λ expressed in micrometers and the temperature T in degrees Celsius. This leads to

$$\frac{dn}{dT} = [n_0(15°C) - 1]\left(\frac{-0.004073}{1 + 0.00732T + 1.34 \times 10^{-5}T^2}\right) . \tag{2.45}$$

The three functions f, q, and s are correction factors to the collimated, weakly absorbed, and nonslewed distortion number N_0 and are given by

$$f(N_A) = (2/N_A^2)[N_A - 1 + \exp(-N_A)] ,$$

$$q(x) = [2x^2/(x - 1)]\left(1 - \frac{\ln x}{x - 1}\right) , \tag{2.46}$$

$$s(N_\omega) = (2/N_\omega^2)[(N_\omega + 1 \ln(N_\omega + 1) - N_\omega] ,$$

where, $N_A = \alpha z$, $x = a/a_0$, a is the beam radius at range z, and $N_\omega = \omega z/v_0$.

Using the expression for distortion number for whole beam blooming, one can test its utility by comparing predictions of Strehl ratio with experimental data and wave optics code simulations. A comparison of this type is shown[29] in Fig. 2.17. Here Gebhardt plots I_{REL} versus the distortion number. This form is equivalent to the power optimization curve of Fig. 2.16 except that now the irradiance is normalized by the zero-power irradiance and the curve is given

in a log-log plot format. Note that the point on the curve with slope -1 defines the critical distortion parameter N_c, which is analogous to the optimal power P_0 previously discussed. At this point, it is easy to show that $I_{\text{REL}} = 0.5$ and that $N_c = 4$.

Thus, for whole beam blooming, uncorrected by adaptive optics, we can use the following expression for Strehl ratio[33]:

$$S_{\text{TB}} = \frac{1}{1 + 0.0625N^2} , \qquad (2.47)$$

where N is given by Eq. (2.42).

Uncorrected Blooming, Ground to Space. We now turn to a discussion of blooming of a beam directed from ground to space.[34] The phase shift induced by the laser beam due to atmospheric heating is

$$\Delta\phi = \frac{2\pi}{\lambda} \int_0^H dh\,\Delta n , \qquad (2.48)$$

where λ is the laser wavelength, H is the altitude at which the phase shift is calculated, and Δn is the change in index induced by the laser. This change is given, for the case of a transverse wind v in the x direction, by

$$\Delta n = \frac{\partial n/\partial T\;\alpha\,\exp(-N_A)}{\rho_0 C_P v} \int_{-\infty}^x I(\xi)\,d\xi , \qquad (2.49)$$

where $I(\xi)$ is the beam irradiance. Consider the case of a flat-top beam for which the integral along the wind can be performed to give

$$\Delta\phi = -\frac{N_D}{2\sqrt{2}\pi}\left[\frac{x}{D/2} + \left(1 - \frac{4y^2}{D^2}\right)^{1/2}\right] , \qquad (2.50)$$

where D is the beam diameter, y is the dimension transverse to the wind direction, and N_D is the cumulative distortion number appropriate for situations where atmospheric parameters are varying along the path, and is given by

$$N_D(H,\theta_z) = \frac{4\sqrt{2}(-\partial n/\partial T)}{\rho C_P}\;\frac{2\pi P(\sec\theta_z)}{D}\int_0^H dh\,\frac{\alpha(h)T_0 t(h)}{v(h)T(h)} , \qquad (2.51)$$

where T_0 is the temperature at $h = 0$, $T(h)$ is the temperature at altitude h, and θ_Z is the zenith angle.

For points near the axis $x,y \ll D/2$, we can expand the square root term in Eq. (2.50) to give

$$\Delta\phi = -\frac{N_D}{2\sqrt{2}\pi}\left(1 + \frac{x}{D/2} - \frac{2y^2}{D_2} - \frac{2y^4}{D^4} + \cdots\right) . \qquad (2.52)$$

The first term is piston; the second is tilt, indicating the beam will bend into the wind; and the third term is defocus. Neglecting higher order terms, we assume that the adaptive optics system ignores piston and we assume tilt is handled by a fast steering mirror. Concentrating on the focus term, an effective focal length for this negative lens is

$$F_B(H) = \frac{\pi}{\sqrt{2}} \frac{2\pi D^2}{N_D(H)} ,$$

$$\simeq \frac{vD^3}{P\alpha H} . \tag{2.53}$$

We have, at this point, the following picture of whole beam blooming without phase correction (see Fig. 2.18). The diffraction-limited beam size is given by $A_D \sim R\lambda/D$, where $R \gg F_B$ is the focus of the beam and the blooming-limited beam radius is $A_B \sim RD/2F_B$. Thus, the blooming Strehl ratio is

$$\mathrm{SR}_{TB2} = \left[1 + \left(\frac{D^2}{2\lambda F_B} \right)^2 \right]^{-1/2} . \tag{2.54}$$

The square root of the expression occurs because the cooling effect of the wind means that all of the beam growth caused by blooming is orthogonal to the wind, and the contribution to the Strehl ratio degradation is from a single dimension.

Corrected Thermal Blooming.   The preceding results are altered significantly if an adaptive optics system is used to correct for whole beam blooming. (We will discuss blooming alone for now. We return to the combined effects of turbulence and blooming later.) Consider the situation[34] in Fig. 2.19. Beacon rays that originate far away are parallel (plane wave) until they encounter the negative lens created by the laser beam. At this point, they are refracted outward and enter the aperture on the ground at an obtuse angle away from the axis. The adaptive optics system conjugates the phase. In ray optics, phase conjugation means retracing the incoming ray direction, so the outgoing corrected laser beam is focused to produce a smaller beam diameter at the position of the negative lens. This increases the strength of the lens, or equivalently, reduces its focal length F_B to a value F_{BC}. From the geometry of the problem, we find the corrected beam size at the lens D_C,

$$D_C = \frac{F_B}{F_B + Z_B} D , \tag{2.55}$$

where Z_B is the height of the blooming layer, i.e., the altitude of the negative lens. The new focal length, F_{BC}, is given by

$$F_{BC} = (D_C/D)^3 F_B . \tag{2.56}$$

Thus, the Strehl ratio for the corrected beam becomes
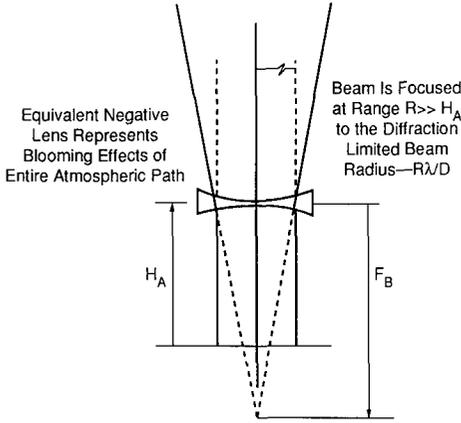
Fig. 2.18 This model of whole beam blooming replaces the extended lens induced by the laser beam with an equivalent negative lens, focal length F_B at altitude H_A.
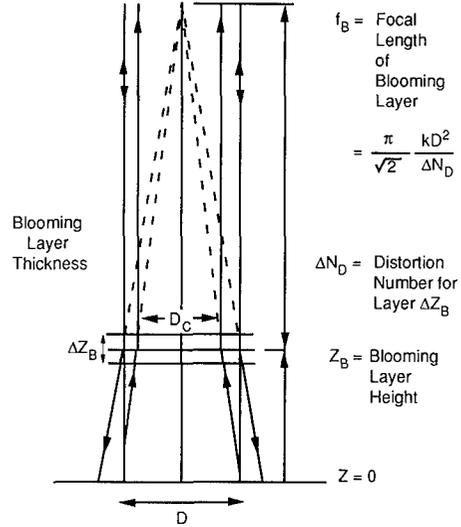
Fig. 2.19 Conjugation of rays from a beacon, considered to be infinitely far away (plane wave), produces a smaller diameter beam at the position of the negative lens, which increases the strength (decreases the focal length) of the lens.

$$\mathrm{SR}_{BC} = \frac{D_C}{D} = \left[1 + \frac{Z_B}{F_B}\left(\frac{D}{D_C}\right)^3 \right]^{-1} . \tag{2.57}$$

Note that this equation has real solutions for D_C/D only if $Z_B/F_B \leq 0.15$. This reflects the fact that if the negative lens is formed too far away (Z_B large) or is too strong (F_B small) the phase conjugate system cannot converge to a stable corrected beam. Such behavior has been predicted by wave optics simulations.[35]

Small Spatial Scale Thermal Blooming.   Recent studies of short-wavelength, large-diameter beams have revealed irradiance patterns unlike any seen before.[36] Small-scale and apparently random structure grows in magnitude with the power of the laser. When this structure becomes a majority of the beam area, the far-field irradiance drops off markedly. Whereas various explanations can be imagined for these phenomena, including (1) numerical/computational errors, (2) diffraction from truncation at sharp edges, and (3) incorrect physics or algorithms in the simulations, we are now convinced that the effects are real and caused by a special thermal blooming phenomenon.

   No new physics is needed to understand the small-scale effects. However, the changing structure of the slant path atmosphere with range does play an important role. The addition of turbulence is critical as a driver of the effects, and the use of adaptive optics can strongly enhance the phenomena. We discuss all of these separate effects in turn and show how they are related.

A Simple Model of Small-Scale Blooming.   We can reduce the problem to its barest essentials by considering some facts about the distribution of absorbers, winds, and turbulence in the atmosphere. Absorbers (molecules and aerosols) and turbulence decrease with altitude, whereas winds generally peak at around 10 km (the jet stream). We are interested in the interaction between turbulence and thermal blooming. Thermal blooming is proportional to the path length times the absorption and is inversely proportional to the wind velocity, as shown in the definition of the distortion number, Eq. (2.42). Thus, if we plot the product of absorption times range divided by wind as a function of altitude, we see that it peaks at between 1 and 2 km, depending on atmospheric conditions at the site (see Fig. 2.20). This peaking is rather pronounced and it defines the *blooming layer*. The turbulence that induces the phase changes in the high-energy laser beam as it leaves the ground is concentrated in a *turbulence layer*. As the beam progresses upward, the induced phase changes become intensity changes resulting from diffraction. When the beam reaches the blooming layer, the scintillations induced by turbulence and diffraction heat the blooming layer nonuniformly with a random small-scale pattern that is similar to the random phase pattern of low-lying turbulence.

We model the atmosphere as a low-level turbulent layer and a higher altitude absorbing layer. This simple version of the atmosphere is suggested by reference to Fig. 2.21, where we see how both the strength of turbulence C_n^2 and of thermal blooming vary as a function of altitude. The former is seen to drop dramatically in the first hundred meters, whereas the latter peaks at the blooming layer altitude between 1 and 2 km. We represent the atmosphere as consisting of a phase-inducing screen at ground level and an absorbing screen at the height of the blooming layer. Initially, we ignore any effects of the atmosphere between these layers, where we calculate only diffraction of the
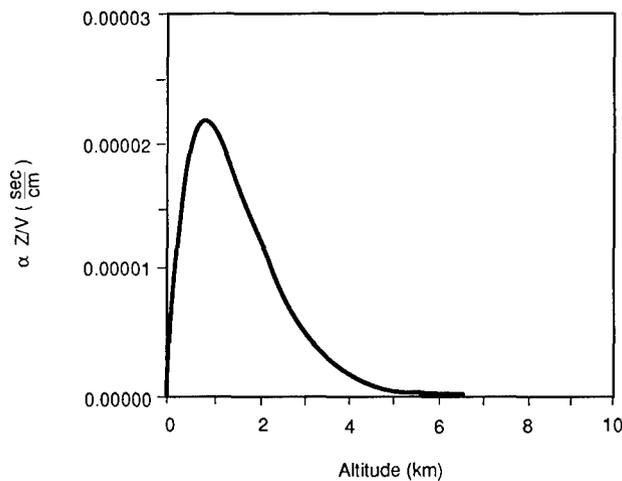
Fig. 2.20   Absorption times range divided by wind velocity. The altitude-dependent terms cause the amount of phase deviation induced by thermal blooming to peak strongly at between 1 and 2 km, depending on site conditions.
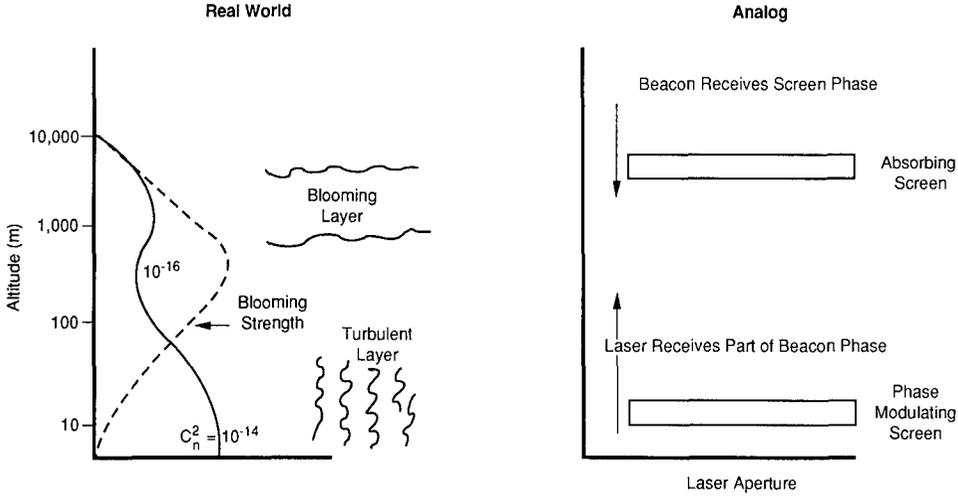
Fig. 2.21 Simplified model of the atmosphere for thermal blooming analysis.

upward traveling laser beam and the downward traveling beacon wave of the adaptive optics system.

Phase Modulating Ground-Level Screen.   Consider a plane wave that passes through a phase-modulating screen on its passage from ground to space. We first treat a single-frequency harmonic modulation, returning later to the more realistic random modulation induced by turbulence. As an example, the highest spatial frequency accessed by a deformable mirror is the inverse of twice the interactuator spacing. When the plane wave emerges from the screen its amplitude (assumed to be unity) is unchanged, but its phase now varies as

$$\phi(x) = \phi_0 + \Delta\phi \cos(2\pi x/d) \ , \tag{2.58}$$

where $\Delta\phi$ is the depth of modulation and d is the period of the phase aberration. Thus, the complex amplitude of the wave is given by

$$f(x) = \exp[i\Delta\phi \cos(2\pi x/d)] \ , \tag{2.59}$$

where we have dropped the constant phase term, $\exp(i\phi_0)$. The depth of modulation induced by the atmosphere is typically very small, so we can expand the complex amplitude, keeping only terms linear in $\Delta\phi$,

$$f(x) \simeq 1 + i\Delta\phi \cos(2\pi x/d) \ . \tag{2.60}$$

This is recognized as a unit amplitude real plane wave plus a cosine modulation in quadrature with the real part.

We now express this field after propagation to altitude $z = z_0$ in terms of its angular spectrum of plane waves.[37] Thus at $z = z_0$,

$$f(x,z_0) = \int_{-\infty}^{\infty} dS\, F(S) \exp\left\{\frac{2\pi i}{\lambda}[Sx + (1 - S^2)^{1/2}z_0]\right\} \, , \qquad (2.61)$$

where $F(S)$ is the Fourier transform of the amplitude at the phase screen,

$$F(S) = \int_{-\infty}^{\infty} dx[1 + i\Delta\phi\,\cos(2\pi x/d)] \exp\left(-\frac{2\pi i}{\lambda}Sx\right)$$

$$= \delta(S) + \frac{i\Delta\phi}{2}[\delta(S - \lambda/d) + \delta(S + \lambda/d)] \, . \qquad (2.62)$$

This expression tells us that in spatial frequency space, the modulation is given by a constant term representing the real amplitude and two traveling waves at $\pm\lambda/d$, which interfere to give the standing wave modulation in quadrature with the real part of the amplitude.

Inserting this result into the expression for the wave at altitude z_0, we have

$$f(x,z_0) = \exp\left\{\frac{2\pi i z_0}{\lambda} + i\Delta\phi\,\cos(2\pi x/d) \exp\left[\frac{2\pi i z_0}{\lambda}\left(1 - \frac{\lambda^2}{d^2}\right)^{1/2}\right]\right\} \, . \quad (2.63)$$

Now, because $\lambda^2 \ll d^2$,

$$f(x,z_0) \simeq \exp\left(\frac{2\pi i z_0}{\lambda}\right)[1 + i\Delta\phi\,\cos(2\pi x/d)\exp(-i\pi z_0\lambda/d^2)] \, . \qquad (2.64)$$

We note immediately two things about this expression. First, the overall factor $\exp(2\pi i z_0/\lambda)$ is just the plane wave optical phase and will be ignored. Second, the last exponential factor within the brackets has the property that, at multiples of the Rayleigh range, $z_0 = Nd^2/(2\lambda)$, it becomes a factor $(-i)^N$; i.e., it converts the modulation, $\Delta\phi\,\cos(2\pi x/d)$, from real to imaginary and vice versa. In other words, it converts phase to amplitude and back to phase again. This is a familiar diffractive phenomenon. It is instructive to put in some numbers at this point. Suppose $d = 10$ cm and $\lambda = 1$ µm, then the Rayleigh range is 5 km. From Fig. 2.21 we see that the height of the blooming layer is typically 1 to 2 km. Hence, substantial conversion of phase to amplitude will have occurred when the modulated wave reaches the blooming layer. This amplitude modulation induces phase changes on the adaptive optics beacon signal via the thermal blooming conversion of amplitude to phase.

Random Phase-Modulating Screen. We now study the effect of a random spectrum of phase perturbations such as those of atmospheric turbulence. Consider a random phase screen where we can analyze the phase of the plane wave emerging from the screen into a cosine Fourier spectrum of spatial components,

$$\phi(x) = \phi_0 \int_{\kappa_{min}}^{\kappa_{max}} d\kappa\,\psi(\kappa)\,\cos(2\pi\kappa x) \, , \qquad (2.65)$$

where ϕ_0 is the rms phase deviation induced by the screen (assumed to be $\ll 1$) and $\psi(\kappa)$ is a random function that, for example, might be the spatial frequency spectrum of atmospheric turbulence, for which $\kappa_{\min} = L_0^{-1}$ and $\kappa_{\max} = l_0^{-1}$. Here L_0 and l_0 are the outer and inner scales of turbulence, respectively.

Because the propagation of these phases to the blooming is approximately linear,[36] the amplitude induced in the blooming layer evolves for each spatial frequency in exactly the manner described here for the special case of a single frequency at spatial scale d^{-1}, where d is the highest scale fed back by the adaptive optics system. Thus, drawing on Eq. (2.64), we have, for the amplitude,

$$A(x,z_0) = \mathrm{Re}[\,f(x,z_0)] = 1 + \phi_0 \int_{\kappa_{\min}}^{\kappa_{\max}} d\kappa\,\psi(\kappa)\,\cos(2\pi\kappa x)\,\sin(\pi z_0 \lambda \kappa^2) \ . \quad (2.66)$$

This expression contains the spectrum induced by the random phase screen weighted at the high-spatial-frequency end of the integration by the phase-to-amplitude term, $\sin(\pi z_0 \lambda \kappa^2)$. Thus, the effect of diffraction is to convert the spectrum to amplitude differently for spatial frequencies near $\kappa_{\max}$ because the term $\sin(\pi z_0 \lambda \kappa^2)$ produces the greatest conversion of phase to amplitude for the smallest spatial scales near $\kappa_{\max}$.

Small-Scale Blooming, No Wind.   We replace the peak blooming effect at the blooming layer with a single sheet that at some altitude z_0 absorbs light of wavelength λ. This absorption is proportional to the absolute square of the complex amplitude, or, from Eq. (2.64),

$$I(x,z_0) \equiv |f(x,z_0)|^2$$

$$= 1 + 2\Delta\phi\,\cos(2\pi x/d)\,\sin\!\left(\frac{\pi z_0 \lambda}{d^2}\right) + (\Delta\phi)^2\,\cos^2(2\pi x/d) \ . \quad (2.67)$$

The nonlinear term proportional to $(\Delta\phi)^2$ produces harmonics of the basic modulation frequency, $1/d$. Because this term is proportional to the square of a quantity already assumed to be small at the outset of this analysis, $\Delta\phi$, we drop this term from consideration. Note, however, that this term is a high spatial frequency source for free-running modes that are beyond the capability of the adaptive optics system to correct. This can occur if the phase modulation is allowed to grow without limit where the $\Delta\phi$ term is no longer small. In this case, Eq. (2.59) cannot be expanded as in Eq. (2.60), but is now expressed as

$$\exp[i\Delta\phi\,\cos(2\pi x/d)] = \sum_{n=-\infty}^{\infty} i^n J_n(\Delta\phi)\,\exp(2\pi i n x/d) \ , \quad (2.68)$$

where $J_n(\cdot)$ is the Bessel function of order n. The field at altitude z_0 is

$$f(x,z_0) = J_0(\Delta\phi) + 2\sum_{n=1}^{\infty} i^n J_n(\Delta\phi)\,\cos(2\pi n x/d)\,\exp(i\pi n z \lambda/d^2) \ . \quad (2.69)$$

Harmonics of comparable intensity beat with the fundamental, giving rise to a spectrum containing appreciable energy. The remaining terms in Eq. (2.67) are a constant piston term that we can neglect and a modulation of the index of refraction at the blooming layer at the fundamental frequency, which is growing linearly with time. The phase induced on the downward traveling beacon plane wave is given by

$$\Delta\phi_B(x,z_0) = 8\left(\frac{\partial n}{\partial T}\frac{1}{\rho_0 C_p}\right)\frac{P}{\lambda D^2}\alpha(z_0)\Delta z I(x,z_0)T \ , \tag{2.70}$$

where n is the atmospheric index of refraction, T is absolute temperature, ρ_0 is the density, C_p is the isobaric specific heat, P is the laser power, λ is the wavelength, $\alpha(z_0)$ is the atmospheric absorption of the layer at altitude z_0, Δz is the layer "thickness," and, because we retain only the second term in Eq. (2.67), we have

$$I(x,z_0) = 2\Delta\phi \cos(2\pi x/d) \sin(\pi z_0\lambda/d^2) \ . \tag{2.71}$$

Thus, we see that the beacon phase is modulated at the frequency that was imposed by the ground layer on the upward traveling laser beam.

To determine the beacon wave at ground level, we repeat this calculation, but now in the downward direction. Neglecting the plane wave optical phase, this gives

$$f_B(x,z_0) = 1 - \Delta\phi_B \sin\left(\frac{\pi z_0\lambda}{d^2}\right) - i\Delta\phi_B \cos\left(\frac{\pi z_0\lambda}{d^2}\right) \ . \tag{2.72}$$

Thus, the beacon wave contains a temporally growing amplitude giving rise to scintillations that are not sensed by a phase-only adaptive optics system. Note that the conversion to amplitude proceeds, initially, linearly in range, whereas the phase is lost quadratically. Thus, significant scintillations can occur even with relatively small phase changes.

Phase Conjugate Instability.   A phase conjugate adaptive optics system is composed of two major features: (1) a beacon beam that passes earthward, picking up phase information about the atmosphere, and (2) a system for conjugating the beacon phase and adding it to the phase of the outgoing beam (see Fig. 2.22). When the beacon light passes down through the atmosphere it senses all of the ambient turbulence that is present whether the laser is operating or not. In addition, it also senses any phase changes induced by the laser beam in the blooming layer. Scintillation-induced density variations are sensed by the beacon. The adaptive optics system gathers and uses this information.

First, consider a localized hot spot in the blooming layer. This is a reduced density region because of the thermal blooming process. Light rays bend away from less dense regions, thus, the beacon rays diverge from this region and arrive at the ground not exactly vertical as they would have been if they had not sensed the hot spot. They form a *tee-pee* of rays (see Fig. 2.23). The adaptive optics sensor takes this information and instructs the outgoing beam to retrace
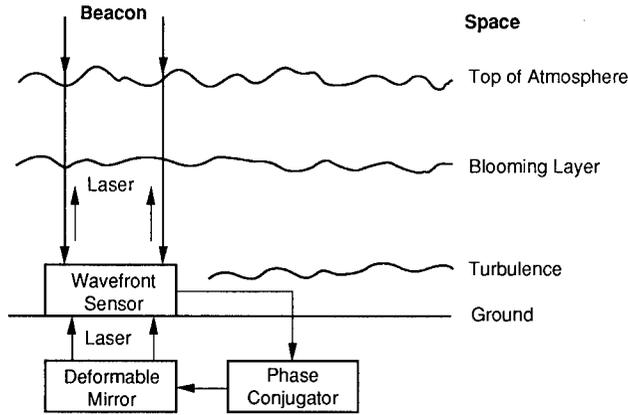
Fig. 2.22 Simplified view of an adaptive optics system used for thermal blooming compensation.
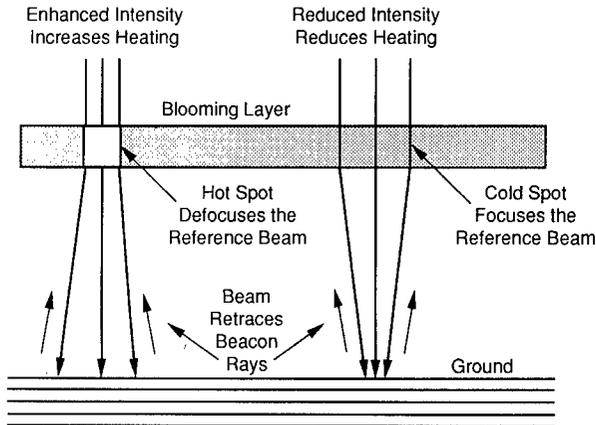
Fig. 2.23 Beacon is defocused by a hot spot. Thus, the rays form a "tee-pee" at the ground. The adaptive optics system forces the laser beam to retrace the beacon rays, causing the hot spot to get hotter.

the incoming beacon rays instead of propagating in a strictly vertical path. By doing so, the laser concentrates energy in the region of the blooming layer containing the hot spot. In other words, the hot spot gets hotter. Similarly, a cooler spot in the blooming layer will tend to get cooler. This enhancement of differential heating induced by scintillations by a phase conjugation adaptive optics system is termed the *phase conjugate instability* (PCI). As is readily evident, this positive feedback mechanism drives the unstable growth of small-scale structure much harder than it would naturally evolve if an adaptive optics system were not present.

Detailed analysis and computation of these phenomena provide us with expressions for Strehl ratio.[38] For the case of a steady wind, the worst-case scenario, we have

$$\text{SR}_{ss} = \exp\left(-3.47\frac{N_\lambda^2}{N_T^{5/6}}\right) , \tag{2.73}$$

where $N_\lambda = N_D/(2\pi)$ is the number of waves of distortion induced by thermal blooming; N_T is the turbulence Fresnel number, given by $N_T = r_0^2/\lambda L$; and L is the height of the blooming layer defined by Fig. 2.21.

Detailed analysis shows that the PCI growth is stronger for small scales of perturbations or scintillations. One is caught in a squeeze. One would like to make the adaptive optics systems capable of reducing the ambient turbulence down to the smallest scales of interest, but one cannot go too far in this scheme without driving the PCI phenomenon. We compromise by accepting some degradation caused by failure to completely correct for turbulence alone, in order to live with an acceptable level of PCI. It appears that this compromise is possible and practical in some systems of interest,[39] especially when we take into account the ameliorating effects of realistic winds.

Effect of Real Winds.   The preceding expression for the Strehl ratio probably overestimates the small-scale blooming problem because it ignores the positive effect of wind fluctuations on suppressing the growth of the structure. For maximum growth rates, the wind has to be uniform in speed and direction throughout the atmosphere. Otherwise, the scintillations no longer line up with the ground-level turbulence "lens array" that focuses and defocuses the laser to form the interference pattern and scintillations in the blooming layer. Hence, anything in the more realistic world that upsets this unique arrangement between turbulence layer and blooming layer dampens the growth of the instability.

The atmosphere is a continuum, and so absorption and thermal blooming are occurring continuously from the laser through the atmosphere. The intensity-to-phase process is occurring at all levels. If there are mechanisms that shift the different levels of the atmosphere with respect to each other, then the growth is strongly decoupled and dampened (see Fig. 2.24). This statement is universally true for both uncorrected and corrected (phase conjugated) beams. The possible mechanisms that contribute to this amelioration are (1) the natural wind changes in speed and direction as a function of altitude. (Natural wind shear displaces and rotates the scintillations with respect to the turbulent eddies near the ground, breaking up the direct link that drives the growth of the structure.) (2) The fluctuations in the wind caused by mechanical turbulence tend to randomly retard or advance the relative positions of scintillations and turbulence lenses. (This phenomenon has been termed *minishear,* first coined and suggested by Szeto and Fried.[39]) (3) The slewing of the beam as it points to a relay mirror or a target, when added to the natural wind, is an additional effective wind shear and alters the value of the wind speed as a function of altitude. (It can add to or subtract from the natural wind and hence can either decrease or increase blooming on the average.)

Detailed analysis gives an expression for the Strehl ratio in the presence of wind fluctuations[38]:

$$\langle \delta V_i(z_1)\delta V_j(z_2)\rangle = \frac{1}{2}\delta_{ij}(\delta V)^2 \exp\left(\frac{|z_1 - z_2|}{l}\right) , \tag{2.74}$$
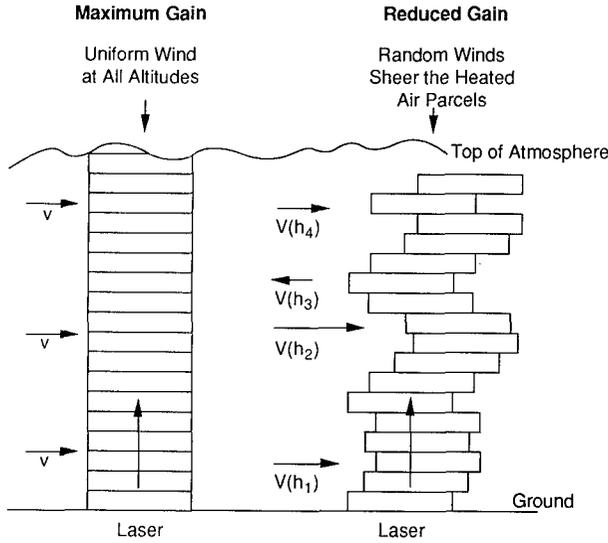
Fig. 2.24 Intensity-to-phase conversion that occurs continuously throughout the atmosphere is misaligned by wind shear, minishear, and slew. This uncouples regions of the atmosphere from each other and strongly dampens the instability growth, both with and without adaptive optics feedback.

where δ_{ij} is the Kroneker delta and l is the correlation length and typically $l \ll L$, the path length. The effect of δV is to separate the propagation path into layers of thickness l, between which the hot and cold small-scale structure is mixed, thereby reducing the growth rate of the instability. In this case, the Strehl ratio becomes

$$SR_{fw} = \exp\left(-14.9\,\frac{N_\lambda^2 l V_W}{N_T^{5/6} N_F^{1/2} L \delta V}\right) \,, \tag{2.75}$$

where $N_F = D^2/\lambda L$ is the whole beam Fresnel number, l is the correlation length of the wind fluctuations, L is the path length, and δV is the wind velocity fluctuation between layers.

A complete picture of the growth of small-scale structure can be summarized. The atmosphere can be thought of as made up of low-altitude turbulence and a higher altitude blooming layer. The low-level turbulence creates random phase changes that are changed by diffraction into random intensity fluctuations at the blooming layer. These intensity fluctuations are changed by absorption and thermal blooming into density changes that create phase changes for the beacon in the blooming layer. If the turbulence layer and the blooming layer have the same vector velocity, the process grows at a high rate. Adaptive optics produces a positive feedback mechanism that enhances the growth rate. The presence of natural wind shear, slewing, or minishear all reduce the growth rate of the turbulence-thermal blooming interaction to acceptable levels, based on current models and data.

2.4.4   Turbulence, Jitter, and Thermal Blooming

In the absence of small-scale instability, the effects of atmospheric turbulence, beam jitter, and whole beam thermal blooming can be analyzed by simple rules for estimating the three effects independently. In general, the response of the atmosphere for the three competing phenomena is extremely complicated. For example, thermal blooming strength is a function of the beam radius and the wind velocity. If beam jitter is low frequency, the beam moves around the atmosphere up to a few beam diameters. This has the effect of a false wind, which can reduce the degree of blooming. Similarly, if the jitter is high frequency, the beam intensity is smeared before the atmosphere has time to respond, effectively increasing the beam diameter. For intermediate frequencies, the beam's own jitter might follow the bending into the wind phenomena or even counteract it. All of these might occur simultaneously.[29]

Atmospheric turbulence has the effect of inducing dynamic tilt, which is an additional component of jitter. Turbulence can also induce beam spreading because of its high-spatial-frequency content. This can increase the apparent beam radius, thus reducing the local intensity and the magnitude of the thermal blooming.

The combined effects of atmospheric turbulence, whole beam blooming, jitter, and diffraction were investigated by Breaux et al.[40] They produced a set of algebraic scaling laws that could be applied to particular cases of atmospheric and diffraction effects. The formulations are based on fundamental principles and are easily employed for analysis of optical systems. The peak intensity of a beam after propagating a distance L is given by

$$I = \frac{P \exp(-\varepsilon L)}{\pi L^2 (\sigma_L^2 + \sigma_B^2)} \ . \tag{2.76}$$

The parameter ε is the total atmospheric extinction, which is the sum of absorption and scattering. The beam spread factor resulting from linear effects σ_L combines the effects of turbulence σ_T, jitter σ_J, and diffraction σ_D. For this illustration, we present the singular case of zero slew, uniform beams, and horizontal path propagation (C_n^2 constant) with no low-frequency beam wander. For the other options the reader is referred to Ref. 40.

The combined beam spreading resulting from linear effects is

$$\sigma_L^2 = \sigma_D^2 + \sigma_T^2 + \sigma_J^2 \ . \tag{2.77}$$

The diffractive spreading is given by

$$\sigma_D^2 \simeq \left(\frac{2}{\pi} \frac{B\lambda}{D}\right)^2 \ , \tag{2.78}$$

where B is the beam quality at the aperture of diameter D. The turbulence effects are divided into two regions. Where $(D/r_0) < 3.0$,

$$\sigma_T^2 = 0.182 \left(\frac{\sigma_D}{B}\right)^2 \left(\frac{D}{r_0}\right)^2 \ , \tag{2.79}$$

and where $(D/r_0) > 3.0$,

$$\sigma_T^2 = \left(\frac{\sigma_D}{B}\right)^2 \left[\left(\frac{D}{r_0}\right)^2 - 1.18\left(\frac{D}{r_0}\right)^{5/3}\right] . \tag{2.80}$$

The spread resulting from beam jitter σ_J^2 is given in units of radians squared. Breaux et al. approximate an effective variance for thermal blooming

$$\sigma_B^2 = 0.0143\sigma_L^2\Psi^{1.178} , \tag{2.81}$$

where Ψ is made up of various parameters representing blooming, beam quality, diffraction, and the other linear effects

$$\Psi = \frac{N_B N_F}{N_Q N_A} \exp\left(\frac{-\varepsilon L}{2}\right) \ln\left[\frac{(N_A + 1)}{(N_A N_C - N_C^2)}\right] , \tag{2.82}$$

where N_F is proportional to the Fresnel number,

$$N_F = \frac{\pi D^2}{2\lambda L} , \tag{2.83}$$

$$N_Q = B\frac{\sigma_L}{\sigma_D} , \tag{2.84}$$

$$N_C = 2.5\sigma_L\frac{L}{D} , \tag{2.85}$$

$$N_A = \left(1 + 6.24\frac{\sigma_L^2 L^2}{D^2}\right)^{1/2} . \tag{2.86}$$

The blooming strength N_B is calculated from the blooming strength equations for the various pulse or cw cases, or it can be simply stated (for the case of collimated cw steady-state blooming in standard dry air) as

$$N_B = 1.84 \times 10^{-8}\frac{P\alpha L^2}{D^3 v} , \tag{2.87}$$

where L and D are in meters, the wind velocity v is in meters per second, the absorption coefficient α is expressed in inverse meters, and the power is in watts.

Substitution of parameters into these equations allows a view of the combined effects of turbulence, whole beam thermal blooming, jitter, beam quality, and diffraction effects for uniform propagation. Comparison with detailed wave optics codes suggests that these expressions are good to a few percent in cases where the threshold for small-scale instabilities has not been exceeded.

2.4.5   Nonatmospheric Sources of Aberrations

Although turbulence is the major source of concern for astronomers using adaptive optics, and thermal blooming is a limiting factor for high-power beam

propagation, adaptive optics is often called to correct for aberrations that enter a system in other ways.

Optical Misalignments and Jitter.   The introduction of tilt or dynamic tilt (jitter) into a system can reduce its performance. In some cases, this performance is only slightly reduced, as in the case of an imaging system where tilt simply moves the position of the image without affecting its quality. These problems often can be accepted. On the other hand, slight misalignments in optics inside a laser resonator can greatly reduce its performance and, in many cases, eliminate the possibility of lasing altogether. Errors in setup or mounting, eternal error sources such as mechanical (acoustic) vibration, or thermal problems can all contribute to a poorly performing system.

Adaptive optics, in a tilt correction mode, can be used to correct for these errors. The problem of maintaining alignment or beam pointing has evolved from early electromechanical telescope and radar control systems to the present-day scanning and control systems that operate in the megahertz range. Tilt in a system can also induce other higher order aberrations. Severe tilts in complicated optical configurations can introduce many high-order aberrations. The elimination of tilt, or pointing error, is the first business of adaptive optics. Control of fast tilt (jitter) or slow tilt (wander or drift) is often all that is needed to bring a system into satisfactory operation.

The Strehl ratio for the effect of jitter is approximately[41,42]

$$SR = \exp\left[-\left(\frac{\pi}{2} \frac{\sigma_{jit} D}{\lambda} \right)^2 \right] , \tag{2.88}$$

where σ_{jit} is a two-axis rms beam jitter over an aperture of diameter D. The one-axis jitter is $\sigma_{jit}/\sqrt{2}$.

Large Optics: Segmenting and Phasing.   Striving for ever better means of collecting faint starlight or focusing laser beams to extremely small spots means requiring larger optics. The manufacture of larger and larger optical elements presents new problems for investigators and new challenges for adaptive optics.

The effects of gravity on massive objects cause a mirror to sag and deform. If this mirror is placed on a movable mount, the direction of the deformation changes with the motion. The sag cannot be calibrated out, nor can it be polished away, because it will change with operation.

The deflection δ_z, normal to the surface, of a circular plate of radius a that is lying horizontally[43] is

$$\delta_z = \frac{3\rho g(1 - \nu^2)}{16Eh^2}\left[\frac{5 + \nu}{1 + \nu}a^4 - \frac{2(3 + \nu)}{1 + \nu}a^2r^2 + r^4 \right] , \tag{2.89}$$

where ρ is the material density, E is Young's modulus, ν is Poisson's ratio, h is the disk thickness, and g is gravitational acceleration. The aberrations for a mirror or large lens under the influence of gravity are a constant piston, a sphere term proportional to r^2, and a spherical aberration proportional to r^4.

Large mirrors can be self-correcting, that is, they can contain the necessary physical elements to adjust their figure and correct aberrations in the other

parts of the system as well.[44] Large optics can be subdivided into smaller segments. These segments then can be coaligned, or phased to act as a larger mirror. The aberrations induced in large mirrors and the systems that are used to phase segments have been under study for a number of years.[45,46] For the purpose of the study of adaptive optics, aberrations that are caused by gravity (low spatial frequency) and aberrations caused by local defect or deformation (high spatial frequency) can be treated in the same manner as similar aberrations on smaller optics, with only the magnification of spatial frequency taken into account.

The aberrations that occur at mismatched edges, however, have spatial frequency content that is very high. These almost infinite spatial frequencies must be treated differently. The correction over the edges can be simply disregarded or the light that impinges these edges can be masked out.

The Strehl ratio for a segmented mirror that is not perfectly phased[41] is

$$\text{SR} = \exp\left[-\left(\frac{N_{\text{seg}} - 1}{N_{\text{seg}}}\right) 4\pi^2 \frac{\sigma_p^2}{\lambda^2} \right] , \qquad (2.90)$$

where N_{seg} is the number of segments and σ_p is the 1-σ, zero mean, rms residual optical path difference (OPD) piston error on each segment.

Thermally Induced Distortions of Optics.   When a beam of light strikes an optical surface, some of the energy is absorbed. For low-power beams, imaging systems, and transparent glasses the absorption is usually never noticed. Transmitting higher powers results in thermally induced aberrations. High-power beams (of the order of kilowatts per square centimeter) can raise the temperature of the optics. Passive or uncooled mirrors and actively cooled mirrors exhibit distortions under high-power illumination.

For a mirror modeled as a flat plate, constrained at the edges, the distortion is a constant plus a function of the radial distance from the center. Thus, if uniform power P is absorbed, the temperature profile results in stresses that cause a characteristic paraboloidal bending of the plate.

$$\delta z = P(C_0' + C_2' r^2) . \qquad (2.91)$$

This is called *bowing distortion,* or *power induced distortion,* and it is proportional to the total absorbed power.

If the intensity profile $I(x,y)$ on the plate is not uniform, regions of local temperature increase result in another form of distortion. The displacement in the out-of-plane direction δ_z is proportional to the absorbed power distribution. With α representing the absorption,

$$\delta z(x,y) = \xi \alpha I(x,y) . \qquad (2.92)$$

The symbol ξ, used in much of the literature, resembles a worm and has been called the *worm factor.* It is a function of material and cooling parameters. Measurements show it to be between 5 and 100 Å W^{-1} cm^{-2}. Because this form of thermal distortion produces a map of the intensity pattern, it is called *intensity mapping distortion.* It is proportional to the intensity distribution rather than the total power.

The thermomechanical response of a plate cannot always follow the intensity pattern exactly. Thermal diffusion tends to smooth the temperature distribution and mechanical response of the plate tends to smooth it more. Those effects are expressed as convolutions (represented by *) between the response of the material to thermal loads R_T, response of the material to mechanical stress R_M, and response to the absorbed intensity forcing function $\alpha I(x,y,t)$:

$$\delta z(x,y,t)_{\text{therm}} = R_T(t) * R_M(t) * \alpha\xi I(x,y,t) \ . \tag{2.93}$$

The Strehl ratio for thermal distortion is derived from the wavefront error it imparts to the beam:

$$\text{SR} = \exp\left[-\left(\frac{2\pi}{\lambda}\right)^2 (\Delta\Phi_P)^2 \right] \ , \tag{2.94}$$

where the mean-square wavefront error is

$$(\Delta\Phi_P)^2 = \frac{\displaystyle\iint_{\text{surface}} 4[\delta z(x,y,t)_{\text{therm}} - \overline{\delta z(t)}]^2 \ dx \ dy}{\text{area of surface}} \ . \tag{2.95}$$

The factor 4 appears because the phase change on reflection is twice the surface deviation. The spatial mean (average piston component) is $\overline{\delta z(t)}$. Because the thermomechanical response of the optical surface is usually much slower than high-frequency variations in the intensity distribution, the adaptive optics system response can have a lower bandwidth than the intensity fluctuations indicate.

Manufacturing and Microerrors. The materials chosen for refractive or reflective optics may have surface errors that figuring and polishing cannot remove. Generally, they are of such low magnitude or high spatial frequency that adaptive optics are useless against them. Occasionally, a large systematic error may occur that produces a low-order aberration. A variation in polishing pressure, for instance, may produce such a defect that can be removed with adaptive optics when the optical element is used.

Even after polishing and coating, there may be roughness in the surface (called *microstructure*) that would affect an adaptive optics system. The microstructure of polished optical materials is expressed in rms roughness,[47] which can range from about 10 Å for fused quartz to over 100 Å for materials such as molybdenum and beryllium. The fraction of incident light scattered, the total integrated scatter (TIS), from a surface with rms roughness δ_{rms} is[48]

$$\text{TIS} \simeq \left(\frac{4\pi\delta_{\text{rms}}\cos\theta}{\lambda}\right)^2 \ , \tag{2.96}$$

where λ is the incident light wavelength and θ is the angle of incidence. For roughnesses approaching even small fractions of a wavelength, we can see that the negative effects of scattering can be severe. Large roughnesses are

also known to be extremely sensitive to incident high powers. If light is multiply reflected in the hills and valleys of the surface, absorption and microthermal mapping can lead to stresses that result in permanent damage to a mirror surface.

Another characteristic of the microstructure, or any other random surface property, is its autocovariance length. Polishing or figuring processes may have some spatial correlation. By calculating the autocovariance $G(r)$ of the surface $z(\rho)$, we can obtain a statistical description of the surface roughness or figure error:

$$G(r) = \lim_{L \to \infty} \int_{-L/2}^{L/2} z(\rho)z(\rho + r) \, d\rho \; , \tag{2.97}$$

where L is the one-dimensional extent of the surface under investigation. Examination of the autocovariance in either one or two dimensions can lead to knowledge about the spatial characteristics of the random surface. A narrow $G(r)$ suggests a very fine grain roughness with no interaction across large regions. This is characteristic of surface microroughness. A wide $G(r)$ suggests some spatial correlation. This occurs mostly as a result of systematic and repeatable figuring or coating processes. A spike in the autocovariance function suggests a strong correlation at one particular separation. A radial-pattern diamond turning machine may produce small defects that exhibit this periodic behavior. A cooled mirror with improperly pressurized coolant channels may also exhibit periodic print-through of the coolant channels that show up as a spike or bump in the autocovariance function.

Even though adaptive optics cannot correct for many of the very high spatial frequency errors, these errors can mask the disturbances that adaptive optics is supposed to correct. These uncorrectable errors therefore fall into the category of irretrievable losses along with atmospheric absorption and scattering. The high-angle scattering induced in an optical train, however, can be a problem because of the stray light. For low-light-level applications, these losses are costly. For high-power applications, the stray light can be damaging to the surrounding environment and catastrophic to delicate sensors.

Aberrations in Laser Resonators and Lasing Media.   Laser resonators present many opportunities for aberrations to arise. The multiple-pass nature of most resonators lets an aberrated propagating wave diffract into intensity fluctuations. The intensity fluctuations can cause further degradation.[49] Optics inside a resonant cavity can become misaligned like optics in any beam train. Tilts, defocus,[50] and astigmatism can appear in resonators[51] and can reach critical levels where the resonator can no longer support lasing.[52]

The laser gain medium can be responsible for a number of aberrations. Because lasing can take place in solids, liquids, gases, or plasmas, any imperfections in these media that result in inhomogeneous distributions of gain can reduce laser performance. Structural imperfections in solid-state lasers or turbulence in liquid or gas lasers[53] are the most common problems. Dynamic gas lasers and free-electron lasers have rapidly moving laser media that provide another source of disturbance to the wavefront.[54]

Because the complex nature of these aberrations and the difficulty in predicting them for any specific device, one usually relies on measurements of

laser output amplitude and phase before specifying an adaptive optics beam cleanup system. A detailed discussion of effects like these that occur in high-energy lasers appears elsewhere in this Handbook.

2.5 WAVEFRONT SAMPLING

A wavefront sensor must have a high-fidelity replication of the wavefront that needs correction. Many adaptive optics systems show a beamsplitter fulfilling the beam-sampling function. The beamsplitters can be actual coated glass components, commonly called *beamsplitters,* or they can represent any device that separates the ongoing beam and the sampled beam. Examples of wavefront samplers are shown in Fig. 2.25.
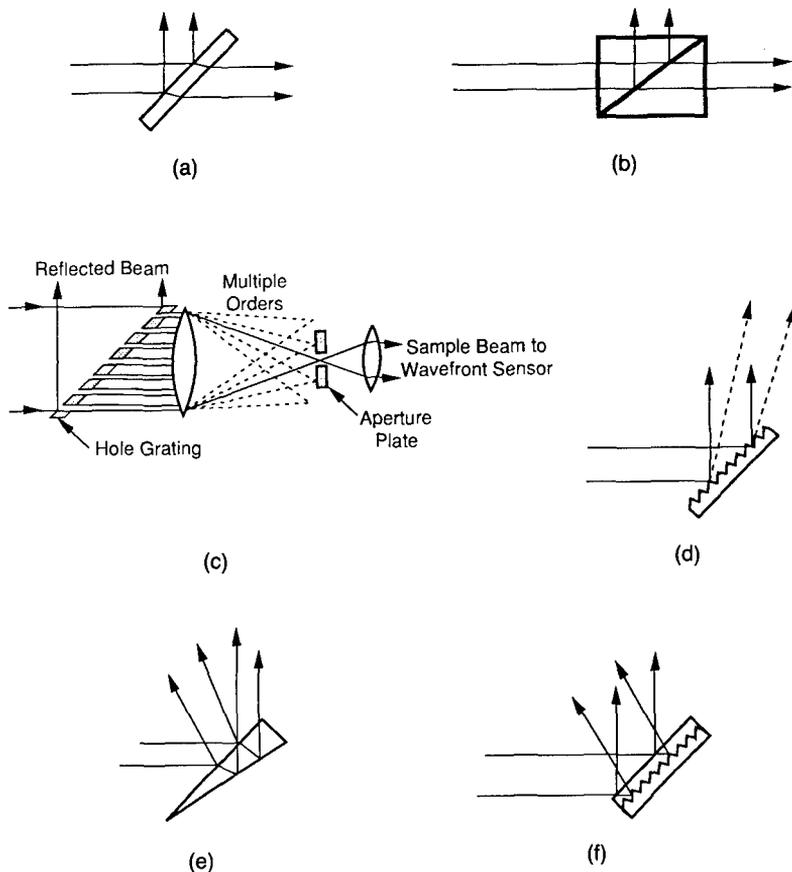
Fig. 2.25 Various wavefront samplers: (a) plate beamsplitter, (b) cube beamsplitter, (c) hole grating, (d) diffraction grating, (e) wedge, and (f) buried grating.

2.5.1   Transmissive Samplers

Commercially available flat plate beamsplitters are thin glass plates with edges parallel to the order of a few milliradians. The surface incident to the beam, usually called the *front surface*, is coated with a film having partial reflection properties for the wavelength of interest. The transmitted portion of the beam exits the back surface, which has a broadband antireflection (AR) coating. The reflection-transmission ratios (T/R) are specified along with the incidence angle and, sometimes, the polarization state. For example, T/R 50:50 ± 5%; 45 deg. Off-the-shelf components have surface flatness that is typically a few waves across the device.

A cube beamsplitter is a matched pair of right-angle prisms. This type of beamsplitter deforms much less than a flat plate when it is stressed by mounting. Specifications for cube splitters are similar to those for plate splitters.

Pellicle beamsplitters consist of a membrane stretched over a frame. Because the membrane is of the order of 5 μm thick, the chromatic and spherical aberration on the transmitted beam are negligible.

Dichroic beamsplitters separate the sample by transmitting one wavelength and reflecting another. They are specified by the amount of light passed at each particular wavelength. For instance T(94%)550 nm/R(98%)1.06 μm/flat plate/45 deg specifies a flat plate, 45-deg incident beamsplitter with visible transmission of at least 94% and IR reflection of at least 98%.

A hole grating is a flat mirror with a regular pattern of holes drilled in it. Most of the beam energy can be reflected from the front surface. By allowing only the zero-order sample through an aperture plate [see Fig. 2.25(c)], the greatly attenuated (10^{-4} to 10^{-6}) replica of the input beam far-field diffraction pattern can be observed.[55,56]

2.5.2   Reflective Samplers

When aberrations are slowly varying, continuous sampling is not required. By driving an acousto-optic cell, alternately creating a grating and removing it, a beam can be alternately directed toward different beam paths. For slowly varying aberrations, such as thermal distortion, a mechanical chopper can be used to alternately direct a sample beam out of the principal path.

A simple wedge or prism can be used when one needs to spectrally separate two beams. If two superimposed beams of wavelengths λ_1 and λ_2 are incident at angle θ_i, a wedge (with wedge angle θ_W) the sine of the angle of the reflected beam from the front surface is

$$\sin\theta_R = n \sin2\theta_W \cos\theta_i + \cos2\theta_W \sin\theta_i .$$

The different directions of the two beams arise because the index of refraction is slightly different for the two wavelengths.

Diffraction gratings are optical elements with a surface that is modulated with a regular groove or slit pattern. They are used in adaptive optics primarily as a method of separating two beams. One of the beams will be passed along (either by reflection or transmission) toward a wavefront sensor. The different orders of diffraction follow the basic grating equation:

$$d(\sin\theta_m - \sin\theta_i) = m\lambda \ , \qquad\qquad (2.98)$$

where d is the separation of the grooves, θ_i is the incident angle measured from normal, and θ_m is the angle of the m'th diffracted order. The plane of incidence is perpendicular to the grooves. The grating properties are usually chosen so that only one diffraction order has significant power.

Because a diffracted beam from a grating is slightly dispersed, a second grating is often used to correct the dispersion. A *grating rhomb* is a configuration where two identical gratings are parallel. The diffracted sample is propagated parallel to the incident beam and perpendicular to the reflected beam.

2.5.3 Hybrids

Many adaptive optics systems cannot use a simple grating or beamsplitter for beam sampling. Transmissive beamsplitters cannot handle high-power flux without significant distortion or destruction. Whereas a wedge works by reflecting a beam from the front surface in a different direction from a beam reflected from the back surface, a *buried mirror* is constructed so that the beam reflected from the back surface is focused differently than the beam reflected from the front surface. The beam reflected from the back surface can be focused toward an off-axis optical element and then transmitted toward the wavefront sensor.

Combining the effects of gratings and the elements of the buried mirror resulted in the concept of a *buried grating*.[57] When the grating is coated with a material that reflects one wavelength and transmits another, the buried grating is used as a beam sampler. If the line spacing of the grating is small (of the order of wavelengths) it functions as a diffraction grating. The sample from this buried short-period grating (BSPG) leaves as a higher order of the grating. When the line spacing of the grating is large (of the order of millimeters) it functions more like a buried mirror. The transmitted beam reaches the grating surface and specularly reflects from each facet. This buried long-period grating (BLPG) is the same as a buried wedge with a faceted back surface. Because the BLPG does not operate by diffraction, but rather by direct reflection, it is insensitive to polarization.

The wavefront sampling error that affects the adaptive optics system performance is composed of two uncorrelated terms. The first term is the wavefront error that is placed on the beam by distortions of the optics on the beam transmission path. This error arises in all the optics, and it is not restricted to wavefront samplers. It is expressed as the variance of the wavefront that is outside the spatial or temporal bandwidth of the adaptive optics system. The error is placed on the beam when it is reflected from or transmitted through the sampler. It can be expressed (in units of distance squared) as a wavefront variance σ^2_{dist}. The second term is the uncommon path wavefront error, which occurs when the fidelity of the sample is not perfect. When the sampled portion of the beam experiences a different wavefront variation than the beam to be corrected, the closed-loop wavefront sensor and corrector will add this variation, or its conjugate, to the beam. Represented by σ^2_{path}, the contributions to sampling error reduce the Strehl ratio by

$$\text{SR} = \exp\left[-\left(\frac{2\pi}{\lambda}\right)^2(\sigma_{\text{dist}}^2 + \sigma_{\text{path}}^2)\right] . \tag{2.99}$$

2.6 WAVEFRONT SENSING

Adaptive optics wavefront sensors must rapidly determine the useful wavefront information from a sample of the beam. Disturbances can occur at rates approaching hundreds of hertz. Wavefront sensors must also be able to measure the information with the necessary spatial resolution to adequately compensate for the aberrations across the full aperture. Two basic types of wavefront information are used. When the wavefront error is expressed in terms of the phase over a small spatial area, or zone, the wavefront error is said to be *zonal*. When the wavefront error is expressed in terms of coefficients of the normal modes of a polynomial expansion over the entire pupil, it is said to be *modal*.

The zonal method divides the wavefront over an aperture into N subapertures. As N approaches infinity, the wavefront is exactly represented. Typically, each subaperture is represented by its piston component and/or two local orthogonal tilt (slope) components.

The modal method divides the wavefront over an aperture into its normal modes. The modes can have any polynomial form. The two most common forms, written in polar coordinates, are the power series representation,

$$\Phi(r,\theta)_{\text{pwr series}} = \sum_{n,m=0}^{\infty} S_{[c]nm}r^n \cos^m\theta + S_{[s]nm}r^n \sin^m\theta , \tag{2.100}$$

or a Zernike polynomial representation, that is orthonormal over a unit circle of radius R',

$$\Phi(r,\theta)_{\text{Zernike}} = A_{00} + \frac{1}{\sqrt{2}} \sum_{n=2}^{\infty} A_{n0}R_n^0\left(\frac{r}{R'}\right)$$

$$+ \sum_{n=1}^{\infty} \sum_{m=1}^{n} (A_{nm}\cos m\theta + B_{nm}\sin m\theta)R_n^m\left(\frac{r}{R'}\right) . \tag{2.101}$$

The form of the radial polynomial $R_n^m(r/R')$ and other restrictions on these representations can be found in Born and Wolf.[12] Both of these common representations have modes corresponding to third-order optical aberrations: focus, tilt, astigmatism, and coma.

The decision to use modal, zonal, or a combination of sensing schemes to measure the wavefront is usually left to the application. If low-order modes (tilt, defocus) are dominant, then modal analysis and corrections should be used. If high-order aberrations are present, the zonal approach should be used. For the case of atmospheric turbulence, both zonal and modal sensing have been used. Winocur[58] shows how the Strehl ratio is improved with many correction zones of characteristic length d_s. His expression is given as

Table 2.1   Residual Errors for Atmospheric Turbulence Correction

σ_M^2	$SR = \exp\left[-\sigma_M^2\right]$
$\sigma_1^2 = 1.0299(D/r_0)^{5/3}$	$\sigma_{12}^2 = 0.0352(D/r_0)^{5/3}$
$\sigma_2^2 = 0.582(D/r_0)^{5/3}$	$\sigma_{13}^2 = 0.0328(D/r_0)^{5/3}$
$\sigma_3^2 = 0.134(D/r_0)^{5/3}$	$\sigma_{14}^2 = 0.0304(D/r_0)^{5/3}$
$\sigma_4^2 = 0.111(D/r_0)^{5/3}$	$\sigma_{15}^2 = 0.0279(D/r_0)^{5/3}$
$\sigma_5^2 = 0.0880(D/r_0)^{5/3}$	$\sigma_{16}^2 = 0.0267(D/r_0)^{5/3}$
$\sigma_6^2 = 0.0648(D/r_0)^{5/3}$	$\sigma_{17}^2 = 0.0255(D/r_0)^{5/3}$
$\sigma_7^2 = 0.0587(D/r_0)^{5/3}$	$\sigma_{18}^2 = 0.0243(D/r_0)^{5/3}$
$\sigma_8^2 = 0.0525(D/r_0)^{5/3}$	$\sigma_{19}^2 = 0.0232(D/r_0)^{5/3}$
$\sigma_9^2 = 0.0463(D/r_0)^{5/3}$	$\sigma_{20}^2 = 0.0220(D/r_0)^{5/3}$
$\sigma_{10}^2 = 0.0401(D/r_0)^{5/3}$	$\sigma_{21}^2 = 0.0208(D/r_0)^{5/3}$
$\sigma_{11}^2 = 0.0377(D/r_0)^{5/3}$	

The subscript M denotes the number of modes corrected for an aperture of diameter D.
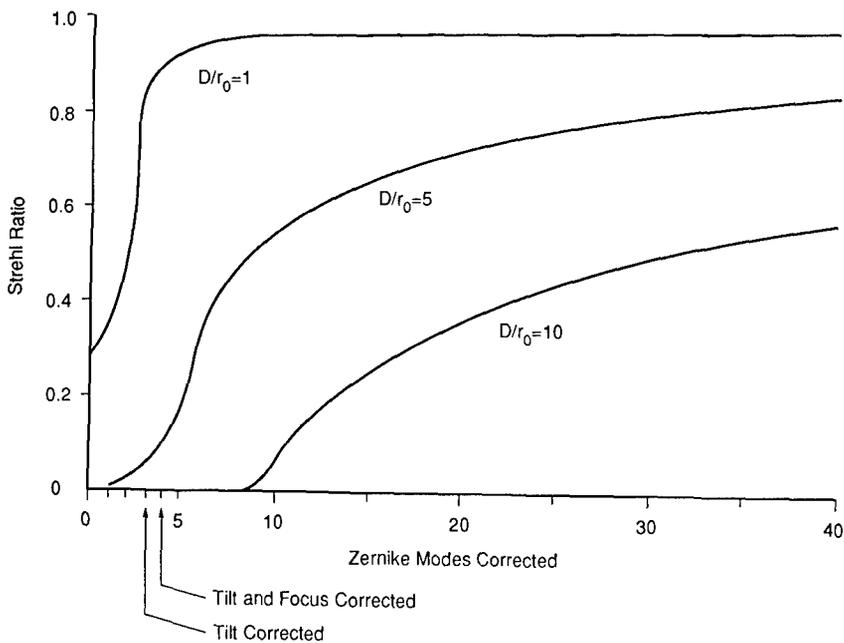
Fig. 2.26 Strehl ratio reduction for atmospheric turbulence compensation with a finite number of Zernike wave front modes corrected.

$$\text{SR} = \exp\left[-(2\pi)^2 0.008\left(\frac{d_s}{r_0}\right)^{5/3}\right] . \tag{2.102}$$

An expression for the Strehl ratio in terms of the number of modes corrected, M,

$$\text{SR} = \exp(-\sigma_M^2) , \tag{2.103}$$

has a similar form.[27] For the first 21 modes, Table 2.1 should be used. For more than 21 modes of correction, one can use the equation

$$\text{SR} = \exp[-0.2944 M^{-(\sqrt{3}/2)}(D/r_0)^{5/3}] . \tag{2.104}$$

This relationship is shown graphically in Fig. 2.26.

2.6.1   Sensing of Lower Order Modes

Tilt Sensing.   Two methods are commonly used to sense wavefront tilt. One makes use of optical interference because the interference of two beams creates an intensity that is proportional to the difference in phase between the beams. If those beams are small adjacent portions of the same beam, then the differential phase, or tilt, in that region can be determined. Another method exploits the characteristic that wavefront tilt displaces a focused image in the direction of the tilt. The displacement of the centroid of the focused image is proportional to the amount of tilt, and it is insensitive to small amounts of higher order aberrations.

The centroid of the intensity pattern $I(x,y)$ is defined as the first moment of the pattern:

$$\bar{x} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x I(x,y)\, dx\, dy \Big/ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(x,y)\, dx\, dy , \tag{2.105}$$

$$\bar{y} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y I(x,y)\, dx\, dy \Big/ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(x,y)\, dx\, dy , \tag{2.106}$$

One can measure the position of the centroid in many ways. One method commonly used in image trackers and wavefront sensors is the *quadrant detector,* or *quadcell.* This device has four detectors that are closely spaced. The intensity pattern is made to focus on the array of four detectors. The difference signals (see Fig. 2.27),

$$\bar{x} = \frac{\int_1 I\, dS + \int_4 I\, dS - \int_2 I\, dS - \int_3 I\, dS}{\sum_i \int_i I\, dS} , \tag{2.107}$$
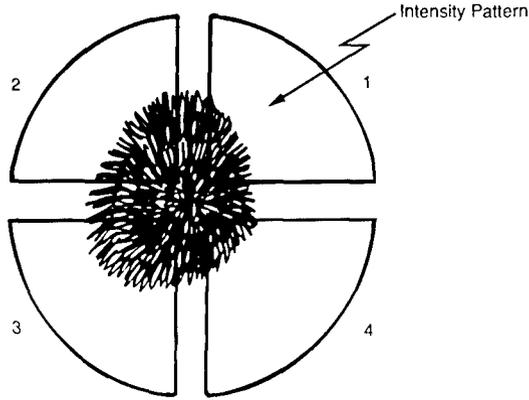
Fig. 2.27   Quadcell detector used for tilt sensing.

$$\bar{y} = \frac{\int_1 I \, dS + \int_2 I \, dS - \int_3 I \, dS - \int_4 I \, dS}{\sum_i \int_i I \, dS} \, , \tag{2.108}$$

are proportional to the position of the centroid. Although quadcell tilt detection is restricted to images that deposit at least some energy in all four quadrants and have nearly uniform intensity distributions in the image plane, the principle is widely used in wavefront sensors.

Resolved images on the quadcell can be used to determine wavefront tilt if errors resulting from object shape and the SNR are not neglected. The error in the angular measurement of tilt for a circular object on a quadcell is given as[59]

$$\sigma_{qc} = \frac{\pi[(3/16)^2 + (n/8)^2]^{1/2}}{\text{SNR}_v} \frac{\lambda}{D} \, , \tag{2.109}$$

where $n = (b\lambda)/(RD)$, with b the object diameter, R the range to the object, D the diameter of the aperture, and SNR_v the voltage SNR. When the object is unresolvable, i.e., b/R is small and $n \to 0$, the centroid measurement error is

$$\sigma_{qc} = \frac{\pi}{\text{SNR}_v} \frac{3\lambda}{16D} \, , \tag{2.110}$$

and when the object is large, $(b\lambda/RD) \gg 3/2$, the error is proportional to object size b:

$$\sigma_{qc} = \frac{\pi b (\lambda/D)^2}{8R(\text{SNR}_v)} \, . \tag{2.111}$$
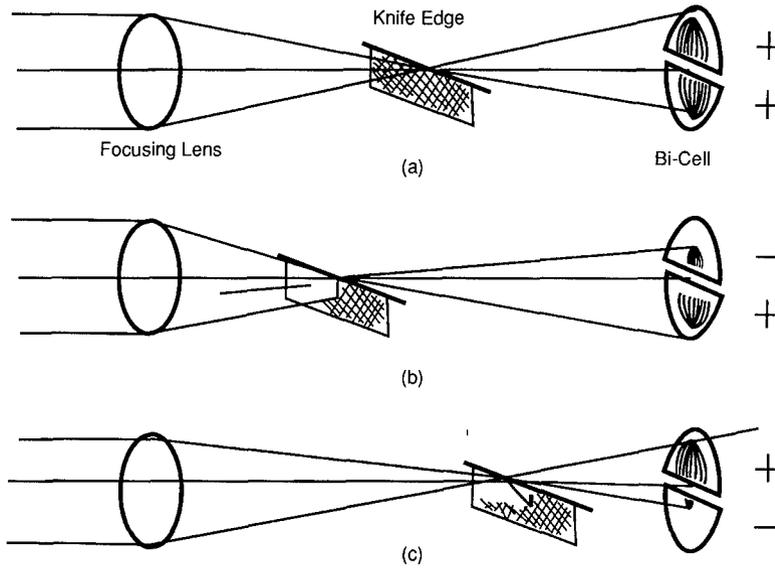
Fig. 2.28 Foucault knife edge test using a bicell for focus sensing.

An error in tilt measurement is related to an equivalent wavefront error over the aperture and, therefore, a reduction in Strehl ratio:

$$SR = \exp\left[-4\pi^2\left(\frac{D\sigma_{qc}}{2\lambda}\right)^2\right] .$$

(2.112)

With the advent of arrays of many small detectors, the methods for determining the centroid have advanced. With enough information from the array, direct calculation of the centroid using Eqs. (2.105) and (2.106) is possible.

Focus Measurement. Because a defocussed image of a point object has light spread out spatially more than the diffraction-limited image, the intensity distribution over a finite region acts as some measure of the defocusing. The Foucault *knife edge* test is used in real-time systems by placing the sharp edge near the focus and blocking half the beam propagating toward a bicell (a pair of adjacent detectors). The difference between the intensities on the two detectors of the bicell is related to the distance between the knife edge and the focal plane (see Fig. 2.28). By proper calibration, the system can be automated and used in real time.

2.6.2 Shearing Interferometer

The wavefront over an entire aperture can be determined from measurements of tilt in smaller regions (zones). This is analogous to fitting a continuous curve to points that represent local slopes. A fundamental modification to a Mach-Zehnder interferometer was introduced by Bates.[60] By tilting the mirrors of a Mach-Zehnder a small amount about their optic axis, the two beams that
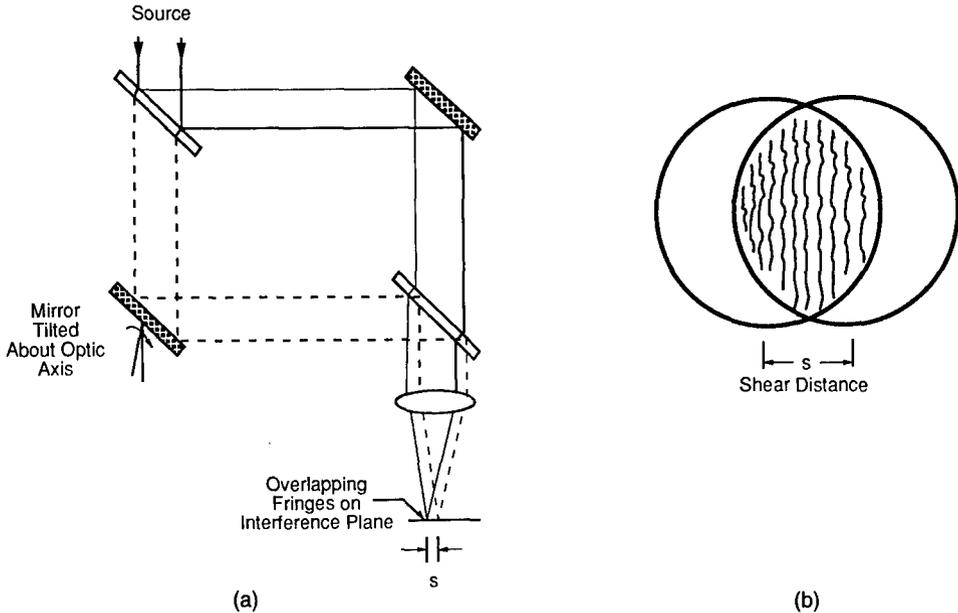
Fig. 2.29   (a) Shearing interferometer. (b) By laterally translating the beam from one leg of the interferometer, the intensity of the fringes represents the local wave front slope of the beam.

form the interference fringes do not totally overlap, as shown in Fig. 2.29(a). Their centers are separated by a distance s, called the *shear distance*. The pattern of fringes appears in the area of overlap; see Fig. 2.29(b). The orders of the fringes are displaced by an amount proportional to the shear. This order displacement Δm can be used to measure the differential OPD (tilt) at any point in the beam:

$$\Delta m \approx \frac{s}{\lambda} \frac{d}{dx}[\text{OPD}(x)] \ . \tag{2.113}$$

Shearing a wavefront makes use of the principle of self-referencing. That is, an interferogram can be produced without the need for a separate plane-wave reference. The resultant interferogram of two sheared wavefronts becomes a map of regions of constant slope of the wavefront. The principal advantage of shearing is the ability to distinguish "bumps" from "holes" in the wavefront. The sign of the slope is a direct realization of the "polarity" of the wavefront.

The process of producing the shear is the topic of many engineering papers. The most notable is the work of Wyant,[61] who uses a double-frequency grating to produce two first-order wavefronts that are laterally sheared and separated from the zero order (see Fig. 2.30).

The beam containing the unknown wavefront is focused onto a grating that has two different line spacings d_1 and d_2 at spatial frequencies ν_1 and ν_2. The angular shear between the two beams is then $\Delta\theta = \lambda(\nu_2 - \nu_2)$. To detect the interference pattern, an array of detectors is placed in the interference plane.
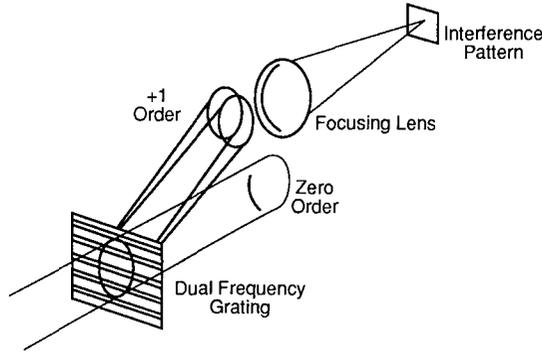
Fig. 2.30 Dual-frequency grating shearing interferometer.

By translating the grating along its optical axis at a velocity v, a Doppler shift occurs that modulates the phase. This application of ac heterodyning causes the signal on each of the detectors to have a sinusoidal pattern proportional to

$$\sin[\omega t + \phi(x,y)] \ , \tag{2.114}$$

where ω is the frequency of the Doppler shift $\omega = 2\pi v(\nu_2 - \nu_1)$. The time when the signal passes through zero can be electronically detected, and that time is proportional to the phase difference in each subaperture.

Shearing interferometers are not limited to lateral shearing nor to coherent light. Some shearing interferometers have been used in white light.[62] The most common variation is the radial shearing interferometer.[63] In this type of wavefront sensor, the beams are separated and magnified by different values. When the beams are superimposed, with their axes coinciding, interference occurs at the area of overlap, within the diameter of the smaller beam.

For a low-light-level, photon-counting shearing interferometer Wyant[64] showed that the variance of the phase estimation is

$$\sigma^2 = \frac{0.81 d_s^2}{s^2 \gamma^2 n} \ , \tag{2.115}$$

where d_s is the subaperture size, γ is the fringe contrast, s is the shear distance, and n is the number of photon counts in a period,

$$n = N \Omega T A \tau R / e^- \ . \tag{2.116}$$

In Eq. (2.116), N is the radiance of the source, Ω is the solid angle of the optics, T is the transmission of the optics, A is the aperture area, τ is the integration time, R is the responsivity of the detector, and e^- is the electric charge. When the effects of noise are factored in, the variance for the shearing interferometer phase estimation is[65]

$$\sigma^2_{\text{WFS(SI)}} = \frac{2d_s^2}{s^2\gamma^2(\text{SNR})^2} \; , \qquad\qquad (2.117)$$

where the SNR for n_N noise events is

$$\text{SNR} = \frac{n}{(n^2 + n_N^2)^{1/2}} \; . \qquad\qquad (2.118)$$

The Strehl ratio reduction for a shearing interferometer is then

$$\text{SR} = \exp\left\{ -\left[\frac{2d_s^2}{s^2\gamma^2(\text{SNR})^2} \right] \right\} \; . \qquad\qquad (2.119)$$

2.6.3 Hartmann Wavefront Sensors

The principle of the Hartmann wavefront sensor is shown in Fig. 2.31. The wavefront is divided by a mask, an array of gratings, or an array of transmissive lenses. Each of the beams in the subapertures is focused onto a detector and, from the lateral position of the spot centroids, the individual subaperture tilt is determined. Modifications to the basic Hartmann sensor are used to overcome drawbacks such as sensitivity to vibration and calibration. These can be done while maintaining the advantages of the Hartmann technique, that is, wide dynamic range, high optical efficiency, white light capability, no 2π ambiguity, and the ability to use continuous or pulsed sources.

To detect the position of the spot, various forms of modulation, detector geometry, and electro-optical processing are used. For astronomy the local wavefront tilt in each coherence diameter r_0 must be measured accurately in each subaperture of diameter d_s, which leads to

$$d_s \leq r_0 \; . \qquad\qquad (2.120)$$

The sensor must not be adversely affected by anisoplanatism in the atmosphere, so the subaperture must be large enough to resolve the isoplanatic patch. See Eq. (2.136).

$$d_s \geq 1.22\lambda \left[2.91 k^2 \int_0^L C_n^2(h) h^{5/3} \, dh \right]^{3/5} \; . \qquad\qquad (2.121)$$

Investigators have analyzed the noise characteristics of wavefront sensors[66] to explore the trades necessary between small apertures, which give high resolution, and large apertures, which maximize SNR.

The most fundamental method of removing alignment errors is the introduction of a well-characterized reference beam. The reference must be distinctive so that the detector processing electronics can discriminate between the signal beam and the reference beam. This is most easily done by modulating the reference and then electronically filtering the two beams after detection. Other methods, such as nutating the beams,[67] or rotating the square quadcells with respect to square subapertures,[68,69] eliminate some of the sensitivity to vibration and to beam or object geometry.
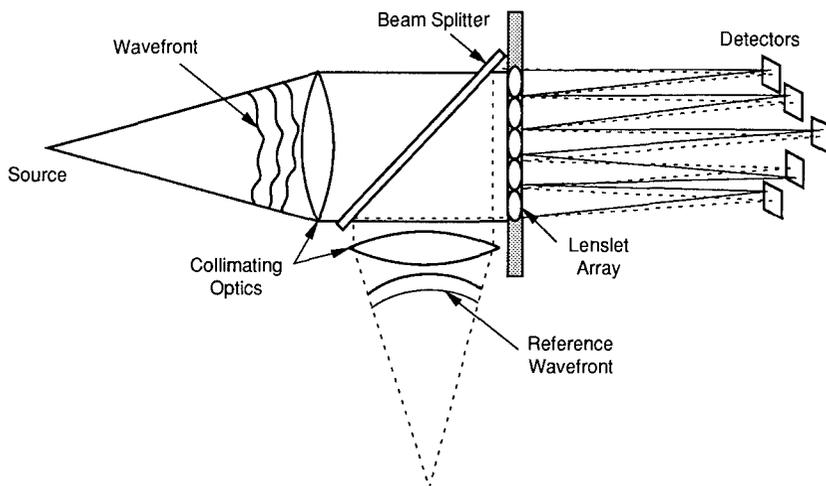
Fig. 2.31 Hartmann sensor divides the beam into subapertures and measures the position of the focused spot that is proportional to local wavefront tilt.

With the introduction of multiple-element staring detector arrays, many new approaches have been demonstrated. Because the basic quadcell has limited linearity and dynamic range, both one- and two-dimensional arrays have been used. One wavefront sensor used at the Kitt Peak National Observatory, Arizona,[70] uses 5×5 regions of a larger array for each of 37 subapertures. The spots nominally have a width of 2 pixels. The centroid is calculated digitally.

The error in wavefront measurement with a Hartmann sensor is derived from the error in measuring the individual subaperture tilts [Eq. (2.111)].

$$\sigma_{\text{tilt}}^2 \cong 0.35 \frac{\lambda^2}{d_s^2 (\text{SNR}_V)^2} \ .$$
(2.122)

The wavefront (WF) variance for a tilted beam in a circular subaperture is

$$\sigma_{\text{WF}}^2 = 4\pi^2 \left(\frac{d_s}{4\lambda}\right)^2 \sigma_{\text{tilt}}^2 \ .$$
(2.123)

When these two equations are combined, the wavefront measurement variance for a Hartmann sensor is

$$\sigma_{\text{WFS(H)}}^2 = 0.35 \frac{\pi^2}{4(\text{SNR}_V)^2} \ .$$
(2.124)

The Strehl ratio reduction for a Hartmann sensor, shown in Fig. 2.32, is then

$$\text{SR} = \exp\left\{-\left[\frac{0.86}{(\text{SNR})^2}\right]\right\} \ .$$
(2.125)
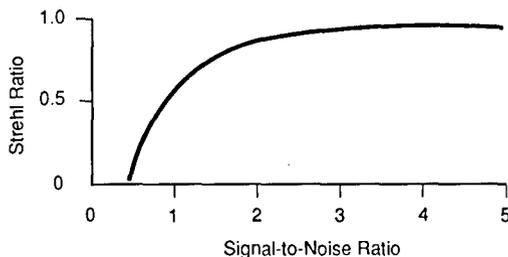
Fig. 2.32 Strehl ratio reduction because of Hartmann sensor SNR.

Roddier[71] and Roddier et al.[72] have shown how the methods of focus sensing, described in Sec. 2.6.1, and the Hartmann subaperture division process can be combined. The method, called *curvature sensing*, measures local wavefront curvature in each subaperture. By comparing the intensity at two equally spaced points on either side of the subaperture focal plane, the local curvature is found. The resultant scalar map of local curvatures and radial slopes at the aperture edge can be used to reconstruct the wavefront.

2.6.4 Multidither Sensing and Correction

If two beams are focused on a target and they are constructively interfering, the intensity will be a maximum. If they are destructively interfering the intensity is a minimum. If the disturbance in the propagation path is changing (relatively slowly), the phase that causes constructive interference can be found by rapidly modulating (dithering) the phase of one of the beams and simultaneously sensing the intensity. Setting the average phase of the beams to cause constructive interference results in an intensity maximization on the target. Systems of this type were described briefly at the beginning of this chapter.

For compensation of uniform atmospheric turbulence over a path length L, many zones of correction are required[73]

$$N_c = \left[\frac{2.01 C_n^2 L D^{5/3}}{\lambda^2 \ln(1/SR)} \right]^{6/5} , \qquad (2.126)$$

where SR is the desired Strehl ratio, D is the aperture diameter, and N_c is the number of correction zones.

Inverting, we find an expression for the Strehl ratio:

$$SR = \exp\left(\frac{-2.01 C_n^2 L D^{5/3}}{\lambda^2 N_c^{5/6}} \right) . \qquad (2.127)$$

To maintain stable control, the search process (the dither) must be much faster than the disturbance variation. Each element in the system attempts to drive

itself and the weighted sum of all others to an intensity maximum.[74] The closed-loop (CL) bandwidth of the atmospheric turbulence correction system must be

$$f_{CL} \cong 0.65 V \left[\left(\frac{2\pi}{\lambda} \right)^2 C_n^2 L \right]^{3/5} , \qquad (2.128)$$

where V is the transverse wind velocity that moves the turbulence across the beam. The range of dither frequencies[75] is

$$f_{max} - f_{min} = 3.2(N_c - 1)f_{CL} , \qquad (2.129)$$

and the maximum dither frequency should be

$$f_{max} = [10 + 1.6(N_c - 1)]f_{OL} , \qquad (2.130)$$

where f_{OL} is the open-loop unity-gain servo bandwidth. The total phase excursion for any channel is

$$\Delta\phi = \frac{3.58}{\lambda} (C_n^2 L D^{5/3})^{1/2} . \qquad (2.131)$$

For very long propagation distances the delay in the signal reaching the detector limits the control bandwidth. Even with lead/lag control compensation, the control signal phase delay must be kept below 30 or 40 deg. The phase delay is proportional to the propagation distance R divided by the speed of light,

$$\phi_{delay} = \left(\frac{2R}{c} \right) 2\pi f_{CL} . \qquad (2.132)$$

The range-bandwidth ($R f_{CL}$) product must remain below 2×10^7. This rules out atmospheric propagation from the ground to space (1000 km) using this method because the bandwidth achievable, 20 Hz, is well below the approximately 100-Hz correction bandwidth required.

2.6.5 Image Sharpening

The goal of most imaging systems is to obtain a "sharp" image. By adjusting the phase of different zones across the imaging system aperture one at a time until a sharp image is achieved the goal can be reached. This differs from multidither techniques because not all phase adjusters are moved simultaneously, nor are they dithered. The *sharpness function S* is calculated after each phase adjustment and the system is converged to maximum sharpness. The sharpness function that is maximized is not always a maximization of on-axis intensity. The integral of the square of the image distribution is commonly used

$$S = \int I(x,y)^2 \, dx \, dy \ , \tag{2.133}$$

as are higher moments of the distribution

$$S = \int x^p y^q I(x,y) \, dx \, dy \ , \tag{2.134}$$

and entropy minimization[5]

$$S = -\int I(x,y) \ln[I(x,y)] \, dx \, dy \ . \tag{2.135}$$

Beckers[76] suggests wavefront control algorithms that maximize the area of the pupil over which the rms wavefront distortion is less than ⅛ wave.

2.6.6   Anisoplanatism

To observe a star or extended object in space from a ground-based telescope, one must have a source of light (a beacon) of sufficient intensity to perform wavefront sensing. Less than 1% of the sky is observable with natural beacons or *guide stars* because they are either too faint or not close enough to the same atmospheric beam path as the object.

Because the atmosphere is not uniform, that is, the disturbances are not isotropic, the adaptive optics system will experience anisoplanatism. If there is significant difference in the path between the incoming wavefront from the guide star and the path between the telescope and the object, the adaptive optics system performance is degraded. The adaptive optics system simply cannot correct for abberations that it does not sense. When the guide star and the object are offset by an angle, the limiting angle, the *isoplanatic angle,* is given by

$$\theta_0 = 0.528 \left[\left(\frac{2\pi}{\lambda} \right)^2 \int_0^\infty dh \, C_n^2(h) h^{5/3} \right]^{-3/5} \ , \tag{2.136}$$

where C_n^2 is a function of altitude h. The Strehl ratio associated with the angular error between the object and the guide star θ is

$$\text{SR} = \exp[-(\theta/\theta_0)^{5/3}] \ . \tag{2.137}$$

If natural guide stars are too faint or displaced too far away from the object, artificial guide stars can be used. First proposed in the open literature by Foy and Labeyrie,[77] but preceded by U.S. government classified research,[78] artificial laser-induced guide stars have been demonstrated.[79] By scattering laser light off nitrogen and oxygen molecules in the atmosphere, an artificial Rayleigh guide star can be placed anywhere in the sky. By time gating the laser pulse and receiving electronics, the guide star can be restricted to a particular altitude. The atmosphere is sufficiently dense for this to occur below about 20 km altitude. Because the artificial beacon is not at infinity, as is the natural guide

star, the light from the beacon does not follow the same path as the light from a natural star. The artificial beacon light appears to come from a point at a different focal distance than a natural guide star; the result is *focal aniso-planatism*. To partially overcome this limitation, artificial guide stars have been produced by resonant sodium absorption in the mesosphere at altitudes between 90 and 100 km. A laser tuned to the sodium resonance at 589 nm can create a beacon with an intensity that is four orders of magnitude higher than a Rayleigh beacon with the same laser power. In addition to creating laser guide stars at higher altitudes, multiple laser guide stars have been created and the wavefront information "stitched together" to reduce anisoplanatic effects.[80]

The Strehl ratio reduction resulting from focal anisoplanatism is a function of the artificial guide altitude H, wavelength, zenith angle, and C_n^2. A linear curve fit based on the analysis by Fried and Belsher,[81] for three atmospheric turbulence models [Hufnagel-Valley 2/7, SLC-day (D), and SLC-night (N)], an imaging wavelength of 0.5 μm, and a zenith angle of 0 deg shows the Strehl dependence:

$$SR = \exp[-(D/d_0)^{5/3}] \; , \qquad\qquad (2.138)$$

where

$d_0 = 0.018H + 0.39$ (H-V model)
$d_0 = 0.041H + 0.29$ (SLC-D model)
$d_0 = 0.046H + 0.42$ (SLC-N model)
D = aperture diameter in meters
H = guide star altitude in kilometers.

The parameter d_0 represents the value of the largest aperture diameter where an artificial guide star adaptive optics system would provide compensation to about $\lambda/6.3$ rms wavefront error.

2.7 WAVEFRONT RECONSTRUCTION AND CONTROLS

An adaptive optics control system usually consists of multiple coupled loops between the wavefront sensor and the correction device. It is assumed that the control follows linear control laws. In many cases, the wavefront must be reconstructed from the wavefront sensor signals. When the reconstruction is explicit, there is a step in the control process where a representation of the wavefront exists. Some controls do not contain this step and directly translate signals from the wavefront sensor to drive signals for the correcting optics.

The wavefront sensor may output information in many forms. A modal sensor releases numbers that are proportional to the normal modes of the wavefront. A zonal sensor outputs numbers proportional to the slopes or magnitude of the wavefront in each zone.

The drive signals to the correction devices have various forms as well. The drive signals to a deformable mirror should be proportional to the amplitude of the influence function at each actuator. The drive signals to a beam-steering mirror are proportional to the tip or tilt of the mirror. The drive signals to an edge-actuated deformable mirror are related to the deformation modes of the

surface. The reconstruction problem is reduced to the fundamental steps of translating wavefront information to correction device drive information.

2.7.1   Solution of Linear Equations

In the wavefront reconstruction and control process, one usually encounters a set of M equations with N unknowns. They can be written discretely as

$$y_1 = a_1 B_{11} + a_2 B_{12} + a_3 B_{13} \cdots a_N B_{1N} \ ,$$

$$y_2 = a_1 B_{21} + a_2 B_{22} + a_3 B_{23} \cdots a_N B_{2N} \ ,$$

$$y_3 = a_1 B_{31} + a_2 B_{32} + a_3 B_{33} \cdots a_N B_{3N} \ , \qquad (2.139)$$

$$\vdots \qquad\qquad\qquad \vdots$$

$$y_M = a_1 B_{M1} + a_2 B_{M2} + a_3 B_{M3} \ldots a_N B_{MN} \ ,$$

or as a matrix equation

$$\mathbf{y} = \mathbf{Ba} \ . \qquad (2.140)$$

The vectors and coupling matrix are defined simply as

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} \qquad \mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} , \qquad (2.141)$$

$$\mathbf{B} = \begin{bmatrix} B_{11} & B_{12} & \ldots & B_{1N} \\ B_{21} & B_{22} & \ldots & B_{2N} \\ B_{31} & B_{32} & \ldots & B_{3N} \\ \vdots & & & \vdots \\ B_{M1} & B_{M2} & \ldots & B_{MN} \end{bmatrix} . \qquad (2.142)$$

The linearity requirement is satisfied when the parameters a_n are constants. The basis functions B_{mn} can have any functional form. They may be wave aberration polynomials or they may be functions representing the coupling between actuator voltages for a deformable mirror and the wavefront at certain points. The basis functions are not necessarily orthogonal or normalized. If this occurs, the inversion of $\mathbf{B}$ can be simplified.

In any case, the solution that is required is an inversion of Eq. (2.140). That is, we must find the parameters a_n that fit the series of basis functions B to the known or measured values y_m. When the number of equations M is less than the number of unknowns, it is an *underdetermined system* and cannot be solved uniquely. When $M = N$ the matrix $\mathbf{B}$ is square and, as long as it is not singular, it can be directly inverted to solve Eq. (2.140): $\mathbf{a} = \mathbf{B}^{-1}\mathbf{y}$. In adaptive optics control system problems one often encounters the case where the number

of equations M is greater than the number of unknowns N. This *overdetermined system* is the basis for much research in numerical techniques, many of which apply to adaptive optics systems.

The problem reduces to calculating the values of the unknowns such that the error is small between the measured, or known parameters y_m, and the actual value of y. An inverse for a nonsquare matrix cannot be calculated directly. An approximation can be determined based on a specific merit function.

The method of least squares finds the minimum of

$$\chi^2 = \sum_{i=1}^{M} \left(y_i - \sum_{k=1}^{N} a_k B_{ik} \right)^2 . \tag{2.143}$$

By differentiating χ^2 with respect to each element a_k and setting $d\chi^2/da = 0$, a set of equations for the a's can be found. In matrix notation, this is equivalent to solving for the vector $\mathbf{a}$ by

$$\mathbf{a} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{y} , \tag{2.144}$$

where $\mathbf{B}^T$ is the transpose of the matrix $\mathbf{B}$. Note that the matrix $\mathbf{B}$, for an overdetermined system, has M rows and N columns. The transpose has N rows and M columns, and the square invertable matrix $\mathbf{B}^T\mathbf{B}$ has M^2 elements. The matrix $(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$ is called the *pseudoinverse* of $\mathbf{B}$. This method, called a *Gauss reconstructor,* is the principal method employed in wavefront reconstruction and control. In some cases, a nonzero piston term in the wavefront reconstruction can cause the matrices to become singular. The effects of noise in the wavefront sensor signals also can cause mathematical difficulties. In these cases, other methods of finding the least-squares solution are used, such as the singular value decomposition[82] (SVD), Jacobi iteration,[83-85] a Gauss-Seidel iteration,[84] or the successive over relation.[84]

2.7.2 Calculating the Wavefront

Calculating Phase from Wavefront Slopes.   The most common adaptive optics reconstruction problem is the determination of the wavefront (OPD or phase) at various points from knowledge of the wavefront slopes at other points on the wavefront plane.[86] The most common wavefront sensors (Hartmann sensors or shearing interferometers) produce output signals proportional to wavefront slopes.

The number of unknown phase points is N and the number of wavefront slope measurements is M. When $M > N$, the system is overdetermined and solutions can be found either by direct least-squares methods, SVD, or other methods previously mentioned. The error between the calculated phase at the various points and the actual phase at those points (that produced the slope measurements) is important. It is not only the merit function for the least-squares solution but it is a physical limitation on the performance of the adaptive optics system.

The problem of phase determination from wavefront slopes depends on the geometry of the problem and various numerical techniques. The phase estimation at N locations is ϕ_n. The measured slopes at M locations are represented

by s^x_m or s^y_m. The superscripts represent the slope in the x or y direction. The phase is to be determined at the points $\phi_{-1,-1}$, $\phi_{-1,0}$, ..., where the subscripts represent the coordinates of the unknown points. A set of equations can be constructed

$$s^x_1 = \phi_{-1,1} - \phi_{0,1} \ ,$$
$$s^y_2 = \phi_{0,1} - \phi_{0,0} \ ,$$
$$\vdots \qquad \vdots \ .$$

These equations can be written in matrix form

$$\mathbf{s} = \mathbf{B}\phi \ , \qquad\qquad\qquad (2.145)$$

where the matrix $\mathbf{B}$ is the *reconstruction matrix*. Often, $\mathbf{B}$ is simply sums and differences based on the geometrical configuration of wavefront sensor positions and phase determination positions. Many of the elements are $+1$ or -1. In some cases, however, $\mathbf{B}$ can be quite complicated. A formal derivation by Wallner[87] uses weighting functions that are dependent on the subaperture and the direction of slope measurement in the subaperture. In actual wavefront sensors the slope measurement is seldom a simple difference between two wavefront points, but usually it is a spatial average over the subaperture. Accounting for noise and the spatial averaging least-squares methods or SVD can be used to invert $\mathbf{B}$ and reconstruct the phase from the measured slopes.

Singular matrices, which arise because of a nonunique solution to the set of equations, can be overcome by numerical methods. If the piston component of the wavefront is disregarded, the solution matrix $\mathbf{B}^{-1}$ can be singular, because there are unlimited solutions that would give the same slope measurements. Adding another row to the matrix $\mathbf{B}$ with each element equal to 1.0 forces the summation of all phase points to a particular value.[88] The addition of another value in the slope vector $\mathbf{s}$ specifies the piston value, usually set to 0. The fixed piston value allows inversion.

$$\begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_M \\ 0 \end{bmatrix} = \begin{bmatrix} \ [\mathbf{B}] \ \\ 1\,1\,1\,...\,1 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_N \end{bmatrix} \ . \qquad (2.146)$$

Calculating Wavefront Modes from Wavefront Slopes.

The problem of reconstructing the phase modes from wavefront slopes is similar to reconstruction of the phase values at various points. The wavefront phase ϕ is described by a polynomial expansion

$$\phi = \sum_{k=1}^{K} a_k Z_k(x,y) \ , \qquad\qquad (2.147)$$

where the coefficients are the a_k and the polynomial basis functions (modes) are $Z_k(x,y)$. A Hartmann or shearing interferometer wavefront sensor measures the slopes of the wavefront at various positions m. There are M total measurements. For illustration, it is assumed that half of the measurements are slopes in the x direction and half are in the y direction. The set of linear equations is

$$\left. \frac{\partial \phi}{\partial x} \right|_m = \sum_{k=1}^{K} a_k \left. \frac{\partial Z_k(x,y)}{\partial x} \right|_m , \tag{2.148}$$

$$\left. \frac{\partial \phi}{\partial y} \right|_m = \sum_{k=1}^{K} a_k \left. \frac{\partial Z_k(x,y)}{\partial y} \right|_m . \tag{2.149}$$

The subscript m implies evaluation at the coordinates (x_m, y_m). This set can be represented by the matrix equation $\mathbf{s} = \mathbf{Ba}$, where the vector $\mathbf{s}$ is the slope vector and $\mathbf{a}$ is the mode coefficient vector,

$$\mathbf{s} = \begin{bmatrix} \left. \frac{\partial \phi}{dx} \right|_1 \\ \left. \frac{\partial \phi}{dx} \right|_2 \\ \vdots \\ \left. \frac{\partial \phi}{dx} \right|_{M/2} \\ \left. \frac{\partial \phi}{dy} \right|_1 \\ \left. \frac{\partial \phi}{dy} \right|_2 \\ \vdots \\ \left. \frac{\partial \phi}{dy} \right|_{M/2} \end{bmatrix} \qquad \mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_K \end{bmatrix} , \tag{2.150}$$

whereas matrix $\mathbf{B}$ is

$$
\mathbf{B} = \begin{bmatrix}
\left.\dfrac{\partial Z(x,y)_1}{\partial x}\right|_1 & \left.\dfrac{\partial Z(x,y)_2}{\partial x}\right|_1 & \cdots & \left.\dfrac{\partial Z(x,y)_K}{\partial x}\right|_1 \\[2ex]
\left.\dfrac{\partial Z(x,y)_1}{\partial x}\right|_2 & \left.\dfrac{\partial Z(x,y)_2}{\partial x}\right|_2 & \cdots & \left.\dfrac{\partial Z(x,y)_K}{\partial x}\right|_2 \\[1ex]
\vdots & & & \vdots \\[1ex]
\left.\dfrac{\partial Z(x,y)_1}{\partial x}\right|_{M/2} & \left.\dfrac{\partial Z(x,y)_2}{\partial x}\right|_{M/2} & \cdots & \left.\dfrac{\partial Z(x,y)_K}{\partial x}\right|_{M/2} \\[2ex]
\left.\dfrac{\partial Z(x,y)_1}{\partial y}\right|_1 & \left.\dfrac{\partial Z(x,y)_2}{\partial y}\right|_1 & \cdots & \left.\dfrac{\partial Z(x,y)_K}{\partial y}\right|_1 \\[1ex]
\vdots & & & \vdots \\[1ex]
\left.\dfrac{\partial Z(\mathrm{x},\mathrm{y})_1}{\partial y}\right|_{M/2} & \left.\dfrac{\partial Z(\mathrm{x},\mathrm{y})_2}{\partial y}\right|_{M/2} & \cdots & \left.\dfrac{\partial Z(\mathrm{x},\mathrm{y})_K}{\partial y}\right|_{M/2}
\end{bmatrix} . \tag{2.151}
$$

The elements of **B** are the derivatives of the polynomial basis functions evaluated at the wavefront sensor subaperture positions. A number of basis sets can be used. There is a requirement that they be linearly independent[89] but it is not necessary that they are orthogonal.[90] And, naturally, there is no requirement for a particular coordinate system. Polar coordinates and slopes measured in the polar directions are equally valid and useful. As before, the solution is found by finding the pseudoinverse of **B**:

$$
\mathbf{a} = [(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T]\mathbf{s} . \tag{2.152}
$$

2.7.3   Calculating Deformable Mirror Drive Signals

Deformable Mirror Drive Signals from Continuous Phase.   Once a phase map or a modal representation is constructed, it is necessary to determine the application of that information for driving a wavefront corrector. A deformable mirror is assumed to be a linear system of electromechanical actuators that deform a mirror surface into a desired shape. A desired surface wavefront $S(x,y)$ is generated by applying the proper amplitude A_i to N actuators. The effect of each actuator on the surface $Z_i(x,y)$ is its influence function. A set of linear equations can be constructed

$$
S(x,y) = 2\cos\beta \sum_{i=1}^{N} A_i Z_i(x,y) , \tag{2.153}
$$

where β is the angle of incidence of the light beam on the mirror. This can be written in matrix form

$$
\mathbf{S} = \mathbf{ZA} . \tag{2.154}
$$

For a surface represented by K points the influence matrix **Z** is K rows and N columns. A pseudoinversion of **Z** is required:

$$\mathbf{A} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{S} \ . \tag{2.155}$$

The actuator commands are then derived from the actuator amplitude vector $\mathbf{A}$ by multiplying by the suitable gain factors and voltage conversions.

Constructing the $K \times N$ influence matrix $\mathbf{Z}$ is not trivial. The effect of each actuator on a point (x,y) in the wavefront plane must be determined. Although the influences often couple, they are assumed to be linear. The elements of $\mathbf{Z}$ can be calculated from theory or determined experimentally by activating each actuator one at a time and determining its response.

Deformable Mirror Drive Signals from a Modal Representation of the Phase.
The deformable mirror drive signals are derived by first inverting the linear sum of influence functions as before

$$S(x,y) = 2 \cos\beta \sum_{i=1}^{N} A_i Z_i(x,y) \ . \tag{2.156}$$

If the wavefront Φ is represented by a sum of modes with basis functions $B_m(x,y)$ and coefficients c_m,

$$\Phi(x,y) = \sum_{m=1}^{M} c_m B_m(x,y) \ , \tag{2.157}$$

the two summations can be equated:

$$S = \Phi = 2 \cos\beta \sum_{i=1}^{N} A_i Z_i(x,y) = \sum_{m=1}^{M} c_m B_m(x,y) \ . \tag{2.158}$$

Using matrix notation and absorbing the $2 \cos\beta$ factor into the $\mathbf{A}$ vector,

$$\mathbf{ZA} = \mathbf{Bc} \ , \tag{2.159}$$

and inverting, the vector of deformable mirror actuator commands is obtained:

$$\mathbf{A} = \mathbf{Z}^{-1} \mathbf{Bc} \ . \tag{2.160}$$

In most cases, $M \neq N$, and the inverse of $\mathbf{Z}$ is found by least-squares techniques.

Deformable Mirror Drive Signals from Wavefront Slopes.   Real-time adaptive optics systems often do not have a need for an explicit representation of the wavefront. In a closed-loop real-time system, the wavefront sensor information must be used efficiently. Calculating the deformable mirror signals directly from the vector of wavefront slopes is commonly used.

The measured slope vector $\mathbf{s}$ is given by

$$\mathbf{s} = \mathbf{B}\phi \ , \tag{2.161}$$

where $\mathbf{B}$ is the reconstruction matrix and ϕ is the wavefront. The phase is corrected by a surface S, which is a sum of actuator influence functions,

$$\phi = \mathbf{S} = \mathbf{ZA} \ . \tag{2.162}$$

The factor $2\cos\beta$ is absorbed into the $\mathbf{A}$ vector. Combining the two equations eliminates the central step of explicit phase representation

$$\mathbf{s} = \mathbf{BZA} \ . \tag{2.163}$$

Matrix $\mathbf{B}$ is $M \times P$, where M is the number of slope measurements and P is the number of phase points. Matrix $\mathbf{Z}$, the influence matrix, is $P \times J$, where J is the number of actuators. The matrix $\mathbf{BZ}$ is an $M \times J$ matrix. Pseudoinversion techniques can be used if it is overdetermined, i.e., the number of slopes M is greater than J. Inverting the matrix $\mathbf{BZ}$ can be done by multiplying the two and performing a pseudoinverse or by inverting each one

$$(\mathbf{BZ})^{-1} = \mathbf{Z}^{-1}\mathbf{B}^{-1} \ . \tag{2.164}$$

The result is a linear conversion of measured wavefront slopes to actuator drive signals.

2.7.4   Temporal Considerations

The delays associated with the matrix calculations appear to reduce the overall correction bandwidth of the adaptive optics system. Although the matrices can be quite large ($N > 1000$), the calculations are manageable. In many cases, the reconstruction matrices and the inverses are system-dependent and need to be calculated only once, prior to operation. The number of control signals needed is only as large as the number of actuators. The number of input signals is usually only slightly larger. Parallel processing is often used to minimize the time delay.

Where the effects of anisoplanatism appear as misinterpretation of the wavefront in a spatial sense, the time delay associated with the control (called the *servo lag*) appears as a misinterpretation of the wavefront in a temporal sense. Simply, if we wait too long to apply the correction, we will apply a wrong correction. The temporal spectrum for atmospheric turbulence is well characterized. The servo lag depends on the response of the atmosphere as well as the response of the electronics and computers.

Strehl Ratio for Servo Lag Error and Bandwidth Specification.   The servo lag error depends on the time constant of the atmosphere at the time the correction is to be applied. The temporal phase difference power spectrum[91] is

$$W_{\Delta d}(r,f) \to 0.065 f^{-8/3} k^2 \int_0^L dh C_n^2(h) v(h)^{5/3} \ , \tag{2.165}$$

where r is the spatial separation of points on the wavefront, L is the range, and f is the temporal frequency; $k = 2\pi/\lambda$.

Note here that this expression was derived for the case of negligible amplitude scintillation, or $f \ll v/(\lambda L)^{1/2}$ and $r \gg (\lambda L)^{1/2}$. If these inequalities

are reversed, then the expression for the power spectrum is simply multiplied[92] by a factor of ½.

Assuming that the corrector system has a high-frequency cutoff f_c the uncorrected portion of the power spectrum $W_{\Delta\phi}(f)$ is given by

$$\sigma_u^2 = \int_0^\infty |1 - H(f, f_c)|^2 W_{\Delta d}(f) \, df \, , \tag{2.166}$$

where we use the high-spatial-frequency zero segment diameter expression because, for atmospheric turbulence, f_c is greater than any of the low frequency break points.

Two choices of servo response function are used. The first is a sharp cutoff binary filter

$$H(f, f_c) = \begin{cases} 1 & f \le f_c \\ 0 & f > f_c \end{cases} . \tag{2.167}$$

The second is a slow roll-off RC filter

$$H(f, f_c) = (1 + if/f_c)^{-1} \, . \tag{2.168}$$

For the binary filter the variance becomes

$$\sigma_B^2 = 0.0196 k^2 \int_0^L dh \, C_n^2(h) v(h)^{5/3} f_B^{-5/3} \, , \tag{2.169}$$

which defines f_b. For the RC filter we have

$$\sigma_{RC}^2 = 0.102 k^2 \int_0^L dh \, C_n^2(h) v(h)^{5/3} f_{RC}^{-5/3} \, , \tag{2.170}$$

which defines f_{RC}. Solving for the cutoff frequencies in the two cases, we have

$$\begin{Bmatrix} f_B \\ f_{RC} \end{Bmatrix} = \left[\begin{Bmatrix} 0.0196 \\ 0.102 \end{Bmatrix} \left(\frac{k}{\sigma_u} \right)^2 \int_0^L C_n^2(h) v(h)^{5/3} \, dh \right]^{3/5} . \tag{2.171}$$

Note that the cutoff frequency (3-dB point) in the case of the RC filter is 2.7 times larger than the binary filter. We adopt a characteristic frequency from the RC case,

$$f_G = \left[0.102 (k/\sigma_u)^2 \int_0^L C_n^2(h) v(h)^{5/3} \, dh \right]^{3/5} , \tag{2.172}$$

where the subscript G stands for Greenwood, and this frequency has come to be known as the *Greenwood frequency*.

To calculate f_G one must have atmospheric data or a model for $C_n^2(h)$ and $v(h)$. A popular choice for this pair are the Hufnagel-Valley model of $C_n^2(h)$ [Eq. (2.19)] and the Bufton wind model[17]

$$v(h) = 2.5 + 18.5 \exp\left[-\left(\frac{h\sin\theta_Z - 12}{5}\right)^2\right],\tag{2.173}$$

where $v(h)$ is in meters per second and θ_Z is the zenith angle. For example, using these choices for $C_n^2(h)$ and $v(h)$, the Greenwood frequency [for $\sigma_u^2 = (\lambda/10)^2$] is 32 Hz.

We may want to know what servo bandwidth is required to achieve a servo lag Strehl ratio of 0.95 for a corrector system operating in an atmosphere that is adequately described by the Hufnagel-Valley and Bufton models. Using the formula $\text{SR} = \exp(-f_G/f_s)^2$, we have the servo bandwidth

$$f_s = \frac{f_G}{(-\log\text{SR})^{1/2}} = 141 \text{ Hz}.\tag{2.174}$$

Servo Lag Error. Now that we have an expression for the Greenwood frequency, we know the Strehl ratio resulting from operating with a servo bandwidth with cutoff frequency f_s.

$$\text{SR} = \exp\left[-\left(\frac{f_G}{f_s}\right)^2 - 2.5\left(\frac{\omega}{4\pi f_s\theta_0}\right)^{5/3}\right],\tag{2.175}$$

where f_G is the Greenwood frequency, f_s is the servo bandwidth, θ_0 is the isoplanatic angle, and ω is the slew rate of the laser beam or the tracker-imager.

2.8   WAVEFRONT CORRECTION

A critical element in an adaptive optics system is the wavefront corrector. The correction to the wavefront can be accomplished by changing either the differential velocity of the wave across the wavefront or by changing the differential optical path length. The former method requires the alteration of the refractive index of the medium through which the wave is passing. These methods, including nonlinear optical phase conjugation, are called *unconventional adaptive optics*.

Conventional wavefront correctors can be modal (tilt or focus mirrors) or zonal (segmented or continuous faceplate deformable mirrors). The construction and operation of these devices is covered in detail elsewhere.[44] We cover the basic performance characteristics that affect the overall utility of the correctors. The most widely applied method alters the optical path difference of the wave using reflective optics in the form of a continuous faceplate deformable mirror.

2.8.1 Deformable Mirrors

A generic representation of a corrector element, in this case a deformable mirror, is shown in Fig. 2.33. The mirror is constructed with a very thin faceplate that is capable of deformation normal to its surface by the application of pushing and pulling action from an array of actuators working against a rigid reaction structure. The actuator is typically a piezoelectric, electrostrictive, or magnetostrictive material that undergoes elongation when an electromagnetic field is applied. The piezoelectric devices typically require application of 2 to 3 kV to achieve strokes of a few micrometers. Corresponding values are a few hundred volts for the electrostrictive devices and 5 to 100 V at relatively high currents for the magnetostrictive materials. The actuators have been miniaturized to allow for as little as 0.5-cm spacing between actuators.[93,94] The deformable facesheets have been constructed from metals (copper, beryllium, molybdenum) and glass (Pyrex, ULE, Zerodur). High-reflectivity coatings, either metallics or multilayer dielectrics, are employed to achieve wavelength-specific performance.

The tasks that deformable mirrors are asked to perform fall into two categories. Low temporal frequency and long-stroke mirrors are used for active figure control of large mirrors. These systems correct for such things as gravitational and thermal loading and seismic jitter. High temporal frequency and short-stroke mirrors compensate for atmospheric turbulence.

When specifying the performance requirements of an adaptive optics deformable mirror, one has to deal with residual errors after correction is applied. This is represented by *fitting error*. A deformable mirror with a discrete set of corrector points cannot exactly represent a continuously deformed surface.

Strehl Ratio for Fitting Error. The active mirror is made up of N discrete actuators. Following Hudgin,[85] let $R_j(x,y)$ be the influence function for the j'th actuator; i.e., it is the phase correction for a unit applied signal. If S_j is the applied signal, then the total applied phase correction is
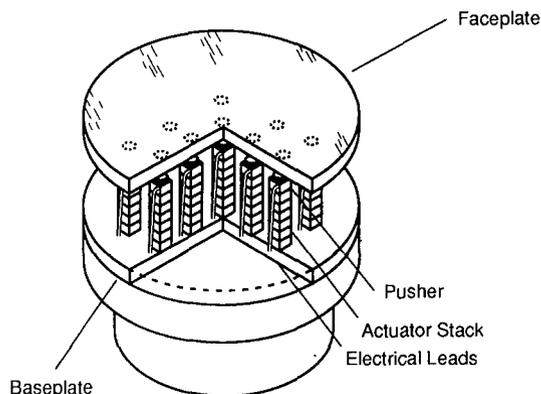
Fig. 2.33 Continuous faceplate deformable mirror.

$$\phi_c(x,y) = \sum_{j=1}^{N} S_j R_j(x,y) \ . \tag{2.176}$$

If $\phi(x,y)$ is the known phase front (e.g., focus) to be matched by the active mirror, then the error variance is

$$\sigma_{\text{fit}}^2 = \frac{1}{A} \int_{\text{Aper}} dx \, dy [\phi(x,y) - \phi_c(x,y)]^2 \ , \tag{2.177}$$

where A is a mirror area. The problem is to choose S_j to minimize σ_{fit}^2 for a given influence function $R_j(x,y)$. Hudgin shows that this is solved by

$$S_i = \sum_{j=1}^{N} C_{jk}^{-1} B_k \ , \tag{2.178}$$

where

$$C_{jk} = \frac{1}{A} \int dx \, dy R_j(x,y) \ , \tag{2.179}$$

and

$$B_k = \frac{1}{A} \int dx \, dy \phi(x,y) R_k(x,y) \ . \tag{2.180}$$

See also Sec. 2.7 and Eq. (2.153). The Strehl ratio is given by

$$SR = \exp[-(2\pi\sigma_{\text{fit}})^2] \ . \tag{2.181}$$

The fitting error is then calculated for specified forms of $R_J(x,y)$ and $\phi(x,y)$. Two influence functions were chosen: (1) a linear interpolation pyramid and (2) a fit to an Itek piezoelectric-actuated mirror. The fit to four low-order errors is given in the following table:

1 Wave rms Error in	rms Fit (waves)	
	Pyramid	Itek Mirror
Tilt	0.0123	0.0338
Focus	0.0407	0.1180
Astigmatism	0.0254	0.0740
Coma	0.0295	0.0745

For example, one wave rms of tilt error can be converted to a residual error of 0.0123 waves rms by a pyramid influence function in a square array of actuators on a circular mirror.

To illustrate, consider a deformable mirror designed to correct for low-order aberrations caused by gain medium aberrations from a gas laser. At equilib-

rium they are measured to be tilt, $\lambda/6$; focus, $\lambda/8$; astigmatism, $\lambda/10$; and coma, $\lambda/15$. What is the Strehl ratio due to the rms fitting error of an Itek mirror? From the table we have

$$\sigma^2_{\text{tilt}} = \left(\frac{0.0338}{6}\right)^2 \text{ waves}^2 \quad \text{SR} = 0.9987 \ ,$$

$$\sigma^2_{\text{focus}} = \left(\frac{0.1180}{8}\right)^2 \text{ waves}^2 \quad \text{SR} = 0.9914 \ ,$$

$$\sigma^2_{\text{astig}} = \left(\frac{0.0740}{10}\right)^2 \text{ waves}^2 \quad \text{SR} = 0.9978 \ ,$$

$$\sigma^2_{\text{coma}} = \left(\frac{0.0745}{15}\right)^2 \text{ waves}^2 \quad \text{SR} = 0.9990 \ .$$

Fitting a Random Phase Front.   When using the mirror to correct for an ensemble of random wavefronts such as would be the case for propagation through atmospheric turbulence, a statistical approach is needed. Hudgin shows that Eq. (2.177), the error variance in the case of atmospheric propagation of the wavefront that is minimized by voltages [Eq. (2.178)] applied to the actuators, becomes

$$\sigma^2_{\text{fit}|\text{min}} = \frac{1}{2A^2} \int dd \ dy \ dx' \ dy' D(r)$$

$$+ \frac{1}{2A^2} \sum_{jk} C^{-1}_{jk} \int dx \ dy \ dx' \ dy' R_j(x,y) R_k(x',y') D(r) \ , \qquad (2.182)$$

where $D(r) = 6.88 r^{5/3}$ for Kolmogorov turbulence. Thus, given the influence functions $R_j(x,y)$ and the structure function $D(r)$, one can evaluate the minimum error variance over the mirror surface after correction with voltage S_i given by Eq. (2.178).

When one substitutes various influence functions, the expression for error variance has the form

$$\sigma^2_{\text{fit}|\text{min}} = \alpha_i \left(\frac{d}{r_0}\right)^{5/3} \ , \qquad (2.183)$$

where d is the interactuator spacing, α_i depends on the specific influence function $R_i(x,y)$ and r_0 is the coherence length.

Hudgin presented curves of $\sigma^2_{\text{fit}|\text{min}}$ versus $(d/r_0)^{5/3}$ for four influence functions. The slopes of the lines give the so-called "fitting error coefficients" α_i. We can therefore create a table of these coefficients for various mirror models. We included the results of other calculations for completeness:

Mirror Fitting Coefficient	Reference
0.229	Gaussian[85]
0.284	Pyramid[85]
0.399	Itek mirror[85]
1.297	Piston only[85]
0.319	Tyson (Gaussian)[95]
0.141	Greenwood and Fried[91]
0.134	Noll[27]
0.130	Fried[26]

Note that, for the cases in Refs. 26, 27, and 91, the error variance assumes independent actuators with no influence between individual subaperatures. Clearly, this assumption underestimates the fitting error.

For example, suppose one desires to compensate for atmospheric turbulence using a deformable mirror of the type built by Itek. Assume a site that has a range of r_0 values varying between 5 and 10 cm for various atmospheric conditions and zenith angles. For a mirror with interactuator spacing of 3 cm, we can find the Strehl ratio reduction for mirror fitting error.

From the table, we see that the mirror fitting coefficient for the Itek mirror is 0.399, thus the residual error will be

$$\text{SR} = \exp\left[-0.399\left(\frac{d}{r_0}\right)^{5/3} \right] , \tag{2.184}$$

with $5 < r_0 < 10$ cm, or

$$0.843 < \text{SR} < 0.948 . \tag{2.185}$$

2.9 SUMMARY

Table 2.2 gives a summary of the Strehl ratio formulas discussed in this chapter.

Table 2.2   Summary of Strehl Ratio Formulas

Effect	Strehl Formula
BASIC STREHL RELATIONSHIP:	
Strehl relationship	2.5
ATMOSPHERIC TURBULENCE:	
Angle of arrival variance	2.31
Anisoplanatism (angular)	2.137
Anisoplanatism (focal)	2.138
Beam spread	2.28
Deformable mirror fitting	2.181
Higher order aberrations	2.103
Log-amplitude variance, plane wave	2.20
Log-amplitude variance, spherical wave	2.21
Log-amplitude variance, saturation regime	2.22
Servo bandwidth	2.175
Tilt	2.40
THERMAL BLOOMING:	
Small scale blooming, no wind	2.73
Small scale blooming, wind fluctuations	2.75
Uplink whole beam blooming, corrected	2.57
Uplink whole beam blooming, uncorrected	2.54
Whole beam blooming, uncorrected	2.47
SENSORS AND OPTICS:	
Hartmann sensor	2.125
Jitter	2.88
Laser resonators	Section 2.4.5
Manufacturing error, power loss	2.96
Multidither correction	2.127
Sampling	2.99
Segmented optics	2.90
Shearing interferometer	2.119
Thermal distortion	2.94
Tilt sensing	2.112
COMBINED EFFECTS:	
Turbulence, jitter, and whole beam blooming	2.76 – 2.87

References

1. J. W. Hardy, "Active optics: a new technology for the control of light," *Proceedings of the IEEE* **66,** 651 (1978).

2. W. T. Cathey, C. L. Hayes, W. C. Davis, and V. F. Pizzurro, "Compensation for atmospheric phase effects at 10.6 μm," *Applied Optics* **9,** 701 (1970).

3. J. E. Pearson, W. B. Bridges, S. Hansen, T. A. Nussmeier, and M. E. Pedinoff, "Coherent optical adaptive techniques: design and performance of an 18-element, visible, multi-dither COAT system," *Applied Optics* **13,** 611 (1976).

4. *New Technologies for Astronomy, Proceedings of the SPIE* **1130** (1989).

5. R. A. Muller and A. Buffington, "Real-time correction of atmospherically degraded telescope images through image sharpening," *Journal of the Optical Society of America* **64,** 1200 (1974).

6. A. C. Claus, "On Archimedes burning glass," *Applied Optics* **12,** A14 (1973).

7. O. N. Stavroudis, "Comments on: on Archimedes burning glass," *Applied Optics* **12,** A16 (1973).

8. H. W. Babcock, *Astrophysical Journal* **107,** 73 (1948).

9. H. W. Babcock, "The possibility of compensating astronomical seeing," *Astronomy Society of the Pacific* **65,** 229 (1953).

10. H. W. Babcock, "Deformable optical elements with feedback," *Journal of the Optical Society of America* **48,** 500 (1958).

11. J. W. Hardy, J. Feinleib, J. C. Wyant, "Real time phase correction of optical imaging systems," in *Topical Meeting on Optical Propagation through Turbulence,* Optical Society of America (1974).

12. M. Born and E. Wolf, *Principles of Optics,* 5th ed., Pergamon Press, Oxford (1975).

13. A. N. Kolmogorov, "The local structure of turbulence in incompressible viscous fluid for very large Reynolds numbers," *Doklady Adademii Nauk SSSR* **30,** 301 (1941).

14. V. I. Tatarskii, *Wave Propagation in a Turbulent Medium,* McGraw-Hill, New York (1961).

15. V. I. Tatarskii, *The Effects of the Turbulent Atmosphere on Wave Propagation,* National Technical Information Service, Washington, D.C. (1971).

16. R. E. Hufnagel, *Proceedings of the OSA Topical Meeting on Optical Propagation Through Turbulence,* Boulder, CO (1974).

17. J. L. Bufton, "Comparison of vertical profile turbulence structure with stellar observations," *Applied Optics* **12,** 1785 (1973).

18. G. C. Valley, "Isoplanatic degradation of tilt correction and short-term imaging system," *Applied Optics* **19,** 574 (1980).

19. P. B. Ulrich, "Hufnagel-Valley profiles for specified values of the coherence length and isoplanatic patch angle," WJSA/MA/TN-88-013, W. J. Schafer Associates (1988).

20. M. E. Gracheva and A. S. Gurvich, *Radiophysics and Quantum Electronics* **8,** 511 (1965).

21. H. T. Yura, "Physical model for strong-amplitude fluctuations in a turbulent medium," *Journal of the Optical Society of America* **64,** 59 (1974).

22. R. F. Lutomirski and H. T. Yura, "Propagation of a finite optical beam in an inhomogeneous medium," *Applied Optics* **10,** 1652 (1971).

23. A. I. Kon and V. I. Tatarskii, *Izvestiya Vysshikh Uchebnykh Zavedenii, Radiofizika* **8,** 870 (1965).

24. D. L. Fried, "Propagation of a spherical wave in a turbulent medium," *Proceedings of the IEEE* **55,** 57 (1967).

25. D. L. Fried, "Optical resolution through a randomly inhomogeneous medium for very long and very short exposures," *Journal of the Optical Society of America* **56,** 1372 (1966).

26. D. L. Fried, "Statistics of a geometric representation of wavefront distortion," *Journal of the Optical Society of America* **55,** 1427 (1965).

27. R. J. Noll, "Zernike polynomials and atmospheric turbulence," *Journal of the Optical Society of America* **66,** 207 (1976).

28. R. J. Sasiela, "A unified approach to electromagnetic wave propagation in turbulence and the evaluation of multiparameter integrals," Report 807, MIT/Lincoln Laboratory (July 1988).

29. F. G. Gebhardt, "High power laser propagation," *Applied Optics* **15,** 1479 (1976).

30. F. G. Gebhardt and D. C. Smith, "Self-induced thermal distortion in the near field for a laser beam in a moving medium," *IEEE Journal of Quantum Electronics* **QE-7,** 63 (1971).

31. B. Edlen, *Metrologia* **2,** 71 (1966).

32. R. Penndorf, *Journal of the Optical Society of America* **47,** 176 (1957).

33. D. C. Smith, "High power laser propagation: thermal blooming," *Proceedings of the IEEE* **65,** 1679 (1977).

34. F. G. Gebhardt, "Twenty-five years of thermal blooming: an overview," *Proceedings of the SPIE* **1221,** 2 (1990).

35. J. Herrmann, "Properties of phase conjugate optical systems," *Journal of the Optical Society of America* **67,** 290 (1977).

36. P. B. Ulrich and L. E. Wilson, Eds., *Propagation of High-Energy Laser Beams through the Earth's Atmosphere, Proceedings of the SPIE* **1221** (1990).

37. J. W. Goodman, *Introduction to Fourier Optics,* McGraw-Hill, New York (1968).

38. V. B. Krapchev, "Atmospheric thermal blooming and beam clearing by aerosol vaporization," *Proceedings of the SPIE* **1221,** 91 (1990).

39. R. K. Szeto and D. L. Fried, The Optical Sciences Company Reports TR-1076, TR-1083, TR-1088, tOSC, Placentia, CA.

40. H. Breaux, W. Evers, R. Sepucha, and C. Whitney, *Applied Optics* **18,** 2638 (1979).

41. C. B. Hogge, J. F. Schultz, D. B. Mason, and W. E. Thompson, "Physical optics of multiaperture systems," *Applied Optics* **27,** 5127 (1988).

42. J. B. Shellan, "Phased-array performance degradation due to mirror misfigures, piston errors, jitter, and polarization errors," *Journal of the Optical Society of America A* **2,** 555 (1985).

43. W. P. Barnes, "Optical Materials—Reflective," in *Applied Optics and Optical Engineering,* Vol. VII, R. R. Shannon and J. C. Wyant, Eds., Academic Press, New York (1979).

44. R. K. Tyson, *Principles of Adaptive Optics,* Academic Press, Boston (1991).

45. J. S. Fender, Ed., *Synthetic Aperture Systems, Proceedings of the SPIE* **440** (1984).

46. V. N. Sintzov and A. F. Zapryagaev, *Soviet Physics-Vspekhi* **17** (1975).

47. J. M. Elson, H. E. Bennett, and J. M. Bennett, "Scattering from optical surfaces," Chap. 7 in *Applied Optics and Optical Engineering,* Vol. VII, R. R. Shannon, J. C. Wyant, Eds., Academic Press, New York (1979).

48. J. O. Porteus, *Journal of the Optical Society of America* **53,** 1394 (1963).

49. G. P. Agrawal and M. Lax, *Journal of the Optical Society of America* **69** (1979).

50. D. Anafi, J. M. Spinhirne, R. H. Freeman, and K. E. Oughstun, "Intracavity adaptive optics: tilt correction performance," *Applied Optics* **20,** 1926 (1981).

51. K. E. Oughstun, "Intracavity compensation of quadratic phase aberrations," *Journal of the Optical Society of America* **72,** 1529 (1982).

52. W. F. Krupke and W. R. Sooy, "Properties of an unstable confocal resonator CO_2 laser system," *IEEE Journal of Quantum Electronics* **QE-5,** 575 (1969).

53. R. C. Harney, "Active laser resonator control techniques," *Applied Optics* **17,** 1671 (1978).

54. A. J. DeMaria, "Review of cw high-power CO_2 lasers," *Proceedings of the IEEE* **61,** 731 (1973).

55. C. B. Hogge and R. R. Butts, AFWL-TR-78-15, U.S. Air Force Weapons Laboratory (1978).

56. L. C. Marquet, *Applied Optics* **10,** 4 (1971).

57. C. Chi, N. B. James, III, and P. L. Misuinas, "Spectral shared aperture component," *Proceedings of the SPIE* **171** (1979).

58. J. Winocur, *Applied Optics* **21,** 3 (1982).

59. G. A. Tyler and D. L. Fried, *Journal of the Optical Society of America* **72,** 6 (1982).

60. W. J. Bates, *Proceedings of the Physical Society* **59,** 940 (1947).

61. J. C. Wyant, "Double frequency grating lateral shear interferometer," *Applied Optics* **12,** 2057 (1973).

62. J. C. Wyant, "White light extended source shearing interferometer," *Applied Optics* **13,** 200 (1974).

63. O. Y. Kwon, "Real-time radial shear interferometer," *Proceedings of the SPIE* **551,** 32 (1985).

64. J. C. Wyant, "Use of an ac heterodyne lateral shear interferometer with real-time wavefront correction systems," *Applied Optics* **14,** 2622 (1975).

65. J. W. Hardy, "Instrumental limitations in adaptive optics for astronomy," *Proceedings of the SPIE* **1114,** 2 (1989).

66. M. Elbaum and P. Diament, "SNR in photocounting images of rough objects in partially coherent light," *Applied Optics* 15, 2268 (1976).

67. J. Feinleib, "Sensor system for detecting wavefront distortion in a return beam of light," U. S. Patent 4,141,652 (1979).

68. J. K. Bowker, "Pulsed laser wavefront sensor," Adaptive Optics Association Technical Report 78-1, Cambridge, MA (1978).

69. C. A. Lors, "I-cubed sensor wavefront correction system," Internal Memorandum, Pratt and Whitney Aircraft (Aug. 28, 1978).

70. J. Allen, B. Baran, L. Schmutz, and J. Yorsz, "Speckle and adaptive optics techniques," *Photonics Spectra*, 97 (Aug. 1988).

71. F. Roddier, "Curvature sensing and compensation: a new concept in adaptive optics," *Applied Optics* 27, 1223 (1988).

72. F. Roddier, J. E. Graves, and E. Limburg, "Seeing monitor based on wavefront curvature sensing," *Proceedings of the SPIE* 1236, 475 (1990).

73. J. E. Pearson and S. Hansen, "Experimental studies of a deformable mirror adaptive optical system," *Journal of the Optical Society of America* 67, 325 (1977).

74. T. R. O'Meara, "The multidither principle in adaptive optics," *Journal of the Optical Society of America* 67, 306 (1977).

75. J. E. Pearson, S. A. Kokorowski, and M. E. Pedinoff, "Effects of speckle in adaptive optical systems," *Journal of the Optical Society of America* 66, 1261 (1976).

76. J. M. Beckers, "Optimization of partial adaptive optics," *Applied Optics* 31, 424 (1992).

77. R. Foy, A. Labeyrie, "Feasibility of adaptive telescope with laser probe," *Astronomy and Astrophysics* 152, L29 (1985).

78. R. Q. Fugate, D. L. Fried, G. A. Ameer, B. R. Boeke, S. L. Browne, P. H. Roberts, R. E. Ruane, G. A. Tyler, and L. M. Wopat,"Measurement of atmospheric distortion using scattered light from a laser guide star," *Nature* 353, 144 (1991).

79. C. A. Primmerman, D. V. Murphy, D. A. Page, B. G. Zollars, and H. T. Barclay, "Compensation of atmospheric optical distortion using a synthetic beacon," *Nature* 353, 141 (1991).

80. D. V. Murphy, D. A. Primmerman, B. G. Zollars, and H. T. Barclay, "Experimental demonstration of atmospheric compensation using multiple synthetic beacons," *Optics Letters* 16, 1797 (1991).

81. D. L. Fried and J. F. Belsher,"Analysis of artificial guide-star adaptive optics system performance for astronomical imaging," Report TR-1157, The Optical Sciences Company, Placentia, CA (1991).

82. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes,* Cambridge University Press, Cambridge (1989).

83. D. L. Fried, "Least-square fitting of a wavefront distortion estimate to an array of phase-difference measurements," *Journal of the Optical Society of America* 67, 370 (1977).

84. B. R. Hunt, "Matrix formulation of the reconstruction of phase values from phase differences," *Journal of the Optical Society of America* 69, 393 (1979).

85. R. H. Hudgin, "Optimal wave-front estimation," *Journal of the Optical Society of America* 67, 378 (1977).

86. R. J. Noll, "Phase estimates from slope-type wave-front sensors," *Journal of the Optical Society of America* 68, 139 (1978).

87. E. P. Wallner, "Optimal wave-front correction using slope measurements," *Journal of the Optical Society of America* 73, 1771 (1983).

88. R. J. Sasiela and J. G. Mooney, "An optical phase reconstructor based on using a multiplier-accumulator approach," *Proceedings of the SPIE* 551, 170 (1985).

89. R. Cubalchini, "Modal wave-front estimation from phase difference measurements," *Journal of the Optical Society of America* 69, 972 (1979).

90. W. H. Southwell, "Wavefront analyzer using a maximum likelihood algorithm," *Journal of the Optical Society of America* 67, 396 (1977).

91. D. P. Greenwood and D. L. Fried, "Power spectra requirements for wave-front compensations systems," *Journal of the Optical Society of America* 66, 193 (1976).

92. D. P. Greenwood, "Bandwidth specification for adaptive optical systems," *Journal of the Optical Society of America* 67, 390 (1977).

93. M. A. Ealey and J. F. Washeba, "Continuous facesheet low voltage deformable mirrors," *Optical Engineering* **29,** 1191 (1990).

94. A. McGovern, Litton/Itek Optical Systems, Lexington, MA, personal communication (1990).

95. R. K. Tyson, "Adaptive optics system performance approximations for atmospheric turbulence correction," *Optical Engineering* **29,** 1165 (1990).

Bibliography

Bareket, N., and C. L. Koliopoulos, Eds., *Wavefront Sensing, Proceedings of the SPIE* **351** (1982).

Becherer, R. J., and B. A. Horwitz, Eds., *Adaptive Optics Systems and Technology, Proceedings of the SPIE* **365** (1982).

Fugate, R. Q., Ed., *Proceedings of the Laser Guide Star Adaptive Optics Workshop,* Phillips Laboratory, Kirtland AFB, NM (1992).

Hardy, J. W., "Active optics: a new technology for the control of light," *Proceedings of the IEEE* **66**(6), 651–697 (1978).

Hardy, J. W., "Instrumental limitations in adaptive optics for astronomy," *Proceedings of the SPIE* **1114**(2), 2–13 (1989).

Ludman, J. E., Ed., *Adaptive Optics, Proceedings of the SPIE* **551** (1985).

Pearson, J. E., R. H. Freeman, and H. C. Reynolds, Jr., "Adaptive optical techniques for wavefront correction," Chap. 8 in *Applied Optics and Optical Engineering,* Vol. VII, R. R. Shannon, J. C. Wyant, Eds., Academic Press, New York (1979).

Roddier, F. J., Ed., *Active Telescope Systems, Proceedings of the SPIE* **1114** (1989).

Special Issue on Adaptive Optics, *Journal of the Optical Society of America* **67**(3) (March 1977).

Special Issue on Adaptive Optics, *MIT Lincoln Laboratory Journal,* **5**(1) (1992).

Tyson, R. K., *Principles of Adaptive Optics,* Academic Press, Boston (1991).

Tyson, R. K. and J. Schulte in den Bäumen, Eds., *Adaptive Optics and Optical Structures, Proceedings of the SPIE* **1271** (1990).

Ulrich, P. B. and L. E. Wilson, Eds., *Propagation of High-Energy Laser Beams Through the Earth's Atmosphere, Proceedings of the SPIE* **1221** (1990).

Sensor and Data Fusion

Alan N. Steinberg
The Analytic Sciences Corporation
Arlington, Virginia

CONTENTS

3.1   INTRODUCTION

> The person who owns a watch always knows what time it is; someone who owns
> two watches never does.

This chapter discusses the architecture issues and techniques involved in combining multisource data and coordinating multiple sensors to enhance system effectiveness.

The integrated design and operation of multisensor systems offer the potential for increased effectiveness and reliability, with possible savings in total system weight and cost, but involves considerations of complementary sensor capabilities, mutual interference, and opportunities for shared assets.

Combining data from multiple sources can enhance measurement accuracies or situation awareness but is subject to issues of sensor alignment, data association, combination of incommensurate features, and accounting for correlated and uncorrelated measurement errors between sensors.

3.1.1   Terminology

Although usage of the principal terms relating to sensor data fusion is far from consistent in the literature, there has been an important attempt at terminology standardization conducted by the Department of Defense (DOD) Joint Directors of Laboratories/Data Fusion Subpanel[1] (JDL/DFS) and the associated Office of Naval Technology (ONT) Data Fusion Development Strategy Panel.[2] We attempt to follow this standard, while citing alternative uses commonly encountered. Table 3.1 lists the symbols and nomenclature used.

In its broadest sense, *sensor fusion* is the integrated use and design of multiple sensors, often referred to as *multisensor integration* (MSI). Such integration is generally for purposes of achieving either better performance or reduced penalties (e.g., cost or weight) than can be obtained with the collection of sensors, each used separately.

An important branch of sensor fusion is *sensor data fusion:* combining, or "fusing," the data from multiple sensors and/or other information sources to obtain an integrated, and possibly more accurate, assessment of the sensed environment. In addition to data fusion, sensor fusion (or MSI) includes improving the interoperability of sensors and designing multifunction sensors and multiple sensor systems that share resources.

Sensor data fusion functions include multisensor target detection, data association, tracking and identification, situation assessment, and situation prediction. For convenience, a *target* is taken to be any object that is subject to detection, identification, or tracking by the use of sensor data. Hence, e.g., cloud formations, weather fronts, and terrain features may be targets of a remote earth observation sensor system.

Because there is no consensus concerning the use of the terms *feature, attribute,* and *characteristic,* we use these interchangeably to indicate nonkinematic aspects of a target that can be measured or inferred. A *parameter* is a quantifiable feature dimension. A target *state vector* includes both kinematics and features at a given moment in time.

The Tri-Service Data Fusion Symposium (DFS) standard uses *identification* (ID) in the broad sense to encompass both the unique individuation of objects ("serial number ID") and classification by type or category (e.g., F-15E, F-15,

Table 3.1   Symbols and Nomenclature

Symbols	Nomenclature
$E[x]$	Expectation (mean or first moment) $:= \bar{x} := \int x\rho(x)\,dx$
$E[x^n]$	nth (noncentral) moment $:= \int x^n \rho(x)\,dx$
$\mathbf{I}$	Identity matrix
M_E	Entropic confidence measure
M_G	Global information measure
M_L	Local information measure
$m(H)$	Evidential mass (Dempster-Shafer)
$\mathbf{P}(k) := \mathbf{P}(k\|k)$	State error covariance matrix $:= E\{[\mathbf{x}(k) - \hat{\mathbf{x}}(k\|k)][\mathbf{x}(k) - \hat{\mathbf{x}}(k\|k)]^T \| Z_k\}$
$\mathbf{P}(k+1\|k)$	State prediction covariance
$P\{\alpha\}$	Probability of event α
P_D	Probability of detection
P_{FA}	Probability of false alarm
$\mathbf{P}_{i,j}(k)$	Cross-covariance between track state estimates i and j at time t_k
r	Target range
$\mathbf{S}(k + 1)$	Measurement prediction covariance $:= E[\bar{\mathbf{z}}(k + 1\|k)\bar{\mathbf{z}}^T(k + 1\|k)\|Z_k]$
$\text{var}(x)$	Variance (second central moment) $:= E[(x - \bar{x})^2] = \int (x - \bar{x})^2 \rho(x)\,dx := \sigma_x^2$
$\mathbf{W}(k)$	Kalman filter gain
$\mathbf{x}(k)$	True target state at time t_k
$\hat{\mathbf{x}}(k) := \hat{\mathbf{x}}(k\|k)$	Estimated target state at time t_k $:= E[\mathbf{x}(k)\|Z_k]$
$\hat{\mathbf{x}}(k + 1\|k)$	Target state prediction
$\hat{\mathbf{z}}(k + 1\|k)$	Predicted measurement
$\bar{\mathbf{z}}(k + 1\|k)$	Measurement prediction error $:= \mathbf{z}(k + 1) - \bar{\mathbf{z}}(k + 1\|k)$
Z_k	Measurement set $\{\mathbf{z}(j), j + 1, ..., k\}$
$^*\mathbf{z}_{i,j}(k)$	Event that measurement (estimate) $\mathbf{z}_i(k)$ is caused by target i
$\mu A(x)$	Degree of membership of x in set A (Zadeh)
$\nu(k + 1)$	Innovation (measurement residual) $:= \bar{\mathbf{z}}(k + 1\|k) := \mathbf{z}(k + 1) - \hat{\mathbf{z}}(k + 1\|k)$
$\rho_x(y)$	Probability density function $\rho_x(y)\,dy := P\{y < x \leq y + dy\} \geq 0$
$\rho(x)$	$:= \rho_x(x)$
σ_x	Standard deviation $:= \text{var}(x)^{1/2}$

fighter aircraft, multiengine jet aircraft, etc.). *Target recognition* is identification in terms of friend, foe, or neutral. *Target characterization* is distinguished from parameter (or feature) *estimation* in that the former is typically an inferred description or labeling rather than a measured value. Generally speaking, characterization is the result of *data* processing, rather than *signal* processing.

3.1.2 Multisensor Integration

The degree of integration of sensor and associated processing functions is an important consideration for the designer of a multisensor system. Richman[3] lists four degrees of multisensor integration, shown in Table 3.2. At one extreme are nonintegrated systems in which individual sensors are installed and employed separately, with no physical, functional, or data integration. At the other extreme are concepts in which the entire sensor suite is designed as a unified system, with maximum physical and processing integration.

Figure 3.1 indicates some of the benefits of integrated approaches to system design. In particular, the use of multiple information sources has resulted in improved system performance in detecting, recognizing, and tracking targets. The adaptive use of diverse sensors can also increase a sensor system's robustness; i.e., its ability to achieve mission objectives when the ordinary use of specific sensors is precluded, whether by equipment failure, adverse reception conditions (resulting from natural causes or intentional countermeasures), or mission restrictions (e.g., the desire to limit for reasons of covertness or electromagnetic interference).

Integration of data processing and control allows multisensor systems to be *opportunistic:* adapting so as to make maximum use of information as available, often being able to operate adequately with noisy or missing data or with failures in individual subsystems.

Table 3.2   Degrees of Multisensor Integration (from Ref. 3)

DEGREE	TYPE OF INTEGRATION	CHARACTERISTICS	EXAMPLE
0	Coincidental	• Separate sensors are "bolted together" on a single platform • Each sensor functions independently • Data from ine sensor is not used in concert with data from other sensors	• Navigation radar • Night vision goggles
1	Signal	• Separate sensors are "bolted together" on a single platform • Data from one sensor is combined with data from other sensors, or used to control the operation of other sensors	• F-14 Radar and TCS
2	Physical	• Sensors suite is designed as an entity • Sensors are either physically collocated, or means are provided to determine the relative sensor orientations • Apertures are shared where possible • Data from one sensor is combined with data from other sensors, or used to control the operation of other sensors	• E-O fire control system (FLIR, DVO, & LRF) • TADS, Apache, LANTIRN, F-16 • Physical integration +AI ≈ gets close to fusion
3	Fusion	• Analysis of data from one sensor is affected by data from other sensors, and vice versa • Data processing exhibits gestalt:"The whole is greater than the sum of the parts" • Physical integration is essential	• Human being
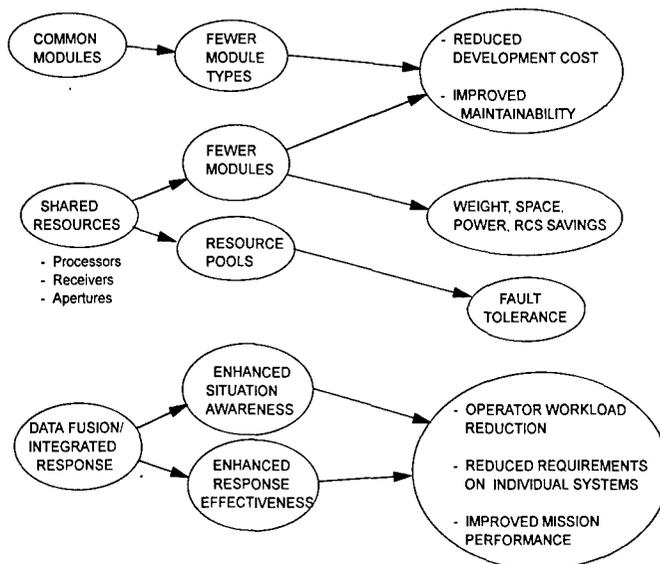
Fig. 3.1   Benefits of multisensor integration.

Additionally, by allocating sensing functions among different sensor types, it is often possible to reduce the required capability of individual sensor, with possible attendant reduction in technology risk and cost. For example, a passive tracking sensor [e.g., an infrared search and track (IRST) or electronic warfare support measures (ESM) system] is inherently more suited to tracking targets in angle than in range. An active sensor [radar or laser radar (LADAR)] can provide complementary measurements. An aircraft IRST can covertly scan a region for threats, identifying potential target tracks. The fire control radar can then be cued to the bearing of a target of interest, requiring only a minimal pulse burst to measure target range and range rate for a complete fire control solution. Once a target is in track, the radar can be used most efficiently, providing update information as required with minimum use of active radiation.

Data fusion issues to be addressed in this chapter are

1. *Data association:* determining whether data from multiple sources relate to the same object (target)
2. *Multisource ID/track estimation:* assuming that reports from multiple sources relate to the same object, combining such data to improve the estimate of the object, or of the overall current or future situation
3. *Collection management:* given a state of knowledge about the sensor environment, allocating multiple information acquisition and processing assets to maximize the utility while minimizing operational costs (such as processing or power loads, interference between functions, or exploitation by hostile systems).

More general issues in multisensor integration to be discussed are

1. *Sensor suite specification:* selecting and allocating performance requirements among elements of a multisensor system
2. *Data processing architecture specification:* achieving required availability of data within system bandwidth, processing and memory limitations
3. *Test and evaluation of fusion systems:* determining criteria and testing means for evaluating the benefits achieved by multisensor integration.

3.1.3 Data Fusion

The U.S. DOD Joint Directors of Laboratories Data Fusion Subpanel (henceforth DFS), has defined *data fusion* as

> a multi-faceted process dealing with the association, correlation, and combination of data and information from multiple sources to achieve refined position and identity estimation, and complete and timely assessments of situations and threats.[1]

> The product of the data fusion process in general is a dynamic integrated situation representation.[4]

It is useful for more general applicability to broaden the notion of *threat assessment* in this definition to that of *event prediction,* to include both natural and purposeful events.

Data fusion has been identified since 1987 by the Congressional Armed Services Committee as one of 21 technologies, for which development is critical to U.S. defense:

> The increasing complexity and speed of warfare, when combined with rapid advances in computer and communications technologies, are forcing a more effective integration of multiple sensors and diverse sources of information. Data fusion is the process of associating, correlating and combining data and information from multiple sensors and sources to achieve refined position and identity estimation and complete timely assessment of situations and threats and their significance.

> Data fusion is an integral part of all military systems from simple weapons to the most complicated large-scale information processing systems. It is essential that the military establishment develop a better understanding of the process of data fusion and a better set of tools to incorporate data fusion in emerging systems.[5]

Figure 3.2 depicts a few of the many applications in which multisensor data fusion plays a significant role. As will be seen, different techniques are applicable to cases such as (a) and (b), involving associating and exploiting the product of similar sensors, as contrasted with cases such as (c) in which sensors differ in their measurement dimensions. Other applications, such as depicted in (d), involve sensors operating in series: sensors with wide fields of view and high target capacities are used to cue low-capacity, high-measurement-quality sensors. As seen in the illustrated examples, a cueing series may involve several stages, each reducing the residual uncertainty in the prior sensors' target state estimation.

An extensive listing of current data fusion systems with U.S. defense applications can be found in Llinas.[6]
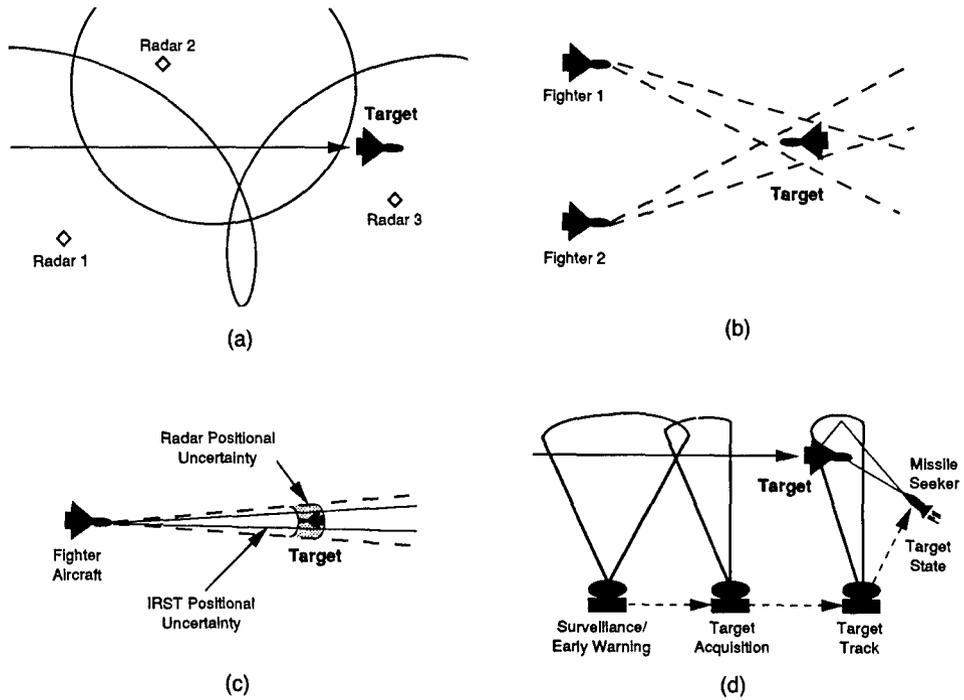
Fig. 3.2   Application examples of multisensor data fusion: (a) similar sensors, air traffic control; (b) similar sensors, internetted fighters; (c) complementary sensors, automatic target recognition; and (d) sequential cueing, air defense.

3.2   DATA FUSION MODELS

Sensor data fusion is a particular case of the general problem of combining information. The relationships of sensor data fusion to other types of multi-source information processing and to single-source (or single-sensor) data processing are treated in this subsection.

3.2.1   Single-Source Data Processing

Sensors or other information sources can be modeled in classical detection and estimation theoretic terms. As shown in Fig. 3.3, a mapping function $F[E(t)]$ transforms received energy $E(t)$ from a K-dimensional space to the sensor's N-dimensional observation space. Process noise $n(t)$ is added to the observed parameters, resulting in the observation vector $\mathbf{X}(t) = F[E(t)] + n(t)$. This received signal is then subject to three processes: detection, estimation, and classification.[7]

   Detection is the process of testing of the observed signal against one or more hypotheses. The observation space is effectively partitioned into discrete regions, one for each hypothesis; the partitioning effects a decision rule.

   In the simplest case, detection reduces to the binary decision of determining whether a target is present (hypothesis H_1) or not (the null hypothesis H_0). The probability density function (PDF) of an observation, $\rho[\mathbf{z}(t)|H_j]$, is called the *likelihood function* of the hypothesis H_j. Given a set of observations X,
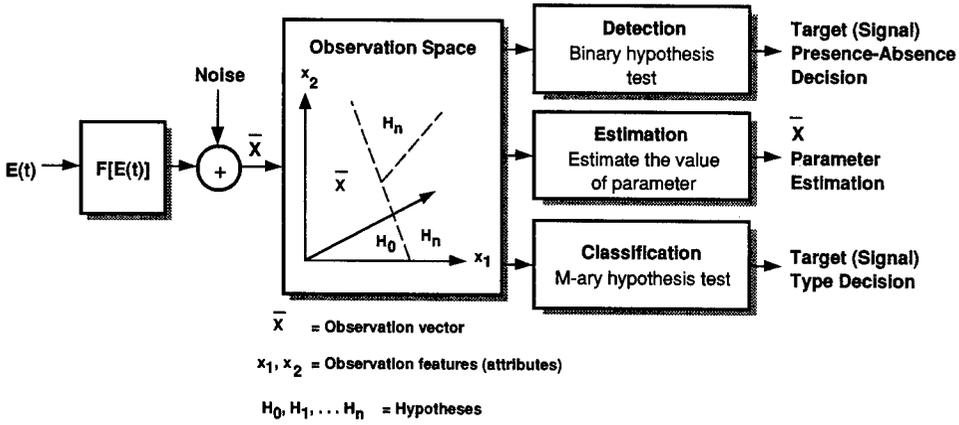
Fig. 3.3   Model of sensor functions.[7]

whose PDFs conditional on H_1 and H_0 are known, the probabilities for the two possible types of errors are defined as follows.

Type 1 error (false alarm):

$$P(\hat{H}_1|H_0) := P(\text{accept } H_1|H_0 \text{ true}) := P_{\text{FA}} \ . \tag{3.1}$$

Type 2 error (missed detection):

$$P(\hat{H}_0|H_1) := P(\text{accept } H_0|H_1 \text{ true}) := (1 - P_D) \ , \tag{3.2}$$

where P_D is probability of detection and P_{FA} probability of false alarm.

According to the Neyman-Pearson criterion, the optimal decision—in the sense of maximum P_D (minimum probability of Type 1 error) for a given allowable P_{FA}—employs the following rule:

$$\Lambda(\mathbf{z}) = \frac{\rho(\mathbf{z}|H_1)}{\rho(\mathbf{z}|H_0)} \begin{array}{c} > \\ < \end{array} \Lambda_0 \begin{array}{c} \Rightarrow \hat{H}_1 \\ \Rightarrow \hat{H}_0 \end{array} \ . \tag{3.3}$$

The decision threshold Λ_0 is set to a desired level of significance of P_{FA}:

$$P[\Lambda(\mathbf{z}) > \Lambda_0|H_0] = P_{\text{FA}} \ . \tag{3.4}$$

The null hypothesis is rejected only if its significance is below the level[8] P_{FA}.

Estimation is the process of determining the value (in some optimal sense) of the parameter $E(t)$ based on the measurement variable $R(t)$. *Classification* is a type of multiple hypothesis testing, in which $R(t)$ is compared against M disjoint hypotheses (classes) $H_0, H_1, ..., H_M$. In a decision-theoretic sense, clas-

sification is a subset of detection. The present distinction is useful primarily in reflecting the different feature spaces employed by conventional sensors in the two processes: detection processing occurs prior to, and classification after, the estimation of signal parameters.

Important trade-offs involve the allocation of these functions within a multisensor processing architecture. Some or all of these functions may be performed either within the individual sensors or in a central processing function.

In relatively distributed fusion systems, each sensor performs detection, estimation, and classification functions. It passes processed track data for association and refinement in a central fusion process.

Alternatively, of detection, estimation, and classification, any or all can be performed on a multisensor basis. At an extreme, raw data measurements (e.g., predetection pixel-level data) are passed from multiple sensors to a fusion node, which performs detection declarations on the basis of the composite data. Assuming that the errors in the various contributing sensors are statistically independent, such a scheme permits detection at SNRs below those that could be processed in single-sensor architectures, because single-sensor false alarms would not correlate between sensors.

The characteristics and relative merits of various data fusion architectures are discussed in Sec. 3.3.

3.2.2 Single- and Multisource Data Association

Multisensor data fusion can be viewed as an extension of the data association problem in processing of single-sensor data sampled over time. In the latter case, a sequence of data measurements (or reports) from the sensor must be associated to ensure that they refer to the same sensed objects (or targets).

If it is decided that two reports refer to the same target, then the next problem is combining the data to form an improved estimate of the target state.

Data association, track filtering, and identification are basic functions in multisensor data fusion as well. The particular problems involved in associating and combining measurements from *multiple* sensors relate to issues of sensor alignment, calibration, and error correlation of commensurate data between sensors and establishing cross-sensor classification templates from noncommensurate measurements.

In fusing commensurate data, multiple measurements of the same parameter set are combined to improve the accuracy or resolution of that parameter set. For example, a high-resolution radar and multiband imaging optical sensor can each estimate a variety of features of a given target. Some of these can be mutually commensurate, either directly or by coordinate transform, such as kinematic track parameters and object spatial dimension, and can be directly combined by filtering techniques to resolve these particular features.

Other parameters are mutually noncommensurate, such as emissivity in various IR bands and radar pulse repetition frequency (PRF) or optical and radar cross sections. Such measurements can be used to evaluate nodes in a classification hierarchy, allowing resolution of higher level features.

We can distinguish a *commensurate fusion* process as one that employs multisource data to estimate the value for a specific node in a (detection or

classification) hypothesis hierarchy. In contrast, a *complementary fusion* process attempts to resolve the value for a parent node based on the estimated values for various subordinate nodes. In the air traffic control and internetted fighter applications depicted in Figs. 3.2(a) and 3.2(b), respectively, the problem is primarily that of commensurate fusion; whereas the ATR example of Fig. 3.2(c) is an example of the fusion of complementary measurements.

3.2.3   Multisource and Multisensor Data Fusion

Goodman[9–12] presents a model of data fusion in terms of a generalized theory of combination of evidence. In this model, data fusion is a process which occurs within decision-making nodes within the context of a command, control, and communication (C^3) system.

In Ref. 12, a generic C^3 system is identified as a network of node complexes of decision makers, human or automated, that interacts with an external structure of friendly and adversarial nodes. As illustrated in Fig. 3.4, each node receives *signals* (i.e., observable effects of external or internal system events), which are modeled as stacked vectors comprised of incoming data from several different nodes. In turn, each node, which may consist of a single decision maker or some coalition of decision makers, processes the data.

This is followed by a response or action taken toward other nodes, within or external to the C^3 system. Associated with each node is (1) a node state, (2) a knowledge base representing its local knowledge of other nodes, and (3) an internal signal processing design. Data fusion plays the central signal processing role in transmitting detected signals to hypotheses formulations, which in turn through algorithm selection leads to an output response to other (friendly or adversary) nodes.

Goodman defines network transactions in terms of primitive relations among the following variables describing a C^3 system: nodes N, hypothesis selections H, detection D of incoming signals S, algorithm selections F, initial node responses R, prior to environmental distortion G, and additive noise Q. To each variable are affixed subscripts for nodes g and h of the C^3 system and discrete time index t_k.

The primitive *intranodal* relations are evaluated in terms of the conditional probabilities on the predecessor relations in the figure; e.g., response = $p(R_{g,k+1}|F_{g,k},S_{g,k},N_{g,k})$.

The primitive *internodal* relationships are the probabilities of nodal selection and the signal and noise/distortion at each node internodal interface. The
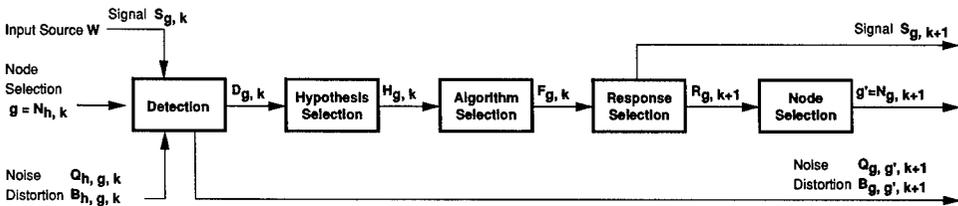
Fig. 3.4   Generic C^3 functional model based on Goodman.[12]

basic internodal analysis is developed via the additive nonlinear regression of relation

$$[S_{g,k+1}|W_{g,k+1} = (h,k)] = G_{h,g,k+1}(R_{h,k}) + Q_{h,g,k+1} \, , \tag{3.5}$$

where variable $W_{g,k+1}$ indicates an original possible source node for a signal at time t_k, given[9] reception by another node at t_{k+1}.

This structure allows that p need not represent ordinary statistical probability but can represent other evaluations, such as heuristic likelihood or possibility.[11] Approaches to uncertainty applicable to data fusion are discussed in Sec. 3.3.4.

The specific problem of sensor data fusion is sufficiently distinct to warrant specific system concepts and techniques. Unlike cases in which data is fused from multiple static information sources (e.g., documents), sensor data fusion is usually characterized by (1) noisy data, possibly with ill-defined statistical properties; (2) nonaligned measurements, i.e., not expressed in a common spatiotemporal coordinate system, and possibly with uncalibrated biases; and (3) constraints on processing and data access resources, allowed processing time, and human operator availability.

3.2.4   DFS Model for Sensor Data Fusion

A standard model for sensor data fusion has been developed by the U.S. DOD Joint Directors of Laboratories/Data Fusion Subpanel (DFS). This panel was established in 1986 as a subpanel to the JDL Technical Panel for C^3:

> The purpose of the DFS is to enhance the efficiency of individual and Joint Service data fusion programs through the exchange of technical information, the enhancement of coordination and cooperation between related research efforts, and by initiating new multi-service cooperative research programs and technology demonstrations.[4]

A version of the DFS functional model for data fusion is shown in Fig. 3.5. As seen, the model distinguishes three "levels" of data fusion.

Level 1 fusion processing consists of single- and multisource processing to develop and maintain estimates of individual objects (targets) in the sensed environment.

Level 2 assesses the set of object estimates and their relationships with one another and with the background environment (e.g., the disposition and movement of forces) to assess the total situation; e.g., an enemy's order of battle (OB). Thus, level 2 processing characterizes higher level structures, organizations, and activity on the basis of spatial, temporal, and parametric relationships and patterns involving level 1 targets and the environment.

Level 3 evaluates the assessed situation in terms of a system's prioritized set of objectives (e.g., those of a military force employing the fusion system for tactical battle planning). At this level, the fusion system estimates the capabilities and intent of the objects and object ensembles to affect or to be affected by the system. In military or intelligence fusion systems, the principal focus of level 3 processing is *threat assessment*. This involves estimating intents, means, and plans of hostile forces. As noted, the more inclusive concept is that of *event prediction*, including both natural and purposeful events.
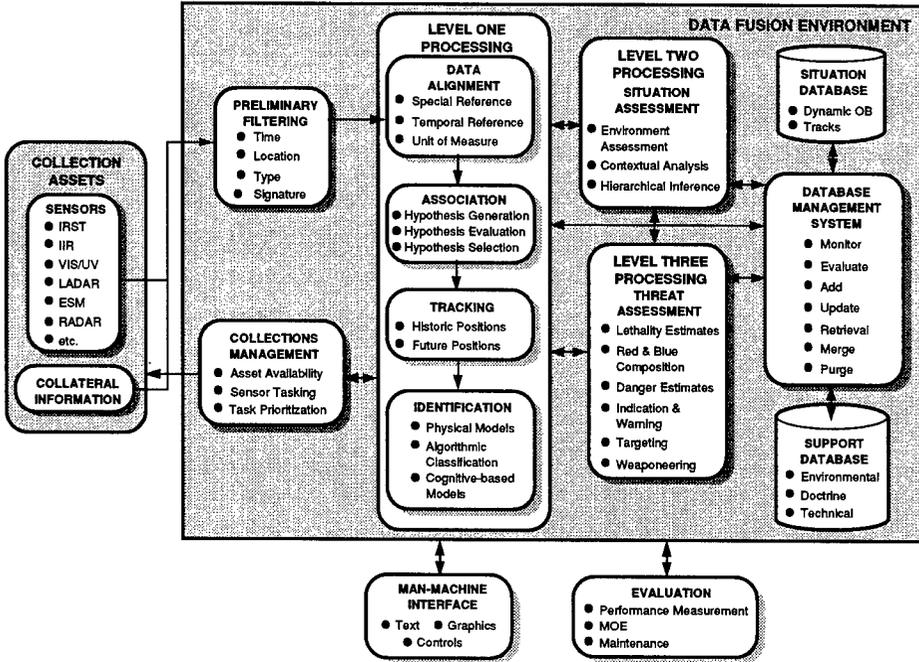
Fig. 3.5   DFS functional model for data fusion.[4]

The level 1 functions tend to involve primarily numeric processing: tracking, correlation, alignment, and association. The higher level functions are generally characterized by symbolic or cognitive types of processing.

A fourth fusion level—variously called process refinement or collection management, as in Fig. 3.5—is sometimes added to this model.[2] This is designed to encompass the dynamic control of sensor and system processes (to include the level 1 to 3 fusion processes) through feedback-based refinements. Referring to collection management as a level of the data fusion process is somewhat anomalous in that the defined levels 1 to 3 involve *estimation* processes, whereas collection management is an integrated *control* process.

To capture the desired flexibility in applying data at each of these levels to specific real-world situations, the DFS panel modified the hierarchical scheme represented in Fig. 3.5 to the open structure shown in Fig. 3.6. In the latter model, based on a knowledge-based *blackboard* architecture, information products can be passed between fusion nodes both within and between processing levels. This permits any of the processing functions to exploit not only its local data base but also those maintained anywhere else in the structure. Thus, a high-level situation assessment process would have the opportunity to access raw sensor data directly, if necessary.

Because this chapter focuses on sensors and their system applications, we concentrate on those fusion functions that directly involve the integration of sensors: level 1 fusion functions and the intimately associated sensor processing and collection management (level 4) functions. Detailed treatment of situation and threat assessment (levels 2 and 3) is beyond the scope of this chapter,
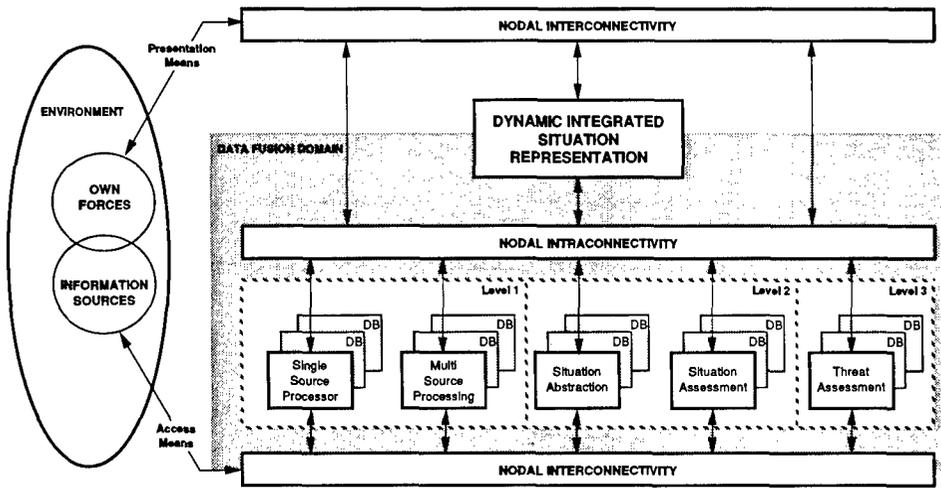
Fig. 3.6 DFS data fusion open architecture model.[4]
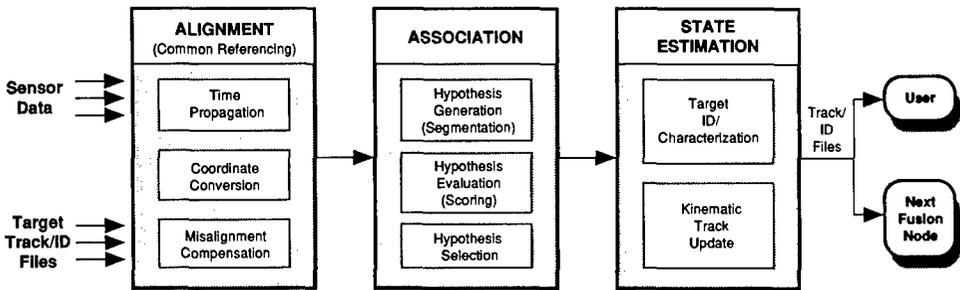
Fig. 3.7 Level 1 fusion node functions.[13]

being nonspecific as to information source (i.e., they are equally applicable to single source systems and to non-real-time information assessment systems as to real-time multisensor systems).

Functions performed in a level 1 fusion node are shown in Fig. 3.7 (following concepts from Ref. 13 and elsewhere). These include

1. *Alignment:* transforming positional/kinematic data from each sensor or other source into a common spatio-temporal coordinate system
2. *Association:* creating multisource fused target reports by generating, evaluating, and selecting hypotheses that sets of sensor reports relate to the same sensed objects
3. *State estimation:* updating the classification/attributive characterization and kinematic state estimation for fused target tracks and combining the multisource information to obtain an improved estimate.

In addition to these level 1 data fusion functions, we discuss the intimately associated (level 4) control function:

4. *Collection management:* coordinating the use of sensors with other mission assets to reduce situational uncertainty, contentions for system assets or detrimental side-effects of system actions.

3.3   ARCHITECTURES

This section discusses several issues relating to the design of realizable data fusion systems. These include factors in the selection of sensors and processing structures to best exploit the opportunities that fusion provides. Additional issues involve the processing burdens imposed by fusion. The number of possible data associations increases exponentially with the numbers of reports or tracks to be associated. Methods for constraining the number of association tests or for reducing the processing burden of each test are discussed in this section.

3.3.1   Sensor Suite Selection and Specification

Available sensors can differ in their capabilities of detecting, characterizing, and tracking specific target types; in the effects of ambient viewing conditions; and in the susceptibility to detection and exploitation by hostile systems. Such factors must be considered in selecting suites of sensors, in employing sensors in-mission, and in utilizing their output.

Requirements for integrating multiple sensors are driven by deficiencies in nonfused alternative architectures. These may include cases (after Harney[14]) where

1. No individual sensor can perform all required functions, or where imposing such requirements on an individual sensor would place excessive performance demands, with attendant unacceptable system cost, risk, weight, or support requirements. There are cases where the resolution achievable from single-sensor identification is inadequate to support required decision making. For example, target recognition beyond visual ranges typically must employ a variety of measurement techniques involving several types of sensors to establish target identities with sufficient confidence to justify weapons release.
2. No individual sensor can provide all required functions under all conditions. For example, some sensors' performance is degraded in low-light conditions, precipitation, atmospheric obscurants, etc. Additionally, operational constraints may restrict the use of particular preferred sensors, requiring the use of supplementary sensors. For example, an aircraft employing stealth techniques, or an air defense system that is susceptible to attack by antiradiation missiles, will restrict its use of radar, demanding increased exploitation of passive sensors, e.g., for range measurement for fire control.
3. Multiple sensors provide graceful performance degradation in the event of component failures.

4. Multiple sensors of different types provide increased resistance to countermeasures.
5. Multiple sensors of different types performing the same function may have improved performance, due to noncorrelated measurement errors.

Table 3.3 compares the performance capabilities and limitations of six major sensor types with applicability to automatic target recognition (ATR) in a combat aircraft.

A common goal of many sensor fusion efforts is to improve target detection, recognition (or identification), or tracking by combining measurements obtained from two or more existing sensors. Figure 3.8 indicates contributions that various aircraft sensors can make to target ID, prioritization, and fire control. This is an example of a hypothetical sensor suite that is synergistic at the feature or decision level; the measurements or reports from multiple sensors are required to develop the required ATR/fire control data.

Harney[14] lists several characteristics of an ideal target recognition feature (which could be a vector involving several measurement dimensions). Such a feature would be

1. *Definable:* amenable to numerical estimation
2. *Discriminatory:* different target classes should exhibit different numerical values of the feature
3. *Universal:* the value of the feature should be invariant for all operating states, environmental conditions and target orientations
4. *Robust:* the feature should be such that its value cannot be readily altered, made unobservable or mimicked
5. *Predictable:* the feature should be susceptible to modeling, allowing estimation when limited target information is available.

Table 3.3   Representative Sensor Performance Matrix for Fighter Aircraft (adapted from Ref. 15)

	FLIR	IRST	LASER RADAR	RADAR	SAR	ESM (RWR/RFI)
Spectral Region	8-12μm	3-5μm	10.6μm	9-11μm	600MHz-60GHz	1-300GHz
Sensor Type	Passive	Passive	Active	Active	Active	Passive
Targeting Function	Detect, Track, Classify, ID	Search, Detect Track, Classify	Detect, Track Classify, ID, MTI	Search, Detect, Track, Classify	Search, Detect, Classify, ID, MTI	Detect, Track, Classify, ID
Processing Type	2-D Imaging	2-D Imaging	3-D Imaging Range, Intensity, Doppler	Non-Imaging Monopulse Polarimetric	Imaging	Non-Imaging
Sensor Frame Rate	1/30 sec./frame 256x256 Pixels	----	10 to 20 sec./ frame 128x128 pixels	<10°/Sec. 1-3° Beam Width 10-20K pulses/sec.	1-50 sec./frame Depending on range	---
Detection Range	<10Km Wide FOV <20 Km Med. FOV <100 Km Nar. FOV	<370Km	<10 Km	<100 Km	<300 Km Depends on power	<200Km
Limitations	Limited by precipitation. Short range search. Large, heavy unit.	Limited by precipitation. Poor air-to-air ranging.	Limited by precipitation and obscurants. Narrow POV requires cueing. Short range.	Detectable by ESM, subject to ECM.	Detectable by ESM, subject to ECM.	Requires radiating target. Poor in-air ranging.

SENSOR	TARGET OR THREAT ID	TARGET OR THREAT PRIORITY	FIRE CONTROL SUPPORT
RWR/RFI	THREAT TYPE	THREAT OPERATIONAL MODE	BEYOND VISUAL RANGE TARGET ID
RADAR		AIRCRAFT RANGE & TIME-TO-GO	TRACK & MISSILE GUIDANCE
IRST	------------	AIRCRAFT & MISSILE TRACK	COVERT TRACK & RAID COUNT
IFF	FRIEND/FOE	--------	FRIEND/FOE
COOP. *	TRACK FILE HAND-OFF	PASSIVE LOCATION	COORDINATED RESPONSE

| MULTISENSOR FUSION | • Object Type
• Friend/Foe
• Raid Count | • Mode
• Lethality
• Event Prediction | • BVR Covert FC
• Coordinated Response |

* Internetted sensor platform (e.g. other fighter or AWACS)

Fig. 3.8   Sensor and sensor fusion functions in a fighter aircraft.

Because, in many applications, this combination of ideal characteristics is not to be had in any single measurement dimension, it is desirable to assemble a composite sensor capability that allows the generation of feature vectors having these qualities. For example, it is not necessary that each measurable parameter be discriminatory, in the sense of sufficing to isolate a particular target class of interest; it is important that the system have a high *a priori* likelihood of measuring feature vectors (intersections of parameters) that discriminate high-interest classes.

The selection of a multisensor suite for a target recognition system can be seen as an attempt to attain this set of characteristics at a system (or feature-vector) level.

Conversely, the selection of sensors with complementary characteristics can reduce the performance requirements (and usually the cost and complexity) for each individual sensor. For example, in several applications, one sensor is used to cue another, distributing the target search and measurement roles compatible with the natural capabilities of different sensor types.

Often cueing is from a surveillance sensor (or early warning or *traffic* sensor) to a target tracking (or *precision*) sensor. The former is characterized by a high probability of detection provided either by a wide instantaneous field of view (e.g., wide field of view staring optical sensor or radar warning receiver) or a high scan rate over a large field of regard (e.g., IRST). Such systems typically are able to handle a large number of targets, although with relatively poor tracking or parametric resolution. Hand-off is to a sensor with greater measurement capability (for example, because of a higher update rate or more sophisticated processing), but at the expense of reduced traffic handling ca-

pability, because of narrow field of view [such as a scanning passive or active electro-optic (E-O) system] and often longer dwell times. See Secs. 3.8.2 and 3.8.3 on sensor cueing design considerations.

3.3.2 Centralized Versus Distributed Fusion Processing

Many systems (to include most single- and multiplatform military systems) are implemented as a hierarchical command, control, and communications (C^3) structure. As such, there are several design decisions regarding the disposition of fusion nodes in the hierarchy.

As noted, several of the principal data fusion functions—namely, report to track association, classification, and track state estimation—are traditional functions in single-sensor data processing systems. Accordingly, there are a number of architectural alternatives for allocating these functions between individual sensors and centralized data fusion processes. Richman[3] lists advantages and disadvantages of centralized and distributed architectures, as shown in Table 3.4.

Table 3.4 Centralized Versus Distributed Sensor Data Processing (from Ref. 3)

ARCHITECTURE	ADVANTAGES	DISADVANTAGES
Centralized	• All data is available to central processor • Smaller variety of standardized processing elements is possible • Choices of sensor location on platform may be less limited • Processor environment is more easily controlled, making qualification easier • Processor maintainability is enhanced by having all processing elements in a common accessible location • Software changes are easier to incorporate into fielded units	• Specialized data buses may be required • Hardware improvements or additions may be difficult • Vulnerability is increased by having all processing resources in one location • Partitioning (limiting the effects of a failure to the subsystem in which it occurred) is difficult • Software development and maintenance are more difficult, as changes relevant to one sensor may affect many parts of the code
Distributed	• Processing elements may be tailored to each sensor, providing improved performance • Existing platform data buses (typically low-speed) may frequently be used • Partitioning is less difficult • Addition of new sensors or modification of old sensors has less impact on system hardware and software	Limited data supplied to central processor may diminish effectiveness of sensor fusion • Severe environment encountered by some sensors may limit choice of processor components and increase cost • Sensor location choices may be more limited • Increased variety of processing elements decreases maintainability, increases logistics support burden, and increases costs

3.3.2.1 Architecture Paradigms. Candidate architecture paradigms, illustrated in Fig. 3.9, include various approaches for distributing data association functions at processing nodes associated with individual sensors or centralizing them for cross-sensor association.

    A. *Fully Distributed, Nonfused Architectures:* In these, each sensor associates its own data into single-source track files, identifies and estimates track states for independent reporting to the higher level situation/threat assessment functions. In many simple systems of this sort, these higher level functions are typically performed by a human operator trying to assimilate multiple data streams as presented on one or more displays.

    B. *Maximally Distributed Fusion Architectures:* In these, single-source association, ID, and tracking are performed independently by each sensor, with subsequent alignment and multisource association by a central fusion process. The latter processes the multisource track files to refine ID and location/kinematics estimation.

        This architecture is the paradigm for *decision level* data fusion. It has the advantage of permitting simple sensor development, test, and integration. Additionally, the data fusion processing load is reduced and fusion can even be bypassed if required for timely reporting (e.g., in a threat warning situation).

        Such an architecture, although a conservative approach to sensor fusion, can nonetheless be rather effective if the sensor processing involves "soft" decisions. In such systems, sensor measurement uncertainties are reported or known to the data fusion process and sensor track files report ambiguous associations or identities; e.g., in terms of the likelihood of candidate hypotheses.

    C. *Distributed Tracking/Centralized Classification Architectures:* In these, individual sensors perform data stream association and track state estimation, but report parametric data for integrated multisource target ID. Such an architecture retains the advantage of distributed tracking: the track state estimation process is performed locally, at rates and with techniques and calibration data tailored to the individual sensor, with reduced data rates and processing load on the central data fusion processor.

        This approach also allows more nearly optimal target ID, by enabling the library association process to involve the full dimensionality of the sensor data, allowing exploitation of correlations of parameters between sensor data streams.

    D. *Centralized Track Fusion Architectures:* In these, track state estimation as well as ID of sensor track files is reserved for the central data fusion process. Although imposing a significant increase in centralized processing load, optimality can be enhanced by filtering tracks based on noise covariance properties of the set of contributing sensors.

    E. *Fully Centralized Fusion Architectures:* In these, observation report "snapshots" are passed directly to a centralized fusion process, for association, tracking, and ID across all sensor outputs. Full optimality
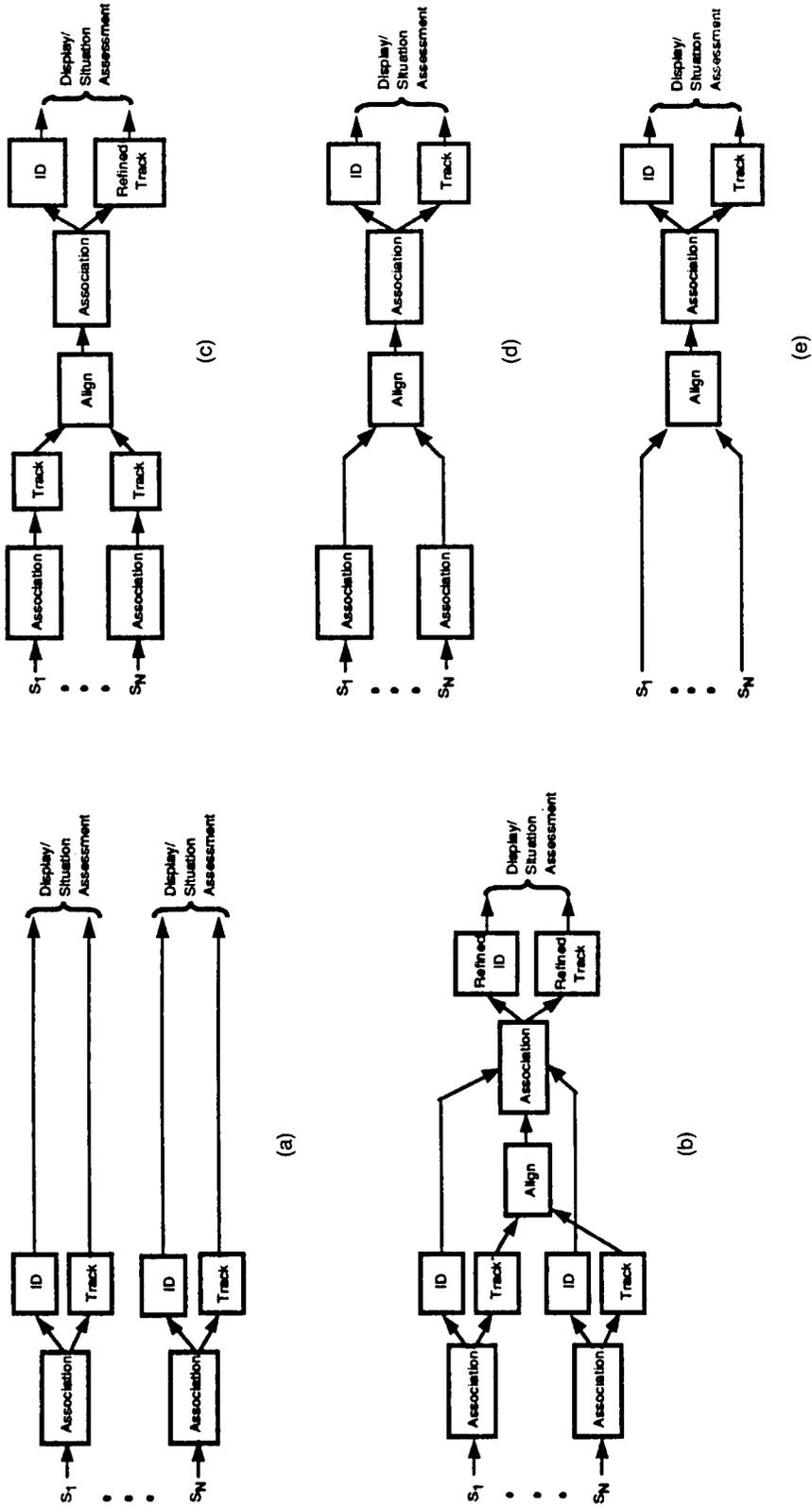
Fig. 3.9 Candidate data fusion architectures: (a) nonfused architecture coincident sensors; (b) fully distributed fusion architecture; (c) distributed tracking, centralized ID; (d) centralized track fusion; and (e) fully centralized, report level fusion.

is attainable in such a system, because cross-sensor attributive as well as kinematic correlations can be exploited both in report-to-report association and in report-to-library association (classification).

Architectures C, D, and E are each amenable to either pixel- or feature-level fusion processing, or to hybrid approaches, in which data from subsets of the sensors are processed differently. Additionally, hybrids among the architecture paradigms shown are possible, with hierarchies of fusion processing nodes.

3.3.2.2   Data Levels for Fusion.

As seen, there are choices concerning the level of single-source processing to be performed prior to multisource association at any given node.

In the specific context of sensor fusion, it has become customary to distinguish fusion performed at pixel, feature, and decision levels, representing the degree to which the sensor data has been processed or interpreted prior to fusion. (These *data* levels should not be confused with the DFS *fusion* levels defined in Sec. 3.2.4).

The alternative data levels for multisensor target ID correspond to alternatives for multisensor target tracking, as shown in Table 3.5. A given fusion system may involve data inputs of all three levels, depending on the types of sensors, the degree of their integration and the specific system requirements.

The following definitions are adapted from Robinson.[16]

Pixel-Level Data Fusion.   The lowest level where data fusion can occur is at the pixel or resolution cell. Hall[17] refers to this as *data-level* fusion, to include of single-dimension pre- or postdetection time-series data [i.e., unprocessed analog-to-digital (A/D) converter output] as well as focal plane data.

Because no sensor information is lost in abstracting data to a mutually commensurate form, pixel-level multisensor processing provides optimal detection and ID performance, often unattainable by other means. For example, pixel-level processing of multispectral LANDSAT data is able to discriminate such features as diseased crop, which are indistinguishable in single-spectral measurements. This level of integration combines the data from different sources prior to individual source detection processing. As such, it has the potential

Table 3.5   Possible Data Levels for ID and Track Fusion

SENSOR OUTPUT/ FUSION INPUT	TARGET ID	TARGET TRACKING
RAW DATA	Pixels, Waveforms	Pixels, Waveforms
MEASUREMENTS	Features	Position Measurements
DECISIONS	Class/ID	Tracks

for extracting multispectral features that would be masked by single-source detection thresholding or feature abstraction. However, pixel-level data fusion generally requires precisely registered sensors and high communication bandwidths.

Feature-Level Data Fusion.   At this level, individual sensors perform target detection and feature extraction processing. Target reports or filtered tracks from multiple sensors are then combined into multisource integrated (MSI) tracks before targets are classified.

In feature-level data fusion, feature vectors z_i provided by various sources are fused into a combined feature vector z_M (the subscript indicates MSI). The fusion process may be as simple as the concatenation of $\{z_i\}$ (for noncommensurate features) or it may involve various decision-theoretic, model-based, or other methods for combining measurements from multiple sources.

Feature-level data fusion is often a practical approach in the many cases in which it is not possible or desirable to combine commensurate data from multiple sources at the pixel level. For one thing, the characteristic features measurable in different types of sensors are often mutually noncommensurate. As an example, a long-wave IR sensor may detect a missile's exhaust plume at long ranges but not the missile body itself, which may be detected by a tracking radar. The features discernible in each sensor's data may be associated to improve the classification and tracking of the missile, even though there is very little added by association at the pixel level.

Even with commensurate features, it may be prohibitively difficult to provide the alignment, processing, and communications capabilities required for pixel-level data fusion. Commensurate postdetection features (e.g., target size as measured by multiple imaging sensors operating in different spectral bands) may be combined by a number of estimation techniques, depending on the availability of prior data, the time variability and statistical stability of the feature, and the degree of statistical dependency between sensor measurements. Techniques applicable to multisource feature estimation are discussed in Sec. 3.5.2 and 3.6.

Decision-Level Data Fusion.   The classification of targets can first be performed for each sensor (or for specific groups of sensors). The results of the completed decisions are combined with the verdict of one classification adding to or detracting from the result of another. A simple example of this would be a rule that at least two sensors must concur in the detection and classification of an object before its classification as a target can be accepted.

More sophisticated approaches involve *soft decision* nodes, whereby sensors qualify their decisions with confidence values (based on one or another model of uncertainty). More than one classification candidate may be reported, allowing resolution at the multisensor level. Each of these approaches has merit, and each should be used when appropriate. Figure 3.10 illustrates advantages and challenges of fusion at pixel, feature, and decision levels. The goal in each example is target classification, with the probability of correct classification being denoted by P_{cc}.

For feature-level fusion in a system with two sensors, the sets of detection features z_1 and z_2 are fused into a combined feature vector z. Per Eq. (3.3), the detection-theoretic optimal feature-level detection criterion for z is given by
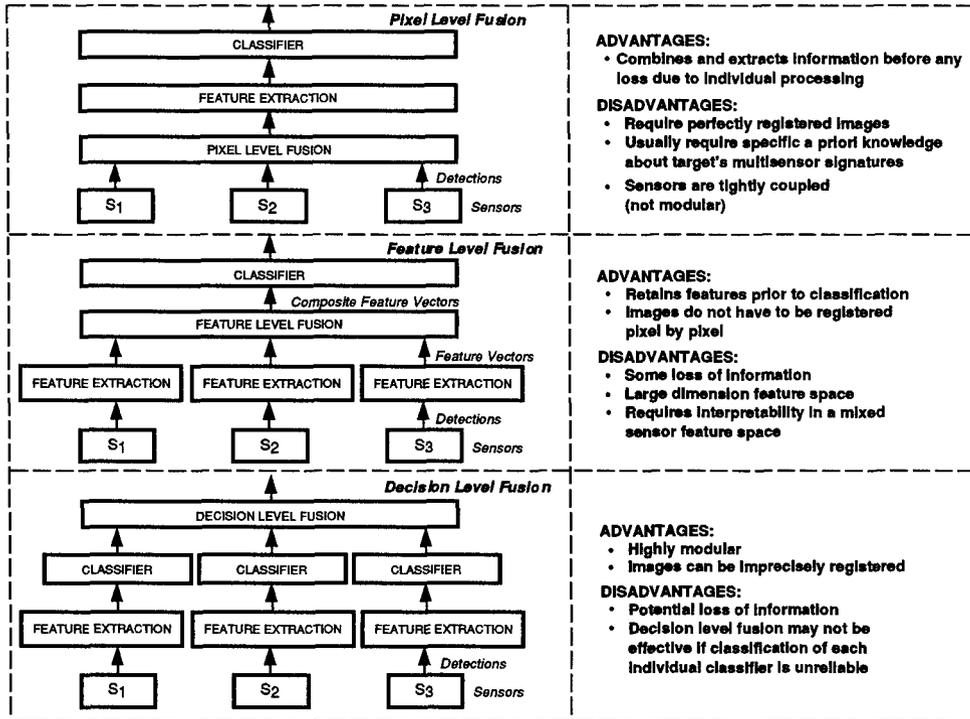
Fig. 3.10   Architectures for fusion at various data levels.[3]

$$\Lambda(\mathbf{z}) = \frac{\rho(\mathbf{z}|K_1)}{\rho(\mathbf{z}|K_2)} \begin{array}{c} > \\ < \end{array} \Lambda_0 \begin{array}{c} \Rightarrow X(\mathbf{z})\,\varepsilon\,K_1 \\ \Rightarrow X(\mathbf{z})\,\varepsilon\,K_2 \end{array} , \tag{3.6}$$

where $\rho(\mathbf{z}|K_1)$ and $\rho(\mathbf{z}|K_2)$ are the *a posteriori* probability densities of $\mathbf{z}$, given target class K_1 and K_2, respectively, and $X(\mathbf{z})$ is the target that generated the measured feature vector $\mathbf{z}$. The decision threshold Λ_0 is given by

$$\Lambda_0 = \frac{c_{11} - c_{22}}{c_{21} - c_{11}} \frac{\rho(K_2)}{\rho(K_1)} , \tag{3.7}$$

where $\rho(K_1)$ and $\rho(K_2)$ are the *a priori* probability densities of K_1 and K_2 and c_{ij} is the cost of misclassifying a signal of class K_j as one of class K_i. The problem of target detection can be viewed as a degenerate case of classification, in which K_1 may represent signal presence and K_2 the presence of noise, clutter, etc.[18,19]

To model the performance attainable in a decision-level fusion system, Robinson and Aboutalib[20] obtained the following sensor detection threshold:

$$\Lambda(\mathbf{z}) = \frac{\rho(\mathbf{z}_1|K_1)}{\rho(\mathbf{z}_1|K_2)} \begin{array}{c} > \\ < \end{array} \frac{\displaystyle\sum_{d_2=K_1}^{K_2} \gamma_1 \int_{\mathbf{Z}_2} \rho(D_2|\mathbf{z}_2)\rho(\mathbf{z}_2|\mathbf{z}_1,K_2)\ d\mathbf{z}_2}{\displaystyle\sum_{d_2=K_1}^{K_2} \gamma_2 \int_{\mathbf{Z}_2} \rho(D_2|\mathbf{z}_2)\rho(\mathbf{z}_2|\mathbf{z}_1,K_1)\ d\mathbf{z}_2} \frac{\rho(K_1)}{\rho(K_2)} \Rightarrow \begin{cases} X(\mathbf{z}_1)\varepsilon K_1 \\ X(\mathbf{z}_1)\varepsilon K_2 \end{cases},$$

$$(3.8)$$

where $\gamma_1 = \{F(D_1 = K_1, D_2, A = K_2) - F(D_1 = K_2, D_2, A = K_1)\}$, and $\gamma_2 = \{F(D_1 = K_1, D_2, A = K_2) - F(D_1 = K_2, D_2, A = K_2)\}$. This inequality provides a decision rule for sensor detection that minimizes the general cost function $F(D_1, D_2, A)$ given an actual target class A. Aboutalib[19] notes that global optimal performance can be achieved only if sensor S_1 is able to evaluate both probability densities $\rho(D_2|\mathbf{z}_2)$ and $\rho(\mathbf{z}_2|\mathbf{z}_1,K_2)$ for all detection feature sets $\mathbf{z}_2$ given $\mathbf{z}_1$. Thus, if we optimize the classification process of each sensor independently (as is usually the case where different sensor systems are developed independently), then the fused detection is generally suboptimal.

On the other hand, if feature vectors $\mathbf{z}_1$ and $\mathbf{z}_2$ are statistically independent (e.g., if the sensors are dissimilar), then $\rho(\mathbf{z}_2|\mathbf{z}_1,K_i) = \rho(\mathbf{z}_2|K_i)$ and the right side of Eq. (3.10) no longer depends on $\mathbf{z}_1$. If so, the sensors are decoupled and local optimization of each sensor can also result in global optimization of the fused decision. In other words, the classification performance of decision-level fusion is equivalent to that of feature-level fusion if the sensors are dissimilar.

An additional drawback of decision-level fusion is that its performance can be very sensitive to the choice of cost function. For example, if a relatively high penalty is imposed for a misclassification, the two sensors will be driven to disagree to minimize the cost function.

For these reasons, feature-level fusion has been recommended for similar sensors, reserving decision-level fusion for dissimilar and independent sensors.[19–21]

Researchers at Northrop have compared the performance of two feature-level techniques, a decision-level approach, and an innovative hybrid design.[22] The techniques were evaluated with live aircraft target data collected by two passive imaging sensors: imaging IR and TV. The figure of merit used for detection performance was $P_D - P_{FA}$ at various target ranges.

The target detection performance that was obtained with various fusion techniques is shown in Fig. 3.11. As seen, decision and feature level techniques were outperformed by a hybrid fusion technique. The latter involved a sequential process with feedback: the first stage, a decision-level fusion process, is triggered if any sensor reports a detection (i.e., using OR logic). Feedback is sent to any sensor that did not report a detection, commanding integration of more frames (to improve SNR), or reduction of the detection threshold. This process is iterated until both sensors report detections. Using this AND logic, the second fusion stage is triggered. The latter is a feature-level process, using the corresponding detection features from the two sensors and a hybrid feature classification tree with reduced threshold values. As seen in the figure, the hybrid fusion technique outperformed decision-level fusion by 66% and feature-level fusion by 11%. It also outperformed either sensor operating independently (OR decision-level fusion) by approximately 20%.
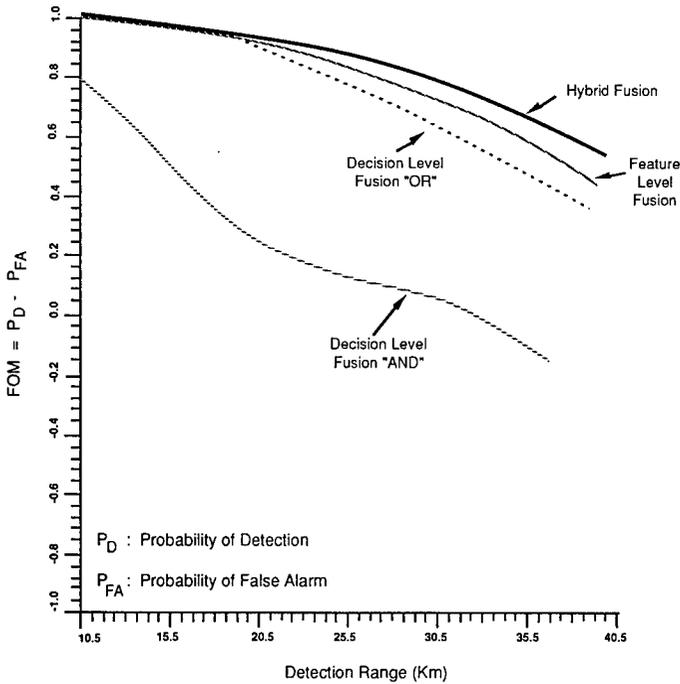
Fig. 3.11   Comparative performance with feature, decision, and hybrid fusion techniques.[22]

3.3.3   Managing the Combinatorial Explosion

Mathematically, the data association problem is a labeled set covering decision problem, in which subsets of the data (i.e., labeled as associated with individual objects or as false alarms) are to be found that group all data from the same object. Because of the differences in detection resolution between data sources, these sets may overlap (e.g., a single radar track may correspond to a cluster of resolved IRST tracks). Because labeled set covering problems are in the class of NP-complete problems, the number of labeling (report association) hypotheses can quickly exceed available processing capabilities as the problem size grows.

A further motivation for constraining the association search process is the level of optimality achievable with a soft decision structure. In the real world of noisy, imperfect sensors and nondiscriminating observable features, a non-truncating decision process soon accumulates an enormous number of working hypotheses regarding report association, track association, and ID. Accordingly, a feasible system must impose some sort of hypothesis truncation or hypothesis pruning strategy.

Several suboptimal approaches suggest themselves to relieve the processing load, with acceptable performance impact appropriate to the particular system environment.

3.3.3.1   Hard- Versus Soft-Decision Fusion.   Sensors with intrinsic data processing capabilities may report their data to the fusion process using a variety of means that influence the latter's design. One important consideration is the selection among alternative means of representing the degree of belief or uncertainty in processed data and the methods by which such data is combined with other sensors' data to make final decisions (e.g., detection, track, or classification decisions on targets.[7]

On the one hand, a sensor or fusion node may hard-limit its report output; such a *hard-decision* node reports only its single best estimate (under some concept of optimality) for each reported measurement or classification. Confidence or uncertainty values may be attached to reported values.

In contrast, a *soft-decision* node may report multiple competing hypotheses for resolution at the fusion or higher processing levels. Each may be tagged with associated belief or uncertainty values. Models of uncertainty and corresponding inference calculi that have been applied to problems of multisource classification include logical algebras, Bayesian single-valued probabilistic, evidential (*possibilistic*) and *fuzzy* inference methods. These techniques are discussed further in Sec. 3.6.1. Soft-decision models for kinematic track state estimation generally employ Bayesian-based probabilistic data association methods (treated in Sec. 3.5.2.2).

Quantitative comparison of the performance achievable with either hard or soft sensor reporting has been treated by several authors.[8,23-25]

As an example, Buede and Waltz[23] compare the performance of various decision rules in a representative automatic target recognition (ATR) application, in which fire-control decision must be made only on the basis of very high confidence target classification. Accordingly, the hard-decision case was represented by a strict reporting rule whereby a sensor can report only decisions that exceed a preestablished confidence threshold. The combined decision is made either on the basis of the first sensor reporting or by plurality vote.

In contrast, the soft-decision implementation involved sensors reporting their partial (or ambiguous) data as soon as signal detection is achieved, with uncertainties represented either as probabilities, evidential masses, or fuzzy variables, etc. The fusion process then combines these values using a decision paradigm appropriate to the given uncertainty model to derive a combined measure that may reduce the uncertainty over that of each sensor.

Simulation results indicated that, for a variety of cases of sensor ambiguity sets, an earlier, longer range decision is achieved by the soft-decision implementation than can be provided by any of the sensors operating independently or by the hard-decision fusion implementation.

3.3.3.2   Fusion Tree Structures.   There are several options in selecting the order in which multiple information streams are to be processed for association (i.e., for ordering the hypothesis-generation process). In general, these options involve deciding which tracks, which sensor outputs, and over which reporting times are to be searched in each successive fusion node.

The resulting ordering forms a *data fusion tree,* providing a top-down functional partitioning of the level 1 data fusion process. Each node in such a tree accepts a batch of reports from one or more sources (either sensors, static or dynamic libraries, or other fusion nodes) and determines which reports cor-

respond to the same object (target) or are false alarms. Fusion tree structures may be fixed in a particular system, with a predefined sequence of searching for candidate associations; or the structure may be adaptive to the current goals and/or data availability.

The selection of the tree structure appropriate to a given application is based on considerations of

1. the desire to avoid overlooking correct hypotheses having low *a priori* likelihood, particularly if such hypotheses have significant consequences
2. the desire to be able to retreat gracefully from selected hypotheses once they prove to be erroneous
3. constraining the exponential generation of association hypotheses
4. requirements for rapid convergence.

Factors 1 and 2 tend to drive the design to greater complexity; ultimately requiring searches over *all* sensor reports over *all* times against *all* tracks. This ideal must typically be constrained by the practical consequences of factors 3 and 4; resulting in a compromise, "knee-of-the-curve" level of performance.

Practical data fusion trees are generally designed to be constrained in time; i.e., processing only recent sensor reports (constrained either by age or by number of scans back). They may assess reports either on a sensor-by-sensor basis (illustrated in Fig. 3.12) or track by track (Fig. 3.13), allowing the search space to be adapted to track update priorities and to sensor reporting rates and predicted rates of significant change.

Most common tracking systems are data driven (Fig. 3.12); i.e., new sensor reports trigger a search for feasible track associations. Each such processing
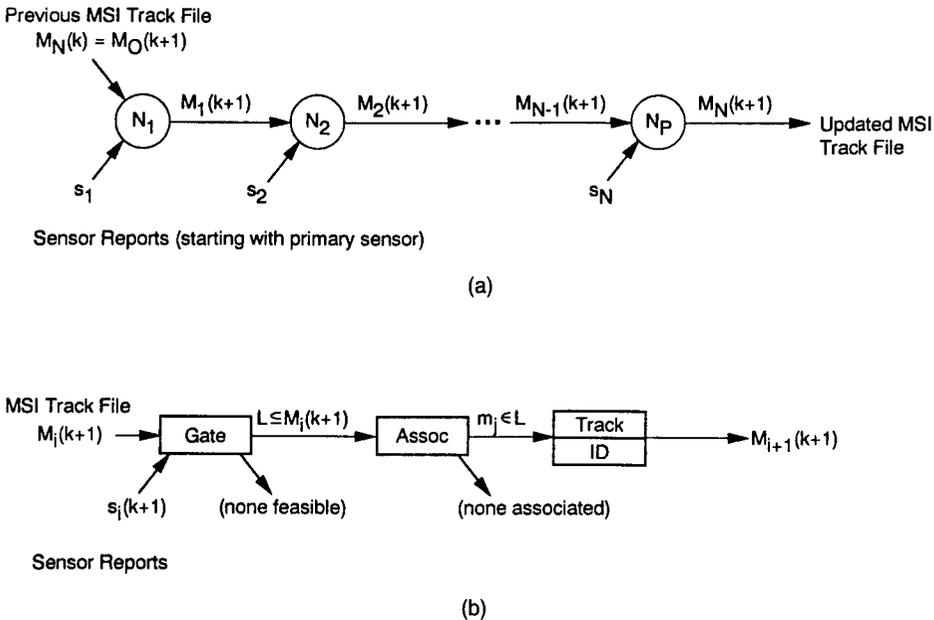
Fig. 3.12 Data-driven fusion structure: (a) tree and (b) node structure.
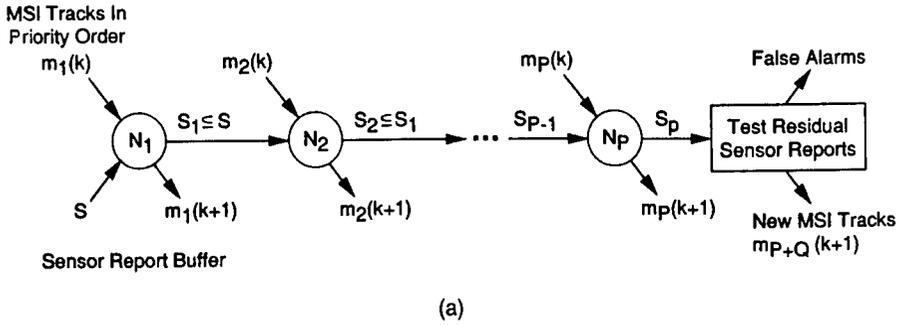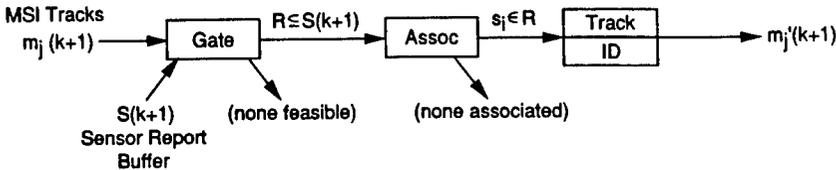
Fig. 3.13   Model-driven fusion structure: (a) tree and (b) node structure.

step may be initiated by one or another triggering rules; e.g., on receipt of a report from any input or from a specific input (e.g., a "clocking" sensor) or at periodic intervals, or when input buffers are full, etc.

Alternatively, a model-driven system (Fig. 3.13) searches for sensor data to associate with given MSI tracks. Thus, search can be constrained by the update priorities and predicted rate of change in individual tracks. Often such systems involve a primary tracking sensor that is used to cue other sensors for supporting data (e.g., target features or event detections).

The search for plausible associations may be constrained within the individual fusion node by the following methods:

1. *Pruning:* elimination of low-value hypotheses
2. *Combining hypotheses:* replacing multiple similar hypotheses with a single near equivalent
3. *Clustering:* partitioning the hypothesis space such that no cross associations are considered feasible
4. *Recovery:* eliminating noise-generated hypotheses.[13,26]

Typical pruning strategies involve either deleting less likely hypotheses (*unbiased pruning*) or deleting less consequential hypotheses (*biased pruning*). In the latter case, those hypotheses are deleted such that their truth or falsity is of little consequence, positive or negative, to the system's objectives. A variant of biased pruning involves pruning by abstraction: a set of hypotheses may be treated as a single hypothesis if there is relatively little differential consequence among them.

3.3.3.3   Architecture Example.   A system design by Bowman and Snashall[13] serves to illustrate some of the factors involved in fashioning a multisensor fusion architecture tailored to a specific, challenging application.

This application involves the automatic, real-time fusion of target data on-board a fighter aircraft; the data are reported by a representative diverse suite of on-board sensors and by other internetted aircraft. In this application, data errors tend to be uncorrelated between sensors and timeliness and processing resources are at a premium.

A periodic update process matched to the display or warning update requirement is employed. Because of limitations in data bus and data link bandwidth and in sensor reporting, multisensor fusion is performed at the feature or decision level. Because of processor and memory limitations, association involves no scan back (i.e., using only current, filtered track reports, or event reports of nontracking sensors).

The data fusion tree is organized in three segments for intrasensor fusion, multisensor fusion of the on-board sensor track files to form a local MSI file, and fusion of the internetted received data with the on-board MSI file. An initial set of nodes perform report-to-report and report-to-track associations within individual sensors (allowing features that are noncommensurate between sensors to be suppressed from further reporting). Each subsequent fusion node associates a particular sensor's track file with the MSI track file accumulated to that point, progressing from the most capable available sensor (i.e., that which generally provides the least track error).

The functional flow for each MSI node in this implementation is shown in Fig. 3.14. Each track in a given sensor's current track file is processed sequentially. The MSI track file is then output for the next fusion node. Individual sensor and MSI track files at each stage are available for higher level situation assessment and display as necessary for rapid response.

In the example, MSI fusion node design, each sensor track undergoes the following processes (Fig. 3.6):

1. *Alignment:* The MSI and sensor tracks are propagated forward to the current update time and sensor reports are transformed into the MSI coordinate system. This includes compensation for misalignments between sensors previously detected in the track association process.

2. *Association:* A data-driven fusion tree is used to associate MSI tracks with sensor tracks. Each fusion node involves the following processes:

   a. *Hypothesis generation:* Simple kinematic gates and feature matches are used to find all of the MSI tracks that could feasibly be correlated with the given sensor track. Past high-confidence associations are thereby confirmed, reducing the burden on the subsequent processing and simplifying tracking in dense environments.

   b. *Hypothesis evaluation:* For each feasible MSI track, a Bayesian maximum *a posteriori* probability (MAP) estimation is made as to the likelihood that the given correlation is correct, employing both kinematic and attributive factors. The highest likelihood value for MSI track association is compared against the likelihood that the sensor track is a new detection or a false alarm (computed on the basis of prior probabilities).
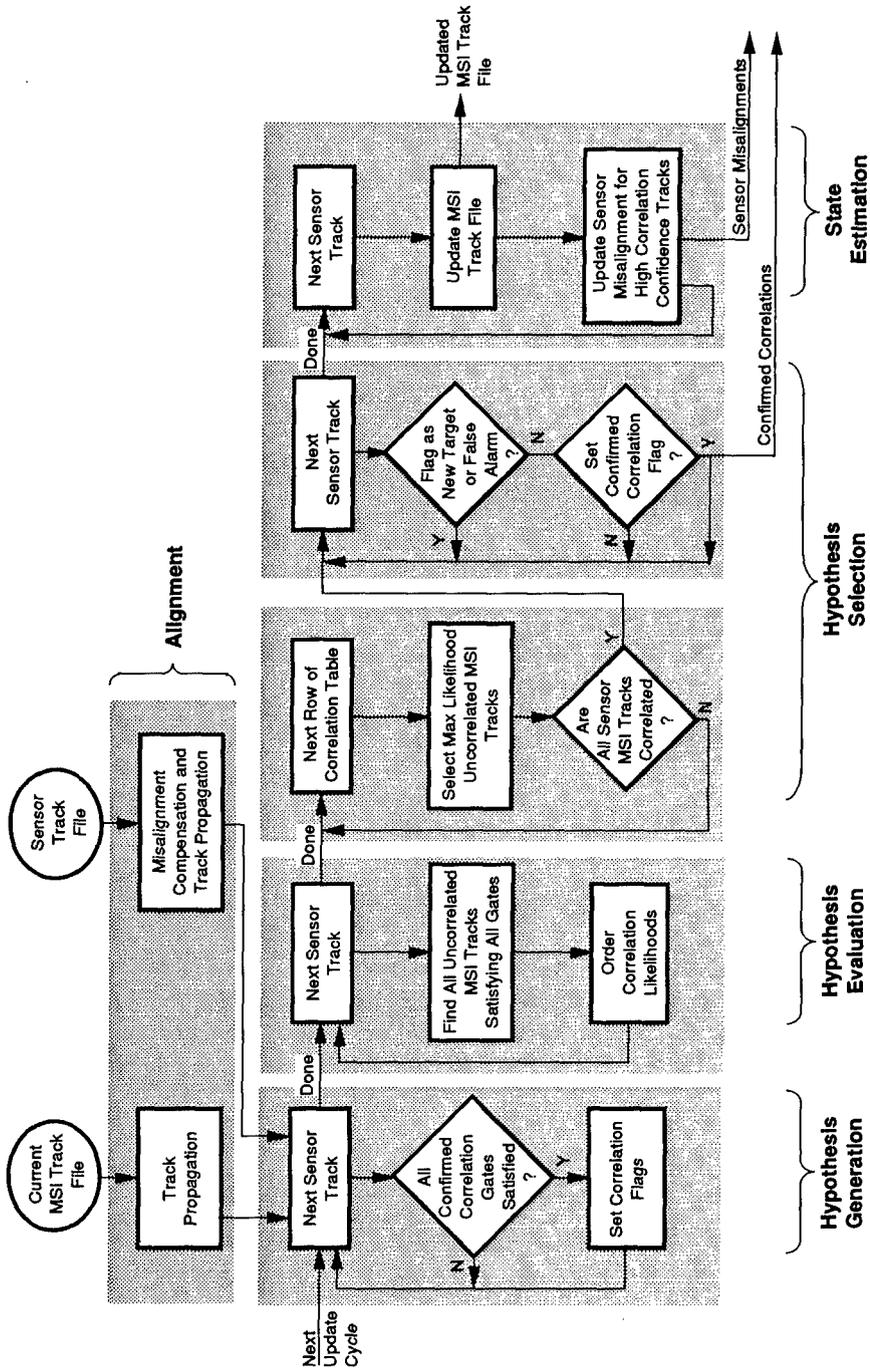
Fig. 3.14 Representative fusion node functional flow.[13]

  c. *Hypothesis selection:* The MAP association hypothesis is selected to update the current MSI file and to update the recursive association confidence. If the latter is sufficiently high, the association between the two tracks is confirmed, allowing feasible track gating and hypothesis to be bypassed in subsequent processing cycles.

3. *Target state estimation:* The kinematic and attribute information of the selected MSI track is updated, using the associated sensor data. Various multisource kinematic filtering and ID techniques can be employed.

3.3.3.4   Parallel Processing.

Parallel processing architectures provide hardware means for coping with the types of massive data processing problems found in multisource data fusion.

As Waltz and Llinas[7] point out, the operations performed in multisource multitarget tracking and correlation contain inherent parallelism that can be used on a parallel machine:

> Functional parallelism occurs in several areas of the data fusion process: (1) multiple sensors receiving reports and preprocessing data in parallel, (2) association of multiple reports in parallel, (3) computation and maintenance of multiple association and ID hypotheses, and (4) parallel searches for possible interpretations of data in knowledge-based reasoning for situation assessment.

In such a system, all stored hypotheses can be processed simultaneously and mathematical computations involving the corresponding probabilities can be performed in parallel. A single-instruction, multiple-data (SIMD) stream machine allows for parallel processing of multiple memory arrays. A bit-oriented SIMD architecture also yields an increase in efficiency because the field lengths of the operands can be specified.

Rosen and Schoenfeld[27] report on developments of parallel architectures for managing the combinatorial explosion in statistical correlation and tracking algorithms. These have been developed for naval command and control applications using contact (track) reporting data from many sources. Soft-decision algorithms maintain several alternative target track hypotheses and associated probabilities of truth, which can be stored and processed in parallel. As contact reports are received, the probabilities are updated to reflect the additional information, and hypotheses with sufficiently small probabilities are deleted from memory. Similarly, new track hypotheses may be generated with each received contact report, which indicate uncertainty in the received data.

In this way, the multihypothesis schemes can retain more information on the continuously changing reported situation; the number of hypotheses that are generated reflects the degree of uncertainty in the track association assessment. However, the number of hypotheses that can be retained in memory and processed in real time is a function of machine efficiency.

The application of parallel processing to multisource, multiple hypothesis correlation and tracking is also addressed by Bowman,[28] who argues for a neural network approach to preserve a great number of competing hypotheses within a realizable processing space.

3.3.4   Adaptivity in Data Fusion

As mentioned, one of the chief virtues of a fused system is its capability for process refinement, i.e., the ability to adapt to varying mission and system conditions. The revision to the DFS architecture (shown in Fig. 3.6) was motivated by desire to ensure the availability of data of all types to sensor/situation assessment functions at all levels of abstraction. Besides providing relatively "raw" low-level data for exploitation at the higher levels, such architectures recognize the value of adapting lower level processes as a function of higher level assessments.

One method is the use of feedback from higher level processes to modify the prior probability estimates employed in lower level detection, estimation, and classification processing nodes. Other examples include sensor misalignment correction based on target range estimates; passive range estimation based on target ID and apparent cross-section measurement; and ID and *a priori* velocity constraints together with measured azimuth rate.[29]

Additionally, feedback from higher level situation assessment and mission management functions allow for prioritized resource allocation, i.e., adaptivity to changes in mission priorities or in system status (e.g., loss or degradation of sensor, communications, or processing capability).

3.3.4.1   Integration Among Levels of Data Fusion.

Such opportunistic adaptivity is afforded by the exploitation of the fused information product to focus or refine the use of the system's sensors. Opportunities for refining data by fusion—whether in a feedback or a feed-forward architecture—are shown in Fig. 3.15:

1. *Detection and correction of misalignment:* Using track associations to isolate correlated tracking errors between sensors. These can be attributed to misalignments between sensors; which can be corrected via calibration. The alignment at infinity of noncollocated sensors can be used to refine range measurements using angle difference techniques, as discussed in Sec. 3.7.2.2.
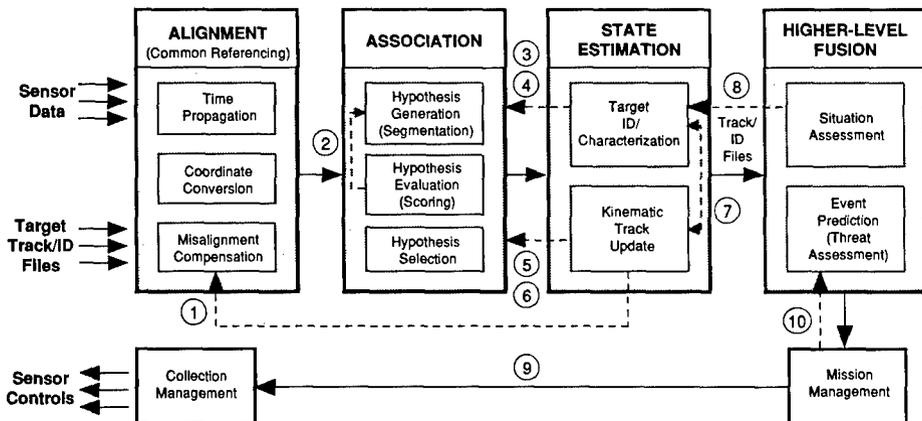
Fig. 3.15 Opportunities for adaptivity in data fusion.

2. *Confirmed association:* Using high-confidence associations among re-
   port clusters as *seeds* for the clustering of new reports. The processing
   burden of clustering N reports around $M \ll N$ confirmed object
   tracks is generally much less than that for clustering within the set
   of received sensor reports: $MN \ll N!$
3. *Predicted features of tracked targets:* Increasing the prior probabilities
   of observable features that are associated *a priori* with the estimated
   target types/states.
4. *Feature inconsistency:* Eliminating candidate track association hy-
   potheses on the basis of inconsistent features reported by the con-
   tributing sensors. This is equivalent to a distributed architecture in
   which the set of feasible ID/track hypotheses reported by one sensor
   does not intersect that reported by another.
5. *Predicted states of tracked targets:* Extrapolating the position and
   kinematic state of estimated targets for association with current sen-
   sor reports.
6. *Track inconsistency:* Eliminating candidate track association hy-
   potheses on the basis of inconsistent track histories.
7. *Track/ID incompatibility:* Eliminating an association hypothesis that
   yields a track that is outside the estimated performance envelope for
   the candidate target type/state.
8. *Situation inconsistency:* Use of contextual data in estimating prior
   probabilities. A higher level situation assessment function (which
   can be implemented with various degrees of automatic and human
   involvement) can evaluate the MSI ID/track reports on the basis of
   statistical or relational properties of sensed and *a priori* data. For
   example, situation assessment can assess the probability of a par-
   ticular type of target H being in the given environment, $P(H)$, or of
   data R from H being reported $P(R|H)$. Such contextual factors can
   include
   a. *Statistical:* Prevalence; e.g., a reported ambiguity between a land-
      based and shipborne emitter hypotheses can usually be resolved
      by recognizing the target's context; similarly, an ambiguity re-
      sulting from a parametric overlap between a specific U.S. and a
      Chinese emitter may be resolved by noting the geographical context.
   b. *Behavioral:* Spatial and temporal histories that discriminate among
      "abstract behavior patterns (e.g., normal activity, threatening con-
      ditions, suspicious activity, erroneous behavior)."
   c. *Relational:* Employing an operational model; e.g., by noting that
      a particular unidentified aircraft is flying in formation with three
      MiG-29's it can be inferred that it is very likely either a MiG-29
      or one of a small class of support aircraft (tanker, escort jammer,
      etc.), which can be tested for by monitoring the formation's tactics.
9. *Sensor cueing/tasking:* Control functions to sensors from the higher
   level assessment and mission management functions can include
   a. cueing a sensor to acquire a target being tracked by another sensor
      (either collocated or remote)
   b. tasking a sensor to search for a predicted target; e.g., to search
      for an object commonly associated with objects of the type of a
      detected object.

10. *Predicted simulation of environment by system actions:* Notification of system actions can be fed back from Mission Management for processing by the event prediction (threat assessment) function to determine likely responses by external systems. A threat system's reactions to system actions could include direct countermeasures—such as jamming, change in emission parameters, or target maneuvers—as well as other detectable actions. These expected reactions may be referred back to the appropriate level 1 to 3 functions in the form of modified prior attribute probabilities or track state expectations/variances. Similarly, the effects of *planned* or *candidate* future actions generated by mission management may be passed to event prediction to assess the likely consequences. In this manner, event prediction can operate in a simulation role supporting the planning function of mission management (see Sec. 3.8 for more discussion of planning in an integrated sensor/response system).

The lower level track associations, states, and IDs can thus be revised on the basis of higher level contextual evaluation.

3.3.4.2   Feedback Versus Feed-Forward Adaptivity.

There is a fundamental architectural issue regarding the desirability of feeding back the refinements in data association, ID, or kinematics estimation achievable by successive levels of information fusing. On the one hand, the decision thresholds and prior probability weightings used in the low level (e.g., single-sensor or MSI levels) can be made much more efficient by correction by higher level processes. However, it is important that errors in lower level estimates not be self-reinforcing or propagated to other sensors. Correcting a relatively accurate sensor with data from a highly biased one will degrade, not enhance, overall performance.

Contextual information can either be fed back from the situation assessment function in the form of modified prior probabilities or can be used to revise the fused track file.

Feedback from sensor data fusion to the individual sensors can enhance sensor detection performance. Detection enhancements can be achieved both by SNR improvements—employing such techniques as cued dwells and matched detection filters—and by lowering the sensor's detection threshold (achieving an acceptable false alarm rate because of the predicted high likelihood of signal presence). An example of the impressive performance improvements achievable by such techniques is given in Sec. 3.3.2.2.

On the other hand, such a mechanism can induce an uncontrolled successive amplification of error via positive feedback. Fusion of two uncorrelated sources, A and B, where the noise output of A is much greater than that of B, can result in a measurement that is noisier than that of B. Feeding back such a result for the purpose of "correcting" B will result in a general degradation of performance.

For this reason, it is generally preferable to refine and feed-forward information in a hierarchical information processing system. Such refinement takes the form of reduced kinematic uncertainty terms and refined confidences regarding candidate association or ID hypotheses.[25,30] Detection thresholds can still be lowered in a multisensor system involving independent sensors, because the resultant increased false alarm reports will tend not to correlate between sensors and will be purged on fusion.

3.4   DATA ALIGNMENT

Alignment is the process of transforming positional/kinematic data into a common spatiotemporal frame of reference. Data from two or more sensors can be effectively combined only if it can be established that they are observing the same scene or target. Alignment procedures are designed to permit association of multisensor data at the decision, feature, or pixel level.

Three straightforward processes are involved in establishing common referencing:

1. *Time propagation:* extrapolating old track location and kinematic data to the current update time
2. *Coordinate conversion:* translating data received in various platform-referenced coordinate system to a common spatial reference system
3. *Misalignment compensation:* correcting for known misalignments or parallax between sensors.

3.4.1   Time Propagation

Before attempting association of new sensor reports with the MSI track file, the latter must be updated to predict expected location and kinematic states of moving targets.

Target motion is modeled by

$$\mathbf{x}(k + 1) = \mathbf{F}(k)\mathbf{x}(k) + \mathbf{G}(k)\mathbf{u}(k) + \mathbf{v}(k) \; ; \tag{3.9}$$

where $\mathbf{x}(k)$ is the state vector at time t_k and $\mathbf{F}(k)$ is the state transition matrix at update time k:

$$\mathbf{F}(k) = \begin{bmatrix} 1 & \Delta t(k) \\ 0 & 1 \end{bmatrix} , \tag{3.10}$$

where Δt_k is the time since the previous update, $\mathbf{G}(k)\mathbf{u}(k)$ is a deterministic unknown error caused by possible target maneuver during the extrapolation interval and $\mathbf{v}(k)$ a process noise sequence (generally assumed to be zero-mean white Gaussian) with covariance $\mathbf{Q}(k)$.

The corresponding measurement model is

$$\mathbf{z}(k) = \mathbf{H}(k)\mathbf{x}(k) + \mathbf{w}(k) , \tag{3.11}$$

where $\mathbf{w}(k)$ is a zero-mean, white Gaussian noise sequence with covariance $\mathbf{R}(k)$.

Using this error model, the state estimate is propagated linearly:

Track state prediction:

$$\begin{aligned} \hat{\mathbf{x}}(k + 1|k) := \; & \mathbf{x}(k + 1) - \hat{\mathbf{x}}(k + 1|k) \\ = \; & \mathbf{F}(k) \, \hat{\mathbf{x}}(k|k) + \mathbf{G}(k)\mathbf{u}(k) , \end{aligned} \tag{3.12}$$

and the state covariance matrix is propagated to reflect the increased state vector uncertainty resulting from extrapolation time lag and possible target maneuver:

Prediction covariance:

$$\mathbf{P}(k + 1|k) := E[\bar{\mathbf{x}}(k + 1|k)\bar{\mathbf{x}}^T(k + 1|k)|Z_k]$$

$$= \mathbf{F}(k)\mathbf{P}(k|k)\mathbf{F}^T(k) + \mathbf{Q}(k) \; ; \tag{3.13}$$

The increased uncertainty due to time lag is given by the term $\mathbf{F}(k)\mathbf{P}(k|k)\mathbf{F}^T(k)$, where

$$\mathbf{P}(k) = \begin{bmatrix} \mathrm{cov}[\mathbf{x}(k)] & \mathrm{cov}[\mathbf{x}(k)\dot{\mathbf{x}}(k)] \\ \mathrm{cov}[\mathbf{x}(k)\dot{\mathbf{x}}(k)] & \mathrm{cov}[\dot{\mathbf{x}}(k)] \end{bmatrix} \; . \tag{3.14}$$

For reported sensor tracks, $\mathbf{P}(k)$ is the sensor track covariance matrix, which is to be made available as an input to the multisensor fusion node. For fused (MSI) tracks, $\mathbf{P}(k)$ is calculated as part of the track fusion process [per Eq. (3.52)].

The factor for the uncertainty resulting from unknown target maneuvers is given as

$$\mathbf{Q}(k) = \mathbf{u}(k)\mathbf{B}(k)\mathbf{B}(k)^T\mathbf{u}(k) \; , \tag{3.15}$$

where

$$\mathbf{B}_k = \begin{bmatrix} \Delta t_k^2/2 \\ \Delta t_k \end{bmatrix} \; , \tag{3.16}$$

and $\mathbf{u}(k)$ is the assumed maximum maneuver acceleration vector.[31]

3.4.2   Coordinate Conversion

The choice of a standard reference system for multisensor data referencing depends on (1) the standards imposed by the system into which reporting is to be made, (2) the degree of alignment attainable and required in the multiple sensors to be used, (3) their adaptability to various reference standards, and (4) the dynamic range of measurements to be obtained in the system (with the attendant concern for unacceptable quantization errors in the reported data).

In most cases of associating data from sensors on-board a single rigid platform, a body-oriented reference system is preferable (at least until the point of external reporting of such data).

For widely spaced sensors on or near the earth's surface, the World Geodetic System 1984 (WGS 84) is currently the most widely accepted and most precisely defined. This earth-centered, earth-fixed coordinate system is used worldwide; e.g., for most U.S. Defense applications and is the standard used by the Global Positioning System (GPS).

A standard coordinate system does not imply that each internetted platform will perform all its tracking or navigational calculations in this reference frame. The frame selected for internal processing is dependent on what is being

solved. For example, whenever an object's trajectory needs to be mapped on the earth, the WGS 84 is a natural frame for processing.

On the other hand, ballistic objects (e.g., spacecraft, ballistic missiles, and astronomical bodies) are most naturally tracked in an inertial system such as the FK5 system of epoch J2000.0. Each sensor platform will need a set of well-defined transformation matrices relating the local frame to the network standard one (e.g., for multiplatform data sensor data fusion).[32]

Passive sensors characteristically measure target angular position far more accurately than range. Hence, the ellipses corresponding to measurement error covariances are very elongated radial to the sensor, and location estimation is most naturally performed in a sensor-centered spherical coordinate system. The large range variances of passive measurements can cause large-scale biases on transformation into a Cartesian coordinate system, because of the nonlinearity of the coordinate transformation:

$$\mathbf{P_x} = \mathbf{J P_w J}^T \ , \tag{3.17}$$

where $\mathbf{x}$ and $\mathbf{w}$ are the target's Cartesian and spherical coordinates, respectively, and $\mathbf{J} = \partial\mathbf{x}/\partial\mathbf{z}$ is the Jacobian of the coordinate transformation. Accordingly, passive sensor data should, if possible, be associated and fused in a sensor-centered spherical representation.[33]

An analogous problem involves active sensors with poor angular accuracy (e.g., range-only radar) for which the same solution is appropriate.

3.4.3 Misalignment Compensation

In many applications, it is not necessary that a pair of sensors be boresighted, or even that they have overlapping fields of view (FOV). So long as track continuity can be assumed, it is often sufficient that the measurements taken with multiple sensors can be transformed into a common spatiotemporal coordinate system.

In some cases, however, it is necessary for a pair of sensors to have simultaneous observation of a target; for example when a wide-FOV sensor is used to cue a narrow-FOV sensor, as discussed in Sec. 3.8.2.

Standard spatial alignment techniques include[3]:

1. *Mechanical design:* requiring rigid structures, which may be incompatible with weight and dynamic load requirements (e.g., in a missile seeker) or with system designs (e.g., with scanning high resolution sensors)
2. *Target boards or boresight modules:* devices in the fields of view of sensors (these are impractical or very complex and expensive in many systems)
3. *Autoalignment:* using low-power signal sources (e.g., light beams) inserted into sensor optical path and focused onto detectors to sense alignment errors (this may be a promising technique for future shared aperture systems).

Data fusion processing adds a powerful alignment technique involving no special hardware and minimal special software. By associating reports on

targets of opportunity from multiple sensors, systematic alignment errors may be detected.

Such techniques have been applied to problems of mapping images to one another (or for rectifying one image to a given reference system). *Polynomial warping* techniques can be implemented without any assumptions concerning the image formation geometries. A linear least-squares mapping is performed based on known correspondences between a set of points in the two images.

Alignment based on targets of opportunity, of course, presupposes correct association and should only be performed with high-confidence associations. A high confidence in track association of point source tracks is supported by (1) a high degree of track cross-correlation (i.e., near-constant offset); (2) reported attributes (features) that are known *a priori* to be highly correlated and to have reasonable likelihood of being detected in the current mission context; and (3) lack of comparable high kinematic and feature correlation in conflicting associations among sensor tracks. Association metrics are treated in Sec. 3.5.2.

Two sensors can be boresighted such that either the centers of their fields of view coincide at infinite range *or* coincide at a finite range. This is illustrated in Fig. 3.16. Only in systems with concentric shared apertures can the centers of the fields of view coincide at all ranges.

In systems with separated apertures, the optimum boresight choice depends on the aperture sizes, angular resolutions, aperture separation, the degree of boresight error that can be tolerated, and the target ranges over which that boresight tolerance must be maintained.[3]

A typical system application is shown in Fig. 3.17, in which the navigation coordinates and target state vector must be transferred from the aircraft's fire control system to missiles carried at various stations on the aircraft.
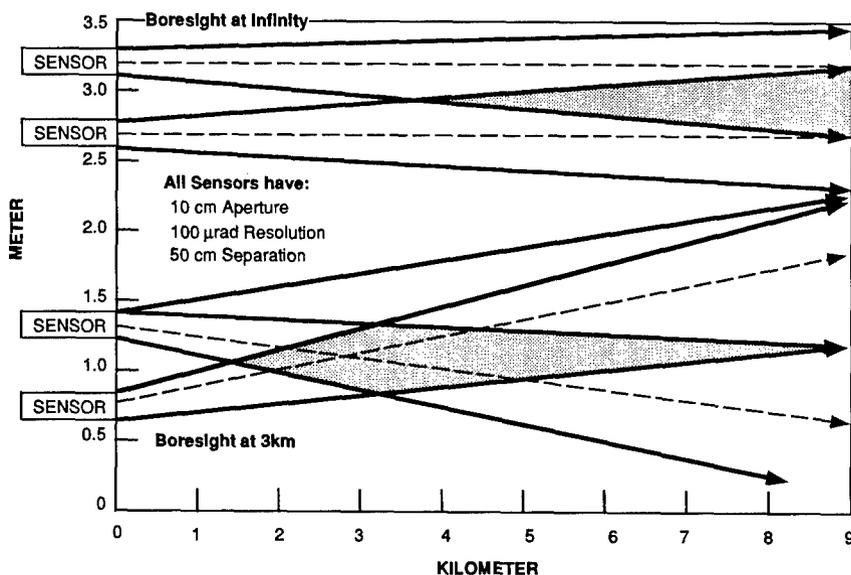
Fig. 3.16   Sensor boresight options.[3]
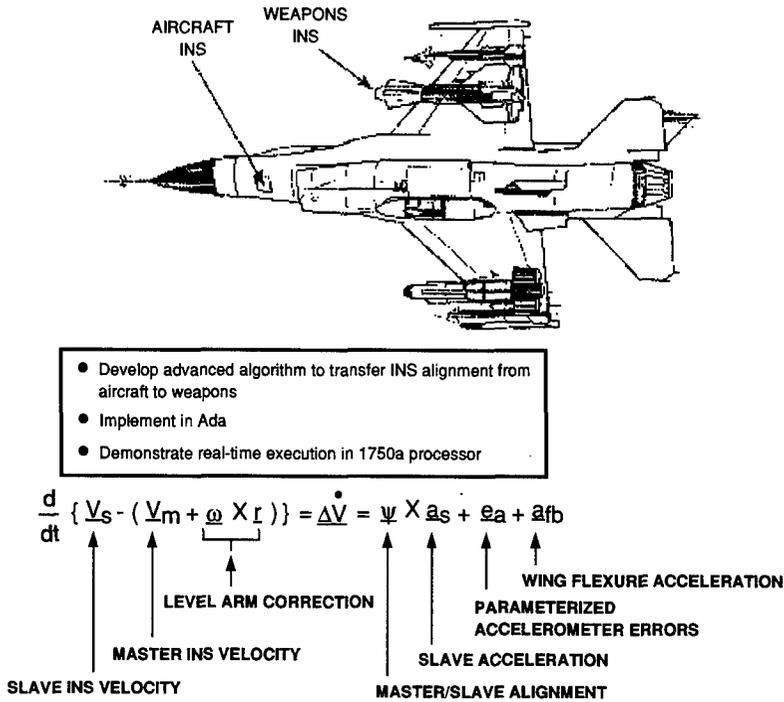
Fig. 3.17   F-16 aircraft/weapon transfer alignment problem.

Significantly more difficult are cases in which target tracks are to be handed from one moving platform [e.g., an airborne warning and control system (AWACS) aircraft] to another (e.g., an interceptor) despite large sensor separations and limited means available for alignment. These issues are taken up in Sec. 3.7.

3.5   DATA ASSOCIATION AND TRACKING

Figure 3.18 illustrates the general problem of data association and tracking. In the example, a set of six-sensor reports (which may be from one or more sensors) are collected in observation times t_1, t_2, and t_3. The problems to be addressed are, first, to determine which sets of reports represent the same sensed object and, second, to use the associated data to estimate the objects' (targets') kinematic states. As illustrated, the reported data may be consistent with three different hypotheses involving two moving targets having generated all reports. The data may also support numerous other hypotheses involving from one to six targets plus false alarms.

Techniques for resolving such association problems include decision metrics based on current track states or track state histories and attribute correlation.

The data association process takes sensor reports as input. These may be instantaneous observation reports or single-source tracks, depending on architectural decisions (see Sec. 3.3.2).
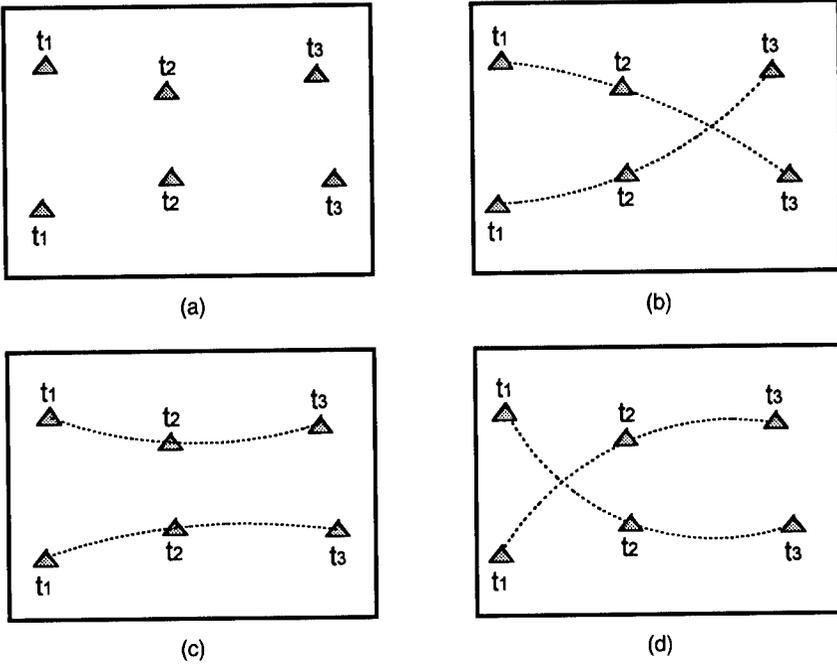
Fig. 3.18 Track association problem example: (a) target location reports, (b) association hypothesis 1, (c) association hypothesis 2, and (d) association hypothesis 3.

Outputs of data association are MSI track files with sets of associated sensor reports (and, depending on the implementation, likelihood scoring for each association) provided for kinematic/ID state estimation. Sensor reports that could not be associated with existing MSI tracks may be clustered and are also provided for track initiation/false alarm processing.

We can distinguish the following types of data association, which may be implemented under various system architectures:

1. *Report-to-report association:* determining that two reports, from the same sensor (look-to-look or scan-to-scan association) or multiple sensors, relate to the same target; used in track initiation
2. *Report-to-track association:* associating reports with current track files, either within a single sensor's track file (as in radar return-to-track association) or to an MSI track file
3. *Track-to-track association:* merging or splitting single-sensor or MSI tracks; merging eliminates duplicate tracks; splitting processes newly resolved tracks or tracks spawned from others [e.g., detection of a missile launched from a tracked aircraft, ballistic missile staging or reentry vehicle (RV) separation]
4. *Track (or report) to library association:* target ID or classification by comparison with a library of characteristics of individual objects or object types.

We follow the DFS partitioning[4] in treating the report-to-report, report-to-track and track-to-track associations as association processes; discussing library associations under identification in Sec. 3.6.

Key functions involved in multisource track association are

1. *Hypothesis generation (segmentation):* Gating techniques can be used to select MSI tracks with which a given sensor report can feasibly be associated. Alternatively, parametric clustering can be used to produce sets of sensor reports that plausibly are from the same target source. Sets of reports corresponding to prior high-confidence associations may be recognized, allowing much of the association processing to be bypassed, except for periodic reconfirmation.
2. *Hypothesis evaluation:* Candidate report-to-track or report-to-report associations are evaluated and confirmed on the basis of location/kinematic correlation and attributive consistency.
3. *Hypothesis selection:* Employing a decision rule to select association hypotheses to be retained for subsequent processing. Either hard- or soft-decision logic may be employed, the latter retaining multiple competing hypotheses, scored for likelihood.

Each fusion node produces one or more sets of sensor reports associated with MSI tracks. These associations may be edited and augmented in subsequent nodes of the fusion tree for input to the state estimation process.

Each of these processes may differ depending on (1) whether observed targets are stationary or moving, (2) the target and clutter densities expected to be reported by the contributing sensors, (3) the degree of attribute discrimination (and therefore association value) likely to be provided by the sensor measurements, and (4) system accuracy requirements and processing constraints.

3.5.1   Hypothesis Generation

A practical fusion system generally needs to constrain the data association search. An exhaustive search, testing all sensor reports from all times against all MSI tracks, could easily become impractical as the number of sensors or the reporting rate increases. As discussed in Sec. 3.3.3.2, practical systems incorporate data fusion trees that systematically test only some sensor reports from some times against some MSI tracks, thereby providing acceptable, if suboptimal, performance. A partitioning process occurs in each fusion node; e.g., using simple gating or clustering techniques to arrange the set of current sensor reports into subsets, each of which plausibly encompasses all the reports relating to one sensed object.

3.5.1.1   Validation Gating.   In processing point-source data from one or more sensors, spatial and kinematic gating techniques are often used to constrain the search space for report or track association.

Target motion and measurement error are assumed to have the form given in Eqs. (3.9) to (3.16). In cases where targets of interest are stationary or where their kinematics are sufficiently predictable and target/clutter densities permit, simple distance gating can be employed for hypothesis elimination as follows. For each pairing of current sensor reports z_S and the expectations z_M (MSI tracks propagated to the update time), the innovation v is obtained,

consisting of the residuals in each measurement dimension: $v = |\mathbf{z}_M(k) - \mathbf{z}_S(k)|$. A distance metric d^2 is defined as the norm of the innovation:

$$d^2 = \frac{1}{N}[v(k)]^T \mathbf{S}^{-1}[v(k)] \; ; \tag{3.18}$$

where N is the measurement dimension and $\mathbf{S}$ is the residual covariance matrix. Because correlation matrices are diagonal,

$$d^2 = \frac{1}{N}\sum_{n=1}^{N} \frac{[z_{M,n}(k) - z_{S,n}(k)]^2}{\sigma_{M,n}^2(k) + \sigma_{S,n}^2(k)} \; . \tag{3.19}$$

A sensor report is considered to be a valid candidate for association with a given MSI track if it falls within a *validation gate,* defined by a threshold δ on d^2:

$$\text{reject } H_{M,S} \text{ if } d^2 > \delta \; , \tag{3.20}$$

where $H_{M,S}$ is the hypothesis that $\mathbf{z}_M = \mathbf{z}_S$. Assuming a Gaussian distribution for the components of $\mathbf{z}$, d^2 has a chi-square distribution with N degrees of freedom. Therefore, the threshold can be set to any desired level of significance α:

$$P(d^2 > \delta|H_{M,S}) = \alpha \; . \tag{3.21}$$

If the MSI and sensor track errors are uncorrelated, the cross-covariance in any measurement dimension is simply the sum of the individual covariances:

$$\text{cov}[z_M(m) - z_S(m)] = \text{cov}[z_M(m)] + \text{cov}[z_S(m)]$$

$$= \sigma_M^2(m) + \sigma_S^2(m) \; , \tag{3.22}$$

where $z_M(m)$ and $z_S(m)$, $m = 1$ to N, are the measurement variables.

In many applications in which independence is not ensured, it is often nonetheless adequate to assume independence for the purposes of initial gating, foregoing optimality in favor of expediency.

These distances can be stored for later use in the hypothesis evaluation process, per Sec. 3.5.2. Gating based on measurement features can be performed as well; however, because of the greater amount of computation involved, feature gate tests are usually performed only in special circumstances: on track pairs that are to be evaluated for initial association and for confirming the association of track pairs that have already been associated with high confidence.

The selection of validation thresholds can be problematical if there is the possibility of large variability in the target kinematics or in the neighborhood target density. If validation gates are too narrow, target kinematics can cause tracks to be lost. If they are too wide, there is an increased likelihood of misassociation or the excessive processing burden on multiple hypotheses (de-

pending on the type of hypothesis selection process). Misassociations can also lead to subsequent track loss, as no consistent track is maintained. In some cases, higher level processing functions may be able to adjust validation thresholds based on contextual information.

Validation gating may be combined with other constraints; e.g., the maximum number of hypotheses to be processed, thereby constraining the processing burden. This is the *k-nearest neighbor* approach. When $k = 1$, this is equivalent to nearest neighbor association discussed in Sec. 3.5.2.1.

Alternative approaches include the use of maneuver detection logic, allowing a gate to be widened sufficiently to prevent loss of track, and the use of discriminating attribute data for hypothesis elimination. Bar Shalom and Fortmann[8] define and evaluate the following methods for maneuver detection: multiple model (white noise with two levels), input estimation, and variable dimension filtering.

3.5.1.2   Report Clustering.

There are applications in which it is impractical to employ a recursive tracking procedure against moving targets, because of insufficient processing resources or insufficient time available for track initiation and filtering.

An alternative approach is to cluster the set of current reports from the contributing sensors—e.g., by simple spatiotemporal proximity thresholding—to create prototracks for further processing. In effect, this data partitioning establishes constraints on plausible association hypotheses, eliminating implausible report pairings from consideration. Techniques for clustering are treated in Sec. 3.6.1.2.

Clustering is especially useful when one sensor can be counted on to provide more accurate tracking than any other. For example, in an integrated aircraft threat warning system, the feature-rich emitter reports from a radar warning receiver may be sorted by clustering around the very precise IRST tracks.

Brandstadt and Maston[34] report on a clustering scheme followed by multiple hypothesis track association designed to solve the problems of short observation times and noisy data. A multisensor targeting problem is addressed, involving an attack helicopter employing FLIR, TV, and millimeter-wave (MMW) radar for target detection, selection, and fire control support. The helicopter "pops up" from behind cover to take quick looks at the battlefield to develop targeting solutions. There is typically insufficient tracking time available in each look to initiate tracks sufficient for look-to-look association.

The approach taken in Ref. 34 is to build clusters of current sensor reports that are feasibly from the same platform (in terms of consistency in report measurements given the individual sensor covariances). The specific technique used differs from more common forms of clustering (i.e., grouping detections that fall within high density regions or within a fixed distance of a given centroid) in being statistically nonstationary: reports are pairwise grouped without the requirement for transitivity. This allows plausible sequences of observations to be associated by means of attributive or coarse kinematic thresholds, without requiring track initiation and recursion.

Feasible report clusters are tested against the mission data base to determine correlations in attributes and feasible kinematic state vectors consistent with the candidate associations of clusters and reports to targets.

3.5.1.3   Scene Segmentation.   In image processing systems, partitioning takes the form of *scene segmentation:* designating regions that may correspond to objects of interest for ID and tracking. A dynamic data base is generated of pixel-registered sensor images, possibly with weightings assigned to designate pixels of high feature recognition value for one or more target types.

Delanoy et al.[35] report on such a scheme employing a rule-based expert system for generation of interest images and for scene interpretation. Database generation includes decisions concerning selection of images and image resolution, based on knowledge concerning the sensor suite, target set, and environmental conditions. Multiple interest images are combined pixelwise, resulting in a single map of interest value. Regions of high interest are thereby identified. Corresponding regions in sensor returns may then be selectively processed at higher resolution for feature detection. The mapping from features to regions helps determine the extraction and identification processes to be employed.

Related functions involve edge and shape detection as a precursor to single- or multisource feature extraction. Matching features using edges, peaks, corners, and saddle points can provide significant computational reduction as compared to correlation matching techniques.[36] Other model-based detection schemes (matched filtering etc.) similarly permit enhanced detection performance under adverse signal-to-noise and clutter conditions.

3.5.2   Hypothesis Evaluation

Several techniques for evaluating associations between measurement reports have been discussed in the literature. Each of these techniques involves testing pairs of data for feasibility and support of association hypotheses (i.e., that they refer to the same object in the sensor measurement space). As such, they may be applied to report-to-report, report-to-track, track-to-track, and library association problems.

Following Ref. 8, we present four types of association techniques with wide applicability. The simplest are nearest neighbor techniques, which use a distance metric, as described in validation gating, selecting the validated measurement that is closest to the predicted measurement as if it were the correct one. A drawback of such a hard-decision approach is that it can eliminate valid measurement candidates.

The remaining techniques discussed avoid this problem in different ways. *Track-splitting* techniques generate multiple association hypotheses whenever there is more than one validated measurement, relying on longer term collection to eliminate unlikely track sequences. The principal concern with such techniques remains that of combinatorial explosion in dense environments.

Bayesian maximum *a posteriori* probability (MAP) techniques avoid hard-decision association. Rather, tracks are updated using the probabilistic combination of all validated measurements. Of Bayesian techniques, so-called *probabilistic data association* (PDA) techniques combine measurements only over the current update cycle, *N-scan* techniques do so over a constrained number of cycles, and *optimal Bayesian* techniques over the entire collection history.

These approaches generally differ in their computational burden and, often, in their performance. As discussed in the following, MAP techniques are generally superior in associating data from sensors with common measurement dimensionality, whereas the chi-square technique is preferable in combining measurements with different dimensionalities.

3.5.2.1 Nearest Neighbor Association.

The simplest approach to data association is that of selecting the validated measurement that is closest to the predicted measurement under some distance metric. This nearest neighbor approach is most applicable with target sets having low densities or predictable kinematics. A fixed threshold (the validation gate) enables initiation of new tracks.

Associate sensor report $\mathbf{z}_s$ with MSI track $\mathbf{z}_m$ if

$$m = \arg\{\min_j[(\mathbf{z}_s - \mathbf{z}_j)^T \mathbf{W}_s^{-1}(\mathbf{z}_s - \mathbf{z}_j)]\} , \text{ and}$$

$$[(\mathbf{z}_s - \mathbf{z}_j)^T \mathbf{W}_s^{-1}(\mathbf{z}_s - \mathbf{z}_j)] < \delta$$

(3.23)

(time indices having been suppressed for clarity).

This distance metric may be simple Euclidean distance (for which $\mathbf{W}_s^{-1}$ is the identity matrix). Or a linearly weighted distance function (*a posteriori* probability) may be used, based on statistically known variances: $\mathbf{W}_s$ being a diagonal weighing matrix for the sensor sources s. Such techniques are most suitable for single-target tracking in sparse environments.

Improved performance in denser environments can be obtained via a quadratic (Mahalanobis) weighted distance based on statistically known covariances among measurement sources: the cross-covariance matrix $\mathbf{M}_{ij}$ for the sensor sources being used for $\mathbf{W}_s$, per Sec. 3.5.4.1. Under the white Gaussian assumption, this is equivalent to a chi-square tail test for N degrees of freedom.

The chi-square test has an advantage over alternative techniques—and particularly the Bayesian (MAP) techniques discussed below—in that it allows natural evaluation of track associations in cases where the dimensionality of track files differs between sensors (e.g., in combining data between an IRST reporting only angle data and a radar reporting angle, range, and range rate). In contrast, a MAP approach would require selecting an arbitrary *a priori* conditional probability for $P(\mathbf{z}|H)$ for the parameters missing in the lower dimensional reports, thereby adding an unwanted bias to the association scoring.

By normalizing for degrees of freedom, the chi-square test is a uniformly most powerful unbiased test for estimating the association of reports with Gaussian error and that may differ in dimensionality and in number of reports per track.[13]

However, because multispectral attributes (or features) are often not commensurate and ID confidences are typically not Gaussian distributed, the chi-square process is generally limited to association via kinematics (with reported features and ID decisions/confidences then employed in some probabilistic or possibilistic manner).

In systems where individual sensors report tracked state estimates, their inputs to the fusion process is autocorrelated both from scan-to-scan and across sensors with each scan.[13] The problem can be treated directly, by compensating

for the autocorrelations.[37] Bowman and Munzer[25] propose a simpler approach, whereby the fused track file is reinitialized with each scan (i.e., to avoid filtering filtered data), ignoring intersensor track correlations both for association and update. Simplified track maintenance logic is employed to maintain consistent track ID across time scans.

3.5.2.2 Track-Splitting Techniques.

A drawback of nearest neighbor techniques (indeed, of all hard-decision techniques) is that any evidence against the selected association is ignored. By definition, any measurement in a validation gate is a valid candidate for association with the given track and the nearest one may not be correct.

One method to protect from such erroneous associations is to split the track into multiple hypotheses every time there is more than one measurement in a validation gate.

An obvious problem with such methods is the generation of unmanageable quantities of track hypotheses. A pruning technique may be employed to eliminate less likely track hypotheses. If $Z_{k,l}$ is the l'th sequence of measurements up to time t_k; the likelihood function of this sequence being a genuine track (i.e., all relating to the same target) is the joint probability density function

$$\rho(Z_{k,l}|*Z_{k,l}) = \rho(\mathbf{z}_{i_1,l}, \ldots, \mathbf{z}_{i_k,l}|*Z_{k,l}) , \tag{3.24}$$

where $*Z_{k,l} := \{Z_{k,l}$ is a correct track$\}$ and $\mathbf{z}_i(j)$ is the i'th measurement at time t_j.

Low-value track hypotheses $Z_{k,l}$ are pruned by means of a threshold on the log likelihood function:

$$\lambda_l(k) := \sum_{j=1}^{k} v^T(j)\mathbf{S}^{-1}(j)v(j)$$

$$= -2 \log\left[\frac{\rho(Z_{k,l}|*Z_{k,l})}{\prod_{j=1}^{k} |2\pi\mathbf{S}(j)|^{-1/2}}\right] . \tag{3.25}$$

Under the linear Gaussian assumption, this likelihood function is chi-square distributed, so that the threshold may be set to any desired degree of significance.

3.5.2.3 Probabilistic Data Association.

Another way to exploit multiple measurements within a validation gate is to avoid selecting any one such measurement for association to the exclusion of the others. Rather, track state updating is performed using a Bayesian weighted sum of all validated measurements, with weights corresponding to the likelihood of being a correct association. In probabilistic data association [or joint probabilistic data association (JPDA) when several targets are being tracked] this likelihood calculation is performed only over the current measurement cycle.

Let β_i be the probability that a measurement $\mathbf{z}_i$ correctly associates with the track for a particular target, given the current measurement set $Z(k)$:

$$\beta_i(k) := P(*z_i(k)|Z_k) \; , \tag{3.26}$$

(β_0 is the probability that none of the measurements was caused by the target).

Using the total probability theorem, the conditional mean target state estimate is

$$\hat{\mathbf{x}}(k|k) = E[\mathbf{x}(k)|Z_k] = \sum_{i=0}^{m_k} E[\mathbf{x}(k)|*z_i(k),Z_k]P\{*z_i(k)|Z_k\}$$
$$= \sum_{i=0}^{m_k} \hat{\mathbf{x}}_i(k|k)\beta_i(k) \; , \tag{3.27}$$

where $\hat{\mathbf{x}}_i(k|k)$ is the updated state estimate conditioned on the i'th validated measurement being correct:

$$\hat{\mathbf{x}}_i(k|k) = \hat{\mathbf{x}}_i(k|k-1) + \mathbf{W}(k)v_j(k) \; , \tag{3.28}$$

where

$$v(k) := \sum_{i=1}^{m_k} \beta_i(k)v_i(k) \; , \tag{3.29}$$

is known as the *combined innovation*.[8]

This update, Eq. (3.28), is nonlinear because of the dependencies of the probabilities β_i on the innovations. Also, unlike standard Kalman filters, the covariance of the estimate depends on these probabilities:

$$\mathbf{P}(k|k) = \beta_0(k)\mathbf{P}(k|k-1) + [1 - \beta_0(k)]\mathbf{P}_c(k|k) + \check{\mathbf{P}}(k) \; , \tag{3.30}$$

and

$$\mathbf{P}(k+1|k) = \mathbf{F}(k)\mathbf{P}(k|k)\mathbf{F}^T(k) + \mathbf{Q} \; . \tag{3.31}$$

Here

$$\check{\mathbf{P}}(k) = \mathbf{W}(k)\left[\sum_{i=1}^{m_k} \beta_i(k)v_i(k)v^T(k) - v(k)v^T(k)\right]\mathbf{W}^T(k) \; , \tag{3.32}$$

and

$$\mathbf{P}_c(k|k) := [\mathbf{I} - \mathbf{W}(k)\mathbf{H}(k)]\mathbf{P}(k|k-1) \tag{3.33}$$

is the covariance of the state updated with a correct measurement.[8]

Probabilistic approaches to data association can be applied to attribute fusion (for target ID) as well as to track state fusion. The classical recursive Bayesian approach to attribute fusion is to update probabilities of classification hypotheses, assuming correct association:

$$P(H_{jm})_k = \frac{\rho[\mathbf{z}_j(k)|H_{jm}] \cdot P(H_{jm})_{k-1}}{\displaystyle\sum_{\forall n} \rho[\mathbf{z}_j(k)|H_{jn}] \cdot P(H_{jn})_{k-1}} \; , \tag{3.34}$$

where $P(H_{jm})_k$ is the estimated probability at time t_k that the j'th track was caused by a target of type m.

In actuality, such a classification probability is contingent on the correctness of associating the measurement $\mathbf{z}_j(k)$ with a target track x_i on which the prior probabilities have been established. Writing this association hypothesis X_{ij}, the conditional classification probability is

$$P(H_{jm}|X_{ij})_k = \frac{\rho[\mathbf{z}_j(k)|H_{jm}, X_{ij}] \cdot P(H_{jm})_{k-1}}{\displaystyle\sum_{\forall n} \rho[\mathbf{z}_j(k)|H_{jn}, X_{ij}] \cdot P(H_{jn})_{k-1}} \; . \tag{3.35}$$

Brandstadt[24] applies the total probability theorem to rewrite Eq. (3.34), scaling the given probability density functions by $P[X_{ij}|\mathbf{z}_j(k)]$, which is the conditional probability that the measurement is correctly associated with a target X_i:

$$P(H_{jm})_k = [P(H_{jm}|X_{ij})_k - P(H_{jm})_{k-1}] \cdot P[X_{ij}|\mathbf{z}_j(k)] + P(H_{jm})_{k-1} \; . \tag{3.36}$$

The probabilistic association $P[X_{ij}|\mathbf{z}_j(k)]$ can be accomplished in any of several ways. In a PDA approach (as employed in Ref. 24), taking all the possible targets and combinations of targets that could have generated $\mathbf{z}_j(k)$, all the probabilities that include X_{ij} are summed to obtain $P[X_{ij}|\mathbf{z}_j(k)]$.

This can be a tedious procedure if there is any likelihood that more than two targets could have generated $\mathbf{z}_j(k)$. There are, fortunately, many applications in which that likelihood is fairly remote; e.g., in tracking submarines or spacecraft, where interference is usually attributable to clutter, not other objects.

In such cases, each measurement tends to be unique and $P[X_{ij}|\mathbf{z}_j(k)] = P[X_{ij}|\mathbf{z}_j(k), 1]$. Under this assumption, this probability may be given as

$$P[X_{ij}|\mathbf{z},1] = \frac{\rho(\mathbf{z}|X_{ij},1) \cdot P(X_{ij})}{\displaystyle\sum_{n=1}^{N} [\rho(\mathbf{z}|X_n,1)] + \rho(\mathbf{z}|X_{nt},1) \cdot P(X_{nt})} \; , \tag{3.37}$$

where $\mathbf{z} = \mathbf{z}_j(k)$, $P(X_n)$ is the prior probability that a particular target X_n will generate a measurement, $P(X_{nt})$ is the probability of a new target detection, and N is the number of tracks selected for evaluation by the segmentation/hypothesis generation function.[24]

In other cases, in which target densities are sufficient that there is a significant probability of misassociation between tracks, other probabilistic data association techniques must be employed. Siu and Pue[38] demonstrate improved tracking performance using JPDA over PDA by fusing multiple source data prior to filtering.

3.5.2.4 N-Scan and Optimal Bayesian Association.

Alternatively, association probabilities can be computed over multiple measurement cycles (N-scan back) or, ideally, over the entire collection history (optimal Bayesian association). The amount of collection history to be processed, of course, can be selected as a trade-off between computation burden and performance.

A maximum likelihood estimator can be employed that selects the hypothesis that maximizes the probability of the value of the target state vector, given the sequence of observed values. Thus, the probability function is the joint probability distribution between the possible values of the random estimated variable and the actual value observed. If a multivariate Gaussian distribution can be assumed for the estimated random variable, the solution is a weighted least-squares estimator, as discussed.

Such techniques in effect combine the virtues of track-splitting and PDA approaches: probabilities of correct association are evaluated over a set of possible track measurement histories l:

$$\beta_{k,l} = P(*\mathbf{z}_{k,l}|Z_k) \ .$$

$$(3.38)$$

The total number of these histories at time t_k is

$$L_k = \prod_{j=1}^{k} (1 + m_j) \ ,$$

$$(3.39)$$

for m_j measurements (plus the possibility of no measurement) at each update time t_j.

For each such history, the track state estimate is given as

$$\hat{\mathbf{x}}_l(k|k) = \hat{\mathbf{x}}_s(k|k - 1) + \mathbf{W}_l(k)[\mathbf{z}_{ik,l}(k) - \hat{\mathbf{z}}_s(k|k - 1)] \ ,$$

$$(3.40)$$

where $\hat{\mathbf{z}}(k|k - 1)$ is the predicted measurement corresponding to history $Z_{k-1,s}$, with covariance $\mathbf{S}_s$. Thus, the combined state estimate at time k is

$$\hat{\mathbf{x}}(k|k) = \sum_{l=1}^{L_k} \hat{\mathbf{x}}_l(k|k)\beta_{k,l} \ .$$

$$(3.41)$$

The covariance on the update state conditioned by each history l is

$$\mathbf{P}_l(k|k) = [\mathbf{I} - \mathbf{W}_l(k)\mathbf{H}]\mathbf{P}_s(k|k - 1) \ ,$$

$$(3.42)$$

and the covariance associated with the combined estimate is

$$\mathbf{P}(k|k) = \sum_{l=1}^{L_k} \beta_{k,l}\mathbf{P}_l(k|k) + \sum_{l=1}^{L_k} \beta_{k,l}\hat{\mathbf{x}}_l(k|k)\hat{\mathbf{x}}_l^T(k|k) - \hat{\mathbf{x}}(k|k)\hat{\mathbf{x}}^T(k|k) \ .$$

$$(3.43)$$

3.5.2.5 Nonestimation Approaches.

The preceding estimation-theoretic techniques by no means exhaust the possible approaches to evaluating and selecting candidate data associations.

Maren et al.[39] report on a data fusion system that employs a neural network for adaptive real-time control of the coefficients of an alpha/beta tracker, with target kinematic information provided by the predicted target position and sensor data. A cooperative-competitive neural network is applied to resolve conflicts between new target detections and existing target tracks. Network operation is unsupervised; the strengths of various nodes adjust to the innovation for candidate association hypotheses. Inhibitions and/or excitations between neurons determine a measure of closeness for each candidate report/track pair, with the strongest value representing the global best pairing set.

3.5.3 Hypothesis Selection

Hypothesis evaluation generates triplets $\langle m,r,s \rangle$, consisting of an MSI track file m, a sensor report r (of whatever type the sensor generates), and an evaluation score (e.g., a likelihood value, evidential mass, or match factor, depending on implementation). These are processed by a hypothesis selection function to produce pairs $\langle m,R \rangle$ involving an MSI track file and a set R of sensor reports to be used jointly in the state estimation for m. Specific declarations by hypothesis selection can include

1. new MSI track (i.e., insufficient correlation between sensor report with any existing track and acceptably low probability of false alarm)
2. updated old track
3. false alarm (erroneous sensor report)
4. inactive track (because of an observed inactive state, observed target destruction, or timed out track)
5. merged tracks
6. split tracks (in which a previously associated set of sensor tracks are no longer found to be consistent with a single target).

The hypothesis selection process can be integrated in a variety of ways with target ID/track estimation functions. Decision criteria have been developed that incorporate association and track state estimation either as (1) a joint process, (2) a sequential process with hard or deterministic association decisions preceding track state estimation, or (3) a sequential process with soft or probabilistic association decisions. Bowman and Munzer[25] have defined maximum *a posteriori* paradigms for these three types of association criteria as follows:

1. *Joint multisensor correlation decision and target state estimation:* The association hypothesis and state vector estimates are jointly optimized, given sensor kinematic/ID reports R:

$$\max_{H,\hat{z}} P(H,\hat{z}|R) = \max_{\hat{z}} [P(\hat{z}|R,H)]P(H|R) \ . \tag{3.44}$$

2. *Multisensor correlation decision:* The association hypothesis is selected optimally and target state vector estimates derived from this hypothesis. This approach is sometimes referred to as *deterministic* data fusion because the association hypothesis is selected unconditionally (e.g., using the nearest neighbor method) before performing the update:

$$\max_{H} P(H|R) \; = \; \max_{H}[P(R|H)P(H)] \; , \tag{3.45}$$

then

$$\max_{\hat{z}} P(\hat{z}|R,\hat{H}) \; .$$

3. *Target state estimation:* The target state estimates are optimally determined, while allowing the association to be unresolved (employing one of the probabilistic data association techniques to average the state update over the different hypotheses):

$$\max_{\hat{z}} P(\hat{z}|R) \; = \; \max_{\hat{z}}\left\{\sum_{H}[P(R|H,\hat{z})P(H|\hat{z})]\right\}P(\hat{z}) \; . \tag{3.46}$$

Criterion 1 has the advantage over criterion 2 of improved state estimation, but at increased computational cost. Criterion 1 is the least computationally intensive of the three and has been the preferred approach in applications with relatively sparse track environments.

Criterion 3 generally has the highest computational burden, because the state estimates must be optimized over the collection of high-probability hypotheses by means of a sum over disjoint correlation hypotheses. This approach is, therefore, usually not used in the many multisensor problems in which the hypotheses are not disjoint. Reid[40] reports improved performance using a sequential Bayesian process with multiple, deferred hypotheses.

Each of these association criteria can be incorporated in a system that corrects or refines decision criteria based on external processes, as discussed in Sec. 3.3.6.

3.5.4   State Estimation

Once it is decided that tracks from separate sources (e.g., two sensor tracks or a sensor track and an MSI track) represent the same target, the next issue is that of combining these estimates to produce a fused estimate.

3.5.4.1   Optimal Techniques.   Assuming that a selected sensor data report $z(k + 1)$ has been correctly associated with a given (single- or multisensor) track, Kalman filtering provides optimal minimum mean-square error (MMSE) track updating. Such is appropriate in a hard-decision fusion structure, where track association is accepted unconditionally. The target state and covariance predictions are made per Eq. (3.12) and (3.13). The corresponding measurement and measurement covariance predictions are

predicted measurement:

$$\hat{z}(k + 1|k) = \mathbf{H}(k + 1)\hat{\mathbf{x}}(k + 1|k) \; , \tag{3.47}$$

measurement prediction covariance:

$$\mathbf{S}(k + 1|k) := E[\tilde{\mathbf{z}}(k + 1|k)\tilde{\mathbf{z}}^T(k + 1|k)|Z_k]$$

$$= \mathbf{H}(k + 1)\mathbf{P}(k + 1|k)\mathbf{H}^T(k + 1) + \mathbf{R}(k + 1) \ . \qquad (3.48)$$

The Kalman optimal linear update scheme (in the mean-squared error sense) is to combine the track state estimates $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{x}}_j$ measurements weighted by the inverse measurement prediction covariance and the covariance between the predicted state and measurement[8]:

$$E[\tilde{\mathbf{x}}(k + 1|k)\tilde{\mathbf{z}}^T(k + 1|k)|Z_k] = \mathbf{P}(k + 1|k)\mathbf{H}^T(k + 1) \ , \qquad (3.49)$$

filter gain:

$$\mathbf{W}(k + 1) = \mathbf{P}(k + 1|k)\mathbf{H}^T(k + 1)\mathbf{S}^{-1}(k + 1) \ , \qquad (3.50)$$

state update:

$$\hat{\mathbf{x}}(k + 1|k + 1) = \hat{\mathbf{x}}(k + 1|k) + \mathbf{W}(k + 1)v(k + 1) \ , \qquad (3.51)$$

covariance of state update:

$$\mathbf{P}(k + 1|k + 1) = \mathbf{P}(k + 1|k) - \mathbf{W}(k + 1)\mathbf{S}(k + 1)\mathbf{W}^T(k + 1)$$

$$= [\mathbf{I} - \mathbf{W}(k + 1)\mathbf{H}(k + 1)]\mathbf{P}(k + 1|k)[\mathbf{I} - \mathbf{W}(k + 1)$$

$$\times \mathbf{H}(k + 1)]^T + \mathbf{W}(k + 1)\mathbf{R}(k + 1)\mathbf{W}^T(k + 1) \ . \ (3.52)$$

New tracks are created as a function of hypothesis selection. The initial state $\mathbf{x}(0)$ of a new track is assumed to be Gaussian with mean $\hat{\mathbf{x}}(0|0)$ and covariance $\mathbf{P}(0|0)$.

In multisensor data fusion, two track state estimates $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{x}}_j$ may be combined using this MMSE optimal linear update scheme. These estimates are combined using measurements weighted by the covariances $\mathbf{S}_i$ and $\mathbf{S}_j$ and the cross-covariance $\mathbf{P}_{ij}$ as follows.

Let $\hat{\mathbf{x}}_i$ and $\hat{\mathbf{x}}_j$ be current target state estimates and $\mathbf{P}_i$ and $\mathbf{P}_j$ associated covariances from two sensors (or from a sensor and a prior MSI track file). Both estimates can be the result of current measurements or one can have been propagated to the current update time. One of the tracks $\mathbf{x}_i$ is selected as providing the "prior" mean. If the association is being performed in a fusion node involving an MSI track and a sensor track (or nontracked event report), then the MSI track estimate is taken as the prior. If association is for the purpose of creating a new MSI track, then the track from the more accurate sensor is selected.

For any dimensions (e.g., target range) that are estimated in the reported sensor track but not yet in the selected MSI track, the sensor estimates can be used for the updated fused estimates.

If the errors in the two track sources are statistically independent, then the fused estimate is simply

$$\hat{\mathbf{x}}_{ij} = \mathbf{P}_j(\mathbf{P}_i + \mathbf{P}_j)^{-1}\hat{\mathbf{x}}_i + \mathbf{P}_i(\mathbf{P}_i + \mathbf{P}_j)^{-1}\hat{\mathbf{x}}_j \ , \tag{3.53}$$

with covariance

$$\mathbf{P}_{ij} = \mathbf{P}_i(\mathbf{P}_i + \mathbf{P}_j)^{-1}\mathbf{P}_j \ . \tag{3.54}$$

If, however, the errors in the two track sources are not independent, the common process noise results in a cross covariance recursion

$$\mathbf{P}_{ij}(k|k) = E[\tilde{\mathbf{x}}_i(k|k)\tilde{\mathbf{x}}_j^T(k|k)]$$

$$= [\mathbf{I} - \mathbf{W}_i(k)\mathbf{H}_i(k)][\mathbf{F}(k - 1)\mathbf{P}_{ij}(k - 1|k - 1)$$

$$\times \ \mathbf{F}(k - 1)^T + Q(k - 1)][\mathbf{I} - \mathbf{W}_j(k)\mathbf{H}_j(k)]^T \ , \tag{3.55}$$

being a linear recursion with initial condition $\mathbf{P}_{ij}(0|0) = 0$.
 The state estimation fusion Eq. (3.53) then becomes

$$\hat{\mathbf{x}}_{ij}(k|k) = \hat{\mathbf{x}}_i(k|k) + [\mathbf{P}_i(k|k) - \mathbf{P}_{ij}(k|k)]$$

$$\times \ [\mathbf{P}_i(k|k) + \mathbf{P}_j(k|k) - \mathbf{P}_{ij}(k|k) - \mathbf{P}_{ij}^T(k|k)]^{-1}$$

$$\times \ [\hat{\mathbf{x}}_j(k|k) - \hat{\mathbf{x}}_i(k|k)] \ , \tag{3.56}$$

and the corresponding covariance of the fused estimate is

$$M_{ij}(k|k) = \mathbf{P}_i(k|k) - [\mathbf{P}_i(k|k) - \mathbf{P}_{ij}(k|k)]$$

$$\times \ [\mathbf{P}_i(k|k) + \mathbf{P}_j(k|k) - \mathbf{P}_{ij}(k|k) - \mathbf{P}_{ij}^T(k|k)]^{-1} \tag{3.57}$$

$$\times \ [\mathbf{P}_i(k|k) - \mathbf{P}_{ij}(k|k)]^T \ .$$

Brandstadt[24] has employed a probabilistic data association approach to extend the Kalman filter update scheme to cases in which there is some uncertainty that data applies to a given track; i.e., where $P[\mathbf{x}_i|\mathbf{z}(k + 1)] < 1$. The association uncertainty can be incorporated into a multisensor tracker by scaling the filter gain $K_i(k)$ by the probability of correct association, substituting the following for Eq. (3.50):

$$\mathbf{W}(k + 1) = P[\mathbf{x}_i|\mathbf{z}(k + 1)]\mathbf{P}(k + 1|k)\mathbf{H}^T(k + 1)\mathbf{S}^{-1}(k + 1) \ . \tag{3.58}$$

3.5.4.2 Suboptimal Techniques.

The Kalman filter—particularly the calculation of cross-covariances $\mathbf{P}_{ij}$—is a computationally intensive process. Often, when the sensors involved are dissimilar, it can be assumed that the cross-covariance is negligible and the procedure of Eqs. (3.47) to (3.52) is employed. Further reductions in processing load are attainable with various suboptimal techniques. The extended Kalman filter applies track filtering techniques to nonlinear systems by means of linear approximations of nonlinear dynamics.

In some cases, updates are required at a fast rate (for instance, in warning of an immediate attack). These updates can be performed more often than the track association procedure itself. If an update is required for a given MSI track, then the most recent sensor estimates can be fused using the latest track association hypothesis. It is not generally necessary (although often preferable from a performance viewpoint) to redo the track association procedure each time a track update is required.

3.6   MULTISENSOR IDENTIFICATION

This section concerns techniques for multisensor feature extraction and target identification. We use *identification* in the broad sense to include target recognition and classification as well as unique individuation.

Because such techniques are widely discussed in the literature (although primarily in the context of single-sensor data processing), we attempt no rigorous definition or thorough survey of the relevant techniques here. Rather, we concentrate on their potential for use in sensor data fusion.

An ID technique's suitability for data fusion applications depends on the following factors.

A Priori **Data Assumptions.**   Feature- and pixel-level target identification presuppose some sort of model by which observed features allow a given target to be discriminated from others. Decision-level fusion must model the decision performance of its contributing sensors. Therefore, every ID technique must be supplied with *a priori* data of some sort, providing expectations concerning targets to be discriminated and the sensor measurement process. Depending on the specific technique employed, these may include parametric probability distributions, conditional probability estimations, and causal or structural models of expected targets. A need for extensive or exhaustive modeling can limit the practical application of particular techniques.

Robustness of Performance with Ambiguous, Contradictory, or Sparse Sensor Data.   This is the complement of the first factor, which concerns robustness with ambiguous, contradictory or sparse *a priori* data. The types of sensor features to be exploited may differ in their universality, or invariance over target operating states and orientation and environmental conditions. For example, historical statistics of signal parameters may not suffice to identify an emitter with pseudo-random parametric agility or reserved "war modes."

Ability To Exploit Diverse Data Types.   The applicability of a technique to multisource data depends on its ability to exploit all the available forms of data; in particular systems, these may include combinations of (1) pixel, feature, and decision-level data; (2) imagery and signal data; (3) data streams with commensurate and noncommensurate features; (4) quantitative parameters as well as discrete (or symbolic) features; (5) features from uncalibrated sources; (6) situational and contextual information (e.g., temporal, spatial, causal and operational relationships between objects and the background environment); and (7) higher level relationships/structures among features and objects.

System Implications.   System factors include requirements on processing, memory and communications hardware, and opportunities for enhancements using special-purpose architectures, e.g., parallel or distributed processing.
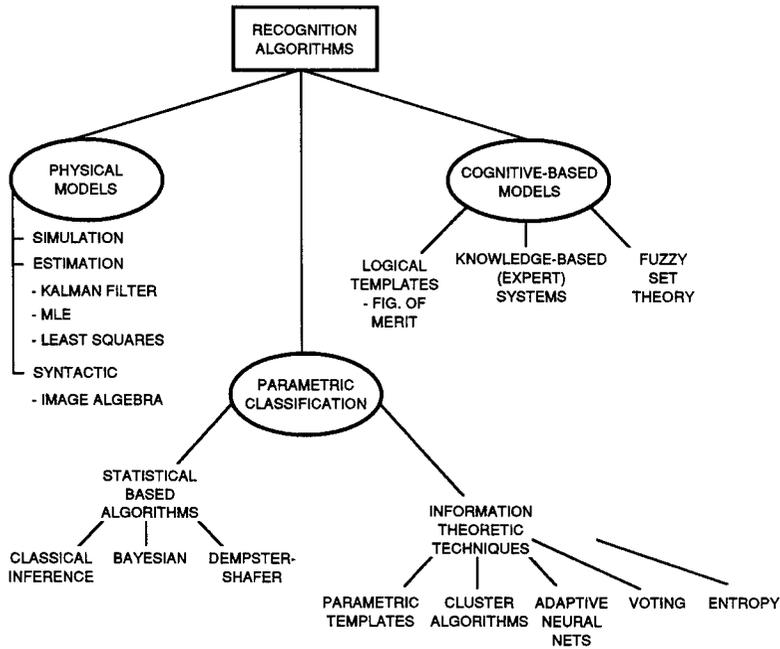
Fig. 3.19   Taxonomy of classification algorithms (from Ref. 7, © Artech House).

Hall and Buede (cited in Ref. 7) have suggested a taxonomy for object identification techniques applicable to multisource data processing. As shown in Fig. 3.19, a distinction is made among (1) techniques that estimate identities *parametrically* (either based on statistical parameter matching on information-theoretic metrics), (2) techniques that identify by comparing reported data with *physical models* of candidate objects, and (3) techniques that emulate higher level *cognitive* processes in identifying objects.

Several techniques, including most statistically based algorithms, are essentially *decision-level* processes: they operate by combining multisource declarations (e.g., target ID declarations), which may be mutually contradictory or incomplete, to infer a decision (e.g., MSI target ID declaration).

Other techniques operate at the *feature level* to infer a target identity or characterization based on measured characteristics. Such techniques include physical models and many implementations of cognitive-based processing.

Finally, there are feature-extraction techniques, which may operate at the *pixel* (or raw-data) *level*. These include various estimation and clustering techniques.

Neural networks have been used to solve both feature and classification problems and hold promise as a means to integrate signal, statistical, and symbolic methods.

In general, statistical parametric methods, although allowing optimal performance (in a MAP sense), are limited by the difficulty of estimating probability density or mass functions. Such methods presume knowledge of parametric distributions, which is often not available in real-world applications.

Nonparametric methods, such as estimation and clustering techniques, often require large numbers of samples. In many cases, processing burdens limit probability density estimation to a small number of dimensions.[41]

Neural network methods are less restrictive in their need for prior statistics and have been shown to operate in very noisy data environments. Their application to target ID, and to pattern recognition in general, has been limited by (1) the processing burden (which is alleviated by the increased availability of massively connected systems) and (2) the lack of explicit performance information or optimality criteria.[41]

In short, an effective system for object identification in nontrivial applications with complex, noisy environments will probably need to use a combination of techniques for various aspects of the problem.

For example, a hybrid identification scheme may use a knowledge- or template-based model to establish the contextual and structural relationships which define the *a priori* data required for target ID by feature extraction or library matching techniques (prior probabilities, parameter statistics, cluster dimensions, discriminant surfaces, etc.).

Clustering or neural network techniques can be used in developing this identification model, assisting the designer in defining measurement-level or more abstract features significant to target detection, recognition, and discrimination. These features can be exploited in-mission either using neural net recognizers or by symbolic or parametric feature extractors; the selection depends on such factors as required response time, performance confidence, processing resource availability, and the quality of *a priori* and sensor data.

A knowledge-based system can serve as a shell for the in-mission identification process, activating specific estimation, statistical, logical, or heuristic processes, as required, to resolve the ID/assessment problem at hand.

These findings are discussed in the following subsections. Additional treatment of processing techniques as they apply to fused ID and assessment may be found in Refs. 7, 17, and 42. Standard sources with more general applicability include Refs. 18 and 43 to 46. Jordan and Choe[41] compare pattern recognition techniques in a formal framework, showing advantages and disadvantages of each.

3.6.1   Parametric Classification

Feature matching is by far the most commonly used process for target identification. Observed feature measurements are compared by one means or another with a library of features for candidate objects or object types. Because features can encompass both discrete symbolic attributes (e.g., two-engined) and continuous variables (such as target emissivity or cross section), parametric classifiers tend either to evaluate the class membership either by parametric distribution or by nonparametric statistics.

A target's unique identity is associated with a set of features that are assumed to apply uniquely. Classes may be defined to correspond to general descriptive characterizations having some significance to the ID system [e.g., "MiG-29 carrying semiactive RF air-to-air missiles," "beyond visual range (BVR) air superiority fighter," and "two-engine jet fighter"].

Often, a straightforward disjunctive inheritance hierarchy is employed, such as that shown in Fig. 3.20. [Abbreviations used in the figure include within
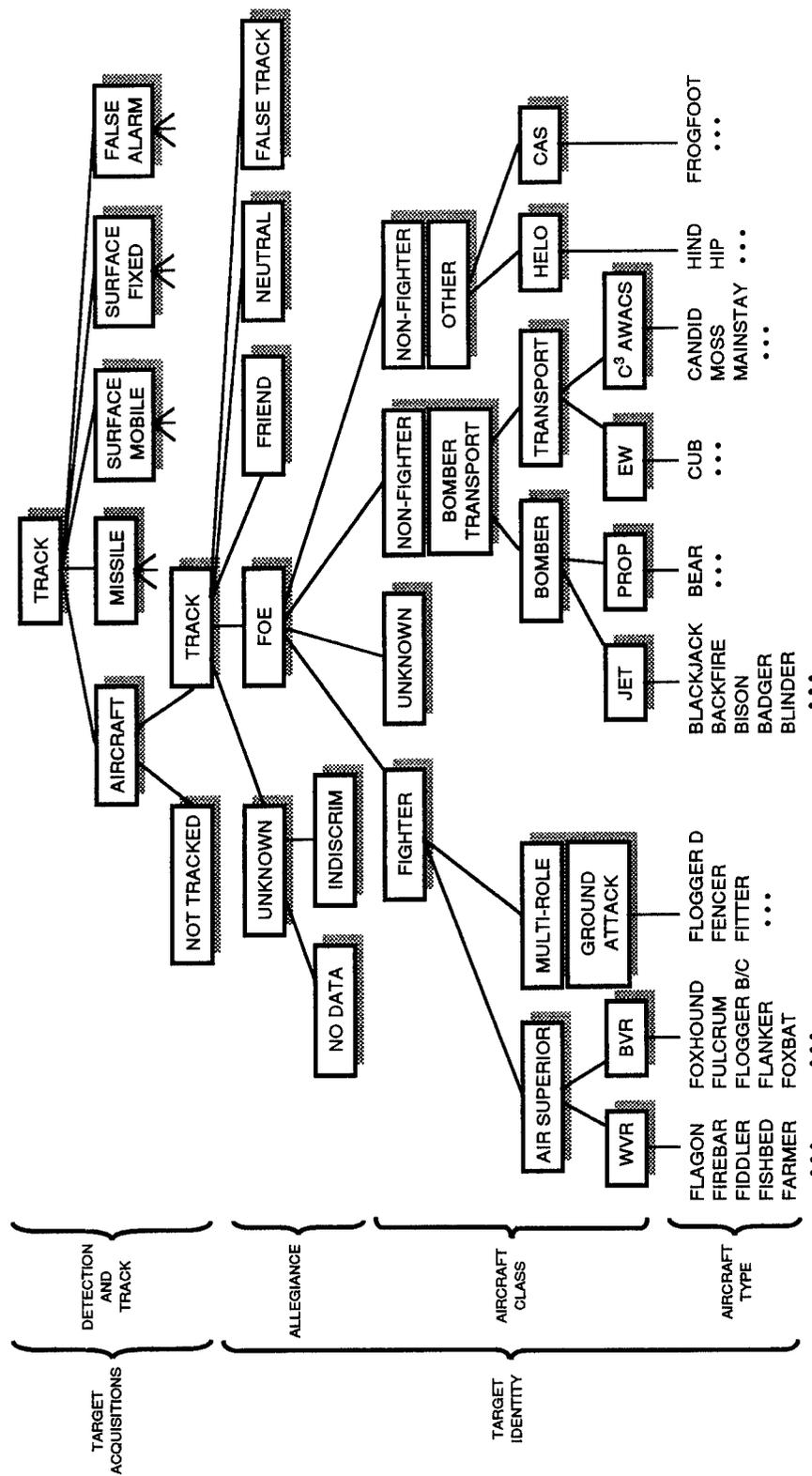
Fig. 3.20 Classification/characterization hierarchy example (adapted from Ref. 51, ©1986 IEEE).

visual range (WVR), early warning (EW), helicopter (HELO), and close air support (CAS).] In such a model, the characterization associated with a lower node provides sufficient conditions for inferring its parent node. More elaborate classification schemes allow multiple inheritances, so that, e.g., both friend and foe fighters could be subordinated to a common set of identifying features.

Two observable features (measurable in the same or in separate sensors) may complement one another by resolving a target's classification to the intersection of two higher level categories. For example, an identification friend or foe (IFF) system may indicate a target to be non-NATO, whereas an imaging infrared (IIR) may identify it as a two-engine jet aircraft and a radar warning receiver (RWR) might provide an ambiguous identification as either a MiG-29 or a particular type of ground radar.

3.6.1.1 Statistics-Based Techniques. We can follow Hall and Buede in grouping classical inference, Bayesian, and evidential reasoning (Dempster-Shafer) methods as basing identification decisions on statistical models of the classes of interest. More recent developments of this type include the generalized evidence processing of Thomoupoulos and Zhang,[47] Thomoupoulos et al.,[48] and Thomoupoulos and Okello[49] and Fixsen-Mahler's extension of the Dempster-Shafer technique.[50,98]

Classical inference techniques assign an observation to one or another class as the basis of the observation's relation to the probability density function (PDF) associated *a priori* with each candidate class.

A discriminant surface is selected between the PDFs of each pair of competing classes H_1 and H_2 to minimize the probability of misidentification errors $P(\hat{x} \in H_1|H_2)$ and $P(\hat{x} \in H_2|H_1)$. These errors are associated with the regions on either side of the discriminant boundary. Classical inference techniques are conceptually and computationally simple. However, they do depend on the availability *a priori* of probability densities, become complex for multivariate data, and can assess only two hypotheses at a time.

A *Bayesian* inference technique selects the maximum *a posteriori* probability (MAP) hypothesis on the basis of sensor reports R (or evidence, generally) and prior likelihood estimates. As illustrated in Eq. (3.35), the Bayesian decision rule is to select the hypothesis H_j, which maximizes

$$P(H_j|R) = \frac{P(R|H_j)P(\mathrm{H}_j)}{\sum_{\forall i} P(R|H_i)P(\mathrm{H}_i)} , \quad \text{where } \sum_{\forall i} P(H_i) = 1.$$

The Bayesian technique has the advantage over the classical approach in that it does not require definition of probability density functions. Therefore, it can be applied to discrete as well as parametric features, and to decision-level data fusion.

The Bayesian technique does require a definition of the hypothesis space into a set of mutually exclusive and exhaustive hypotheses H_1, ..., H_N and requires an assignment of prior probability to each such hypothesis. It also requires assumption of conditional (*a posteriori*) probabilities $P(R|H_i)$ of

each observed type of evidence (e.g., sensor report) conditional on each hypothesis H_i.

Therefore, the Bayesian approach suffers from artificiality in the many cases where there is no basis for assigning prior probabilities. Also the requirement that the competing hypotheses be mutually exclusive yet exhaustive tends to force system designers to create artificial classification categories, ignoring potentially significant dependency relations in natural or operationally significant categories.

Evidential reasoning was developed by Dempster and expanded by Shafer as a method to represent uncertainty in *a posteriori* likelihood estimation.[17,51,52] Principal advantages of the Dempster-Shafer (D-S) formulation over Bayes's are in the intuitive representation of evidential certainty and the ability to represent logical dependencies among hypotheses.

Allowable D-S hypotheses constitute the power set 2^{Ω} of a *frame of discernment* Ω, defined as a set of elemental propositions (i.e., mutually exclusive and exhaustive descriptions relating to a universe of discourse). Each elemental proposition is assigned an evidential "mass" (or weighting) $m(x) \leq 1$. Thus, for example, the identity of a sensed target x is represented by a D-S *body of evidence* $B_x = \{m_1(H_1), ..., m_a(H_a), m_{\ominus}(\ominus)\}$, where $m_{\ominus}(\ominus)$ is the weight of the null hypothesis $\ominus$ (namely, the degree to which we believe that none of the propositions H_i applies to x). This reduces to the Bayesian formulation in cases where the propositions in B_x constitute the elemental set Ω and $m_{\ominus}(\ominus) = 0$ (i.e., where there is no uncertainty as to the assignment of evidential masses). Evidential weightings for two bodies of evidence

$$B = \{m_i(H_i)|i = 1, ..., a\}, \qquad C = \{n_j(K_j)|j = 1, ..., a\} \ ,$$

are derived using Dempster's rule of combination:

$$r_{ij} = m_i n_j \frac{\alpha_{DS}(H_i, K_j)}{\sum\limits_{i=1}^{a} \sum\limits_{j=1}^{b} m_i n_j \alpha_{DS}(H_i, K_j)} \ , \tag{3.60}$$

where the D-S *agreement function* $\alpha_{DS}(H_i, K_j) = 1$ if H_i and K_j are consistent; 0 otherwise.

D-S evidential mass reduces to Bayesian probability in cases where all hypotheses considered are mutually exclusive and exhaustive over a given universe of discourse (i.e., where there is no general level of uncertainty).

Integration into a fusion system involves relating association metrics to probability masses. For many commonly used association methods, such as the k-nearest neighbor method (Sec. 3.5.2.1), it is not obvious how to make meaningful mass assignments. Hunt et al.[55] have developed methods for such assignments. Given a distance ordering among feature vectors yielding the k target tracks nearest to a sensor report feature vector, a mapping is made to D-S mass vectors as follows. Each candidate association $H_i = (\mathbf{x}_i = \mathbf{x}_s)$, $i = 1, \ldots, k$, contributes the same total mass m_i, which is distributed among the various disjunctive association hypotheses (subsets of Ω); m_i is split between

its own singleton classification hypothesis H_i and the disjunction of all remaining singleton hypotheses:

$$m_i = m_0 = m_i(H_i) + m_i\left(\bigcup_{j=1}^{k} H_j - H_i\right) , \quad j \neq i .$$ (3.61)

The self-contributing mass fraction $m_i(H_i)$ increases monotonically as a function of the feature distance d_{iS} relative to the other candidates d_{iS} (although the weightings of this function are not specified in the reference). Simulation performance with this approach is reported to be appreciably better than attained with individual sensors and marginally better than a Bayesian approach (with priors assigned).[55] On this basis, the mass assignment technique can allow the nonparametric nearest neighbor method to be integrated into a viable multisensor classifier, one not dependent on the availability of the *a priori* statistical information required by a Bayesian process.

The Bayesian and D-S formulations have complementary shortcomings: Bayesian evidence theory involves representation for uncertainty in *a priori* knowledge but has no way to represent evidential uncertainty; the converse is true for D-S theory. Fixsen and Mahler[50] proposed a modification of D-S to relieve these difficulties and Mahler[98] has recently produced a most satisfactory generalization, called *conditioned Dempster-Shafer* (CDS). CDS is applicable to cases of both ambiguous evidence and ambiguous priors (and, indeed, to several models of uncertainty). This is achieved by generalizing the D-S agreement function of Eq. (3.60) in terms of different Sugeno uncertainty measures ρ:

$$\alpha_\rho(H,K) = \frac{\rho(H \wedge K)}{\rho(H)\rho(K)} ,$$ (3.62)

where "$\wedge$" indicates logical conjunction and a Sugeno uncertainty measure is some function ρ defined on all subsets of 2^Ω such that $S \subseteq T \subseteq 2^\Omega \Rightarrow \rho(S) \leq \rho(T)$, $\rho(\ominus) = 0$, and $\rho(2^\Omega) = 1$. This formulation reduces to orthodox D-S if we define $\rho(H) = 1$ if H is consistent; 0 otherwise. Mahler[98] shows that an equivalence to Bayesian probabilistic combination is obtained if ρ is defined as a probability measure with a distribution over Ω. However, CDS is more powerful than either Bayes or D-S in that other Sugeno functions ρ may be chosen to represent any distribution of uncertainties concerning *a priori* knowledge or *a posteriori* evidence.

The D-S formulation operates on a richer domain of hypotheses than does Bayes's: 2^Ω (i.e., all possible partitionings of Ω) rather than directly on Ω (i.e., on a single exhaustive partitioning). Therefore, it tends to be much more computationally intense. Gordon and Shortliffe[47] have adapted the D-S approach to improve computational efficiency. They do so by partitioning the reasoning space 2^Ω into a hierarchical hypothesis structure, thereby constraining the required search. Fixsen and Mahler[50] and Maurer et al.[99] have defined entropy-based similarity metrics among hypotheses, which allow relatively similar hypotheses to be combined with a minimal loss of information content. The combinatorial explosion can thereby be constrained, by replacing a large body of evidence with one of any required smaller size, which minimally increases entropy.

3.6.1.2   Information Theoretic Approaches.   Several nonstatistical methods have been employed in single- and multisource identification to map from observable features to object ID or characterization. As such, they can be of use in cases where the stochastic properties of the observable data are unavailable. Many of these techniques are adaptive to the characteristics of the observed data, including various clustering methods, entropy-based algorithms, and adaptive neural networks.[57,58]

Information-theoretic techniques involve the use of a similarity metric, or measure of closeness in measurement space. Therefore, some sort of adaptation is needed to apply such methods to discrete, or symbolic, features. Zadeh's[59,60] fuzzy logic is one applicable method, as discussed in Sec. 3.6.3.

Often such techniques find use in developing classification paradigms by detecting characteristic and discriminating properties of the target set. For example, the appearance of a new type of combat aircraft developed by a hostile country requires modifications to the ID processes used in threat warning, air defense, and ATR systems. Parametric cluster analysis is useful in revealing characteristic differences between the new target and others. These discriminants in the observable data can be used to refine the ID paradigms to include the new target type.

Clustering Methods.   Clustering is used in many applications to develop models and, therefore, discrimination criteria, within a feature space. These statistically nonparametric methods are related to the so-called unsupervised or feature-based techniques of pattern recognition.[7,43,61]

Desarathy[44] has defined a taxonomy for clustering techniques, employing a four-dimensional vector: density/similarity, aggregative/disaggregative, crisp/fuzzy, and self-learning/exhaustive search/user input.

Density-based approaches, not being sensitive to individual samples, are suited for high-volume-data scenarios, involving low-dimensional feature space.

In contrast, similarity-based approaches are applicable to higher dimensionality feature spaces and can be made sensitive to individual outliers. They are more computationally intensive, and therefore are more limited to moderate volume scenarios. Aldenderfer and Blashfeld[62] have identified four major types of similarity metrics:

1. correlation coefficients, such as Pearson's product-moment correlation coefficient
2. distance measures, such as Euclidean and Mahalanobis distance
3. association coefficients for binary variables, such as Jaccard's or Gower's coefficient
4. probabilistic similarity coefficients.[7,62,63]

Entropy Methods.   Information-theoretic techniques can be used to establish the information content of sensor data. Such evaluation is important in establishing quantitative metrics for evaluating the contribution of individual sensor measurements or fusion nodes to target ID. Entropy-based performance measures are discussed in Sec. 3.9.1.

As with clustering methods, entropy-based classifiers make no attempt to model the structure or utility of observable features. Rather, feature sets are evaluated only in terms of their rarity.

Neural Networks. Neural networks are processing structures designed to detect and associate patterns in data in ways similar to those used by biological nervous systems. Neural networks involve simple processing nodes that are massively interconnected, each node performing a nonlinear transformation on its inputs. The weightings that a node applies to its inputs are subject to modification via *learning functions.* A diversity of network structures, node transform, and learning functions have been examined.[64] These include designs where learning is subject to supervision (e.g., Hopfield and Hamming nets) and those in which the system learns in an unsupervised manner (e.g., systems by Carpenter-Grossberg and Kohonen).

Neural networks provide the advantage of being relatively unconstrained by limitations in *a priori* statistical or parametric information. They also can be suitable to operation in dense, noisy environments: a given neural structure is relatively immune to the problem of combinatorial explosion, trading off resolution as the quantity of competing hypotheses grows.[28,65]

Recent applications of neural networks to data fusion include ship[66] ID, sonar pattern recognition,[67] and aircraft[28] ATR. Bowman[28] presents early results indicating that neural nets can outperform traditional clustering techniques for multisource classification, particularly in instances of noisy or missing sensor data. Other defense and fusion applications are discussed by North[68] and cited in Ref. 7.

3.6.2 Physical Models

Physical model ID attempts to select the best causal model consistent with the observed data. Techniques of this type include simulations, syntactic image models, and such estimation techniques as Kalman filtering.

Estimation techniques have very limited applicability to the mainly symbolic data typical of feature- and decision-level ID problems. However, they may have a place in feature extraction in pixel (or data) level data fusion. For example, estimation can be useful in inferring ID features based on track kinematics; e.g., extracting a missile's guidance type (and, therefore, to resolve ID) from its multisensor track history.

Predictive models based on the underlying physical characteristics of potential targets are especially applicable where high-confidence *a priori* target parametric data are unavailable. This can occur either where the required intelligence data is lacking or where targets may employ signature control or signature masking techniques.

In such cases, target ID may require relating observed data to modeled intrinsic physical limits. For example, the beam width and frequency range of a target radar system are restricted by the basic physics of its antenna aperture.

Track association, as well as ID, may sometimes require reliance on hypotheses concerning underlying physical causes. This can occur when errors in the kinematic state are highly correlated between sensors; e.g., in associating multiple target reports provided by passive physically separated sensors (for which instantaneous range estimates are highly ambiguous).

If the sensors are such that reported features are not directly comparable, track association may yet be predicated on consistency in higher level features inferred from the diverse sensor data. At best, these will be definable features that are stable and predictable, per Harney's[14] criteria listed in Sec. 3.3.1.

Inferred multisensor features, applicable to aircraft track association and ID include (1) mass/aerodynamic features such as drag and lift coefficients; (2) multispectral imagery, to include mutually registered multispectral radar cross section (RCS), thermal and irradiance maps; and (3) characteristic object spatiotemporal interrelationships (e.g., flight formations and tactics).

3.6.3   Cognitive-Based Models

There are a number of approaches that attempt to replicate, in one way or another, the processes by which humans recognize or classify objects. These techniques employ symbolic representations of universes of discourse and either rigorous or heuristic inference approaches. Correspondingly, such approaches have taken the form of representing the structure of pertinent data sets either (1) in a closed form logical structure (such that the likelihood of membership of an object in various classes is strictly determined by the deductive inference structure, given the available data) or (2) in the form of an open set of heuristics to be searched by an *inference engine* control process to reach a suboptimal but generally sufficient conclusion.

Templating Methods.   Traditional logical templating methods evaluate sensor data against a pattern of logical conditions for membership in each candidate class. Diverse inferencing techniques can be employed as appropriate to apply diverse types of information to problems of data association, ID, or situation assessment. The logic employed may be traditional bivariate first-order predicate calculus, or generalizations thereof, either to allow greater power in temporal reasoning[69,70] or to incorporate considerations of uncertainty in the deductive process.

Templating can be extended to organizational hierarchies or other complex entities; e.g., command and control structures or integration levels within a complex system. Diverse *ad hoc* methods for combining information using the relational structures in the sensed environment have been developed under the name of figure of merit (FOM) techniques. Such inference procedures apply to object-oriented situational data bases; allowing diverse contextual information to be used in solving local problems, such as estimating the ID, state, or activity of individual objects.

By defining entities in terms of relational structures, likelihood measures (or figures of merit) can be calculated for alternative ID or state hypotheses. The general procedure is to establish a network of dependency relationships among entities of interest, allowing the inferencing of entity state or event relationships.

The approach is eclectic in that it allows different inference processes to be employed as appropriate to specific types of problems: parameter measurement, data association, target ID, characterization of higher level systems or organizational structures, prediction of event likelihood and timing, etc. Figures of merit (which can be based on diverse modes of inference) combine and propagate through the relational network by a weighted sum process. In this way, data of diverse types as well as information beyond local measurements and nonstatistical estimates of conditional likelihoods can be applied.

To be sure, such systems of inference are not optimal in any sense, nor are they necessarily internally consistent. Therefore, their validation is a matter of considerable simulation and continual refinement.

FOM techniques also allow the detection of higher level patterns of activity and therefore find use in situation assessment and event prediction (DFS fusion levels 2 and 3). FOM techniques have been employed in several generations of U.S. Army tactical intelligence systems, including battlefield exploitation and target acquisition (BETA), LOCE, and all source analysis system/enemy situation correlation element (ASAS/ENCE).

As an example, the BETA system employs an entity hierarchy for relating received data to the overall tactical intelligence picture. The hierarchy includes single elements (e.g., individual assemblies—radios, guns, engines), compound elements (composites of simple elements such as weapon platforms), and complex elements (composites of simple and compound elements—artillery batteries, command posts, etc.).[71,72]

Fuzzy Logic.   As a generalization of modal logics applicable to target ID, that developed by Zadeh from his fuzzy set theory allows representation not only of uncertainties deriving from noisy or incomplete data but also of the uncertainty in the conditions for membership in one class or another.[59,60]

Information-theoretic techniques involve the use of a similarity metric, or measure of closeness in measurement space. Therefore, some sort of adaptation is needed to apply such methods to discrete, or symbolic, features. In many cases, a fuzzy set representation can provide the wanted link between symbolic and parametric inference: degrees of applicability of a description and relative similarities between objects in nonparametric features can be represented in a closed-form deductive system.

In fuzzy set theory, the strict binary determination of set membership is replaced by a function $\mu A(x)$, which defines the degree of membership of an entity x in a set A. The notion of "degree of set membership" is subject to more than one interpretation. For example, $\mu A(x)$ can be taken to signify (1) the possibility, or likelihood in some nonstatistical sense, that x is a member of A or (2) the degree to which the description (or attribution) A applies to x.

Under the former interpretation, the membership function can be employed to represent the uncertainty in assigning noisy data to one or another discrete class. The decision boundary in such a case may itself be drawn very precisely, as in the problem of determining whether an aircraft with inaccurate navigation equipment has strayed across a well-defined international boundary. In this interpretation, Zadeh's membership function $\mu A(x)$ is a probability density function and can be seen as a generalization of Bayesian uncertainty.

In the latter sense of degree of membership, $\mu A(x)$ allows a logical representation of imprecise natural-language descriptions such as *large, close, blue,* etc. Zadeh and others have suggested several intuitively or behavioristically derived density functions to capture the heuristics involved in applying such fuzzy terms.[59,60]

Knowledge-Based Methods.   The field of artificial intelligence (AI) has developed sophisticated techniques for representing higher orders of intelligence and knowledge in estimation and control systems. These include inference, search techniques, and knowledge representation schemes that relate directly to the integration of information.[42]

In particular, so-called knowledge-based system architectures have been developed. These are characterized by the strict separation of *knowledge* from

intelligence or reasoning ability. Intelligence is embodied in a so-called inference engine; knowledge resides in a knowledge base.

Such a structure allows information to be represented in a variety of ways, which give the potential for very diverse, complex realms of discourse to be modeled and reasoned about in relatively natural and insightful ways. Among representation schemes that have potential use in various levels of multisource data fusion are the following:

1. *Production rules:* roughly equivalent to an implication in propositional logic. Typically, a production rule has the form IF (evidence) THEN (hypothesis). Useful techniques include *forward chaining,* i.e., moving from a set of evidence through a sequence of production rules to reach a conclusion, and *backward chaining,* i.e., searching through sequences of rules to find evidence that will support a given hypothesis.[42] Generally, production rules map into a classical propositional or first-order predicate logic. However, more powerful logical frameworks can and have been employed, including various modal logics, temporal logics, higher order predicate logics, and nonmonotonic (i.e., not necessarily consistent) logics.[42,73]

2. *Frames:* data structures that group knowledge about particular objects and situations. Objects of various types are modeled in terms of characteristic structures, with distinguishing (or otherwise useful) attributes maintained in specific nodes (called *slots*).

3. *Semantic nets:* networks of frames, representing relationships among objects at various levels of abstraction.

4. *State transition networks:* finite state automata (dynamic model) representations of systems that are subject to change over time. States are represented by nodes; arcs between nodes associated with logical statements represent conditions for state transition. The use of this type of representation is illustrated in Sec. 3.8.5.

Such flexibility of representation permits a knowledge-based system to integrate and use diverse types of knowledge and to do so in ways that were not necessarily foreseen by the system designer. Frames or other object-oriented structures can be used to represent the system's current estimate of perceived objects in its environment: their identity, current operating mode, and kinematic state. Other types of data structures may allow a system to exploit such diverse aspects of the situation as

1. the schedule and priorities of various mission objectives
2. the time of anticipated events
3. expectations concerning detectable objects (potential targets) in the mission environment
4. objectives, priorities, tactics, communication, and coordination of such objects
5. these systems' situational awareness
6. the quality of received sensor data
7. the expected outcome of planned or anticipated actions, including the quality of data expected with a proposed sensor assignment/allocation.

Knowledge-based systems allow the use of any of the formal inferencing schema discussed. They are also well suited for implementing heuristics; e.g., in sit-

uations where prior probabilities or likelihood relationships are difficult to quantify in any consistent way. Early knowledge-based *expert* systems, such as Mycin[74] and Prospector,[61] used informally applied *certainty factors* to represent degrees of belief. These led to inconsistent (i.e., noninvertible) procedures, which were judged to be acceptable in applications where individual probability assignments tended to be very small.[42]

As shown in Fig. 3.15, hypothesis evaluation for data association or target ID can often benefit by applying very diverse sorts of contextual data. A blackboard-type of knowledge-based architecture can be employed to implement the required flexibility in the processing system (as in the revised DFS data fusion model shown in Fig. 3.6).

In a blackboard implementation, all knowledge relating to a broad problem domain is embodied in individual knowledge sources (KSs). A KS can be viewed as an independently acting object, repeatedly examining a common data area— the blackboard—to determine whether the conditions for the KS's execution are present. The blackboard involves a knowledge space containing both permanent knowledge and current hypotheses concerning the threat environment. The blackboard also contains information that determines which KSs to trigger. Message passing is used for interfacing among control processes.

A typical approach to interpreting real-world information is to use a combination of data-driven (forward-chaining) techniques and model-driven (backward-chaining) techniques. As Garvey[42] notes, model-driven approaches tend to be most useful when the models are restrictive and the data is noisy. Noisy data requires effective data acquisition and data management techniques based on an understanding of the system's goals and its environment. Methods for adapting the collection management process to these factors are treated in Sec. 3.8.

Another virtue inherent in the modular character of the system's knowledge base is the capability to develop prototype system designs rapidly, allowing evaluation and refinement of candidate decision processes.

3.7 MULTIPLATFORM FUSION

The discussion so far has been generic as to the spatial collocation of sensors involved in fusion. We now consider the particular issues and benefits involved in the integration of spatially separated sensors.

Benefits for integrating data collected from spatially separate platforms include (1) the ability to maintain track continuity, despite obstruction by the earth's horizon or other local obstacles or obscurants (including, in some cases, intentional sensor countermeasures); (2) the added resolution in kinematic state estimate and ID by the use of complementary sensors (i.e., with independent errors); (3) the opportunity for state estimation resolution afforded by parallax and (if applicable) independent motion of the sensor platforms; and (4) opportunities for semiactive operation (i.e., one platform passively processing the reflected energy from a target illuminated by another platform).

Issues of particular concern in multiplatform data fusion, in addition to those involved in single-platform fusion, are:

Spatial Alignment. The alignment problem with spatially separated sensors differs from that with collocated sensors because of the greater difficulty in

boresighting. The problem of track association caused by uncertainties in sensor alignment has been a major obstacle in developing effective fusion systems involving distributed sensors.

The U.S. Air Force Rome Laboratory has for several years been conducting data fusion experiments, using fixed ground-based radar and infrared sensors across the northeastern United States to track aircraft (including both cooperating and incidental aircraft traffic). Although a highly successful program, numerous problems have had to be overcome relating aligning sensor data.[75]

It is anticipated that the availability of highly accurate GPS positional data will permit improved common referencing of spatially separate sensors by mutual tracking of known geographic features or cooperative targets.

Temporal Alignment.   Bandwidth constraints, packet switching delays, and speed of propagation all contribute to reporting latencies that must be accounted for in data association. For example, delays of several seconds are to be expected in data passed across a joint tactical information distribution system (JTIDS) network.

Constraining Communications Traffic to the Available Channel Capacity. An important consideration for a distributed sensor system is the communication involved in transferring information between the sensor platforms. This can lead to a trade-off between allocating system resources for computation versus communications. Also, the potential cost of transferring all sensor reports and the benefits of performing local data processing tends to favor the sensor-level tracking approach. Compromise techniques have been proposed in which selected observations are passed among sensor platforms, allowing the tracking accuracy achieved by multiplatform fusion to be attained by distributed data processing.[47-49]

Reducing Correlated Errors in Similar Sensors.   Many applications involve networks of similar sensors. In such cases, sensors and associated processing are subject to correlated measurement and decision errors. For example, deception countermeasures such as spoofing or signature control will likely produce similar errors in each reporting radar in an internetted air defense system. Techniques must be employed to prevent correlated errors from reinforcing one another in a multisensor system, thus degrading the performance attainable in an individual sensor. In one approach,[13] each reporting platform flags its internetted reports to identify the contributing sensor process and enabling the use of maximum-entropy techniques in associating parametrics from similar sensors. For dissimilar sensor contributors, independent Bayesian statistics can be employed [per Eq. (3.35)].

Calibration of Sensor Declarations.   A practical issue involved in integrating data from multiple processing systems is that of weighing the validity of each measurement and decision source. Systems that have been developed independently may provide the benefit of noncorrelated errors, but often at the cost of uncertainty in the quality of the reported data. Such systems will likely differ in their detection, track association and discrimination thresholds, their parameter measurement accuracies, and tracking filter characteristics. Communication channel constraints usually preclude detailed reporting of factors that contribute to errors in individual measurements. Often confidence values

are tagged to reports or even to individual measurements, but the heuristics used to assign such values are typically not applied consistently across multiple sensor platforms.

3.7.1 Multiplatform Fusion Architectures

The designer of a multiplatform data fusion system must decide the following architectural issues:

Distributed Versus Centralized Data Base.   One common approach to multiplatform fusion involves each site maintaining two sets of track files. A global track file (GTF) contains tracks on all targets, with track information being communicated from other sites. A second, local track file (LTF) contains tracks on those targets that are being actively tracked by that site. Thus, the global track file is a fused composite of the local track files, which gives each site an overview of the entire situation. It aids in local sensor allocation and in the determination of which targets should be included in the local track file. (Of course, global is a misnomer; there may be a hierarchy of target tracking data bases, each corresponding to the current state of knowledge of a particular node in a C^3 structure.[47–49])

Should each platform maintain both a local, autonomously derived, LTF as well as a GTF? Should all tracks be maintained at all sites or should each site only maintain a selected subset of target tracks?

Distributed Versus Local Data Fusion.   Should each sensor platform perform association of its LTF to the GTF currently available to that platform? Or should such association occur only in a central processing node or nodes?

Communication of Autonomous Versus Updated Global Data.   If LTF-to-GTF association is performed in each sensor platform, should the result be disseminated the network in the form of a corrected GTF, or should LTF (with GTF association) be passed?

Local Use of Global Versus Local Data.   How is each platform to use the GTF and its LTF in performing subsequent track association and processing? The limited perspective of a local track file is often mitigated by its currency compared to received global data. Data inconsistencies between old data received at a given net node and current local data can be resolved either by a distributed set of heuristics or by restricting the fusion tree structure to a strict hierarchy (with a suboptimal but predictable error in the data).

Allocation of Collection Management Functions.   A centralized collection management function permits the optimal utilization of available resources, the ability to avoid mutual interference, and opportunities for combined measurement (e.g., passive ranging). On the other hand, it is generally desirable that each individual sensor platform have the capability to operate autonomously, requiring that it be able to determine collection priorities. This implies some sort of scheme for delegating responsibilities, perhaps based on a hierarchy of prioritized network-wide goals.[76]

These issues typically are addressed in the context of the requirements and constraints of specific implementations. As an example of the way such design issues have been addressed, we again examine the data fusion architecture

developed by Bowman and Snashall[13] for use by internetted fighter aircraft, focusing on the architecture's multiplatform features.

Figure 3.21 depicts the fusion tree employed in each fighter aircraft for processing on-board sensor data, integration of track data received from the network, and outputs to the network. The fusion tree was structured to distribute processing functions as much as possible to (1) reduce communications bandwidth (because the bandwidth limitations of current JTIDS links are a major constraint in this application); (2) permit graceful performance degradation if individual sensors, processors, or platforms drop out; and (3) reduce the complexity and cost of modifications as sensors are upgraded or replaced.

Additionally, the fusion tree was designed to partition the problem for reduced algorithmic complexity. For example, the highest confidence associations are performed first to reduce the search space for subsequent processing.

The processing in each node of the fusion tree follows the general flow shown in Fig. 3.14. After common reference processing for coordinate conversion, alignments, and time propagation, both report-to-track and report-to-report hypothesized associations are generated and evaluated. After association hypothesis selection, track maintenance is performed and ID/track state estimates are generated.

The tree has a node for each sensor to perform its own report-to-track association and ID/track file generation [data from nontracking sensors—e.g., IR or UV missile launch detector (MLD)—are fused with tracks at the multisensor level].

Subsequent fusion nodes perform association among groups of on-board sensors that support particular mission functions (e.g., offensive, defensive, and navigation functions) and that generally involve common sensing and reporting priorities, common targets of interest, and commensurate features. That portion of ground ID/track processing, which is easily partitioned from the air data processing is accomplished independently and concurrently, using a separate ID library and lower order kinematic filtering.

In the fusion tree shown in Fig. 3.21, updates of each aircraft's local MSI track file (LTF) are shared across the data link to generate the internetted global track file (GTF). An aircraft receiving LTF data from other aircraft will associate that data with its on-board LTF to create a local version of a global track file. In turn, it will pass data from its LTF (not its local GTF). However, an output LTF track will be tagged to indicate the off-board LTF tracks with which it has been associated. To reduce the communications and associated processing load, each platform passes only the on-board LTF data which "significantly improve" the LTF data received from other platforms.

An alternative approach would be for each aircraft to transmit GTF updates, doing so whenever its LTF data have "sufficiently improved" the GTF. The benefits of this approach are the use of a common data base and reduced message traffic. Criteria for determining sufficient improvement would have to be defined in terms of the particular mission priorities and communications constraints but would include creations of new tracks, elimination of false or duplicate GTF tracks, "significant" improvements in track kinematic or ID state estimates, and measurements to be used in multiplatform passive ranging (discussed in Sec. 3.7.2). Maurer et al.[99] have developed a graph-theoretic metric for measuring "significant" LTF/GTF differences that address many of these
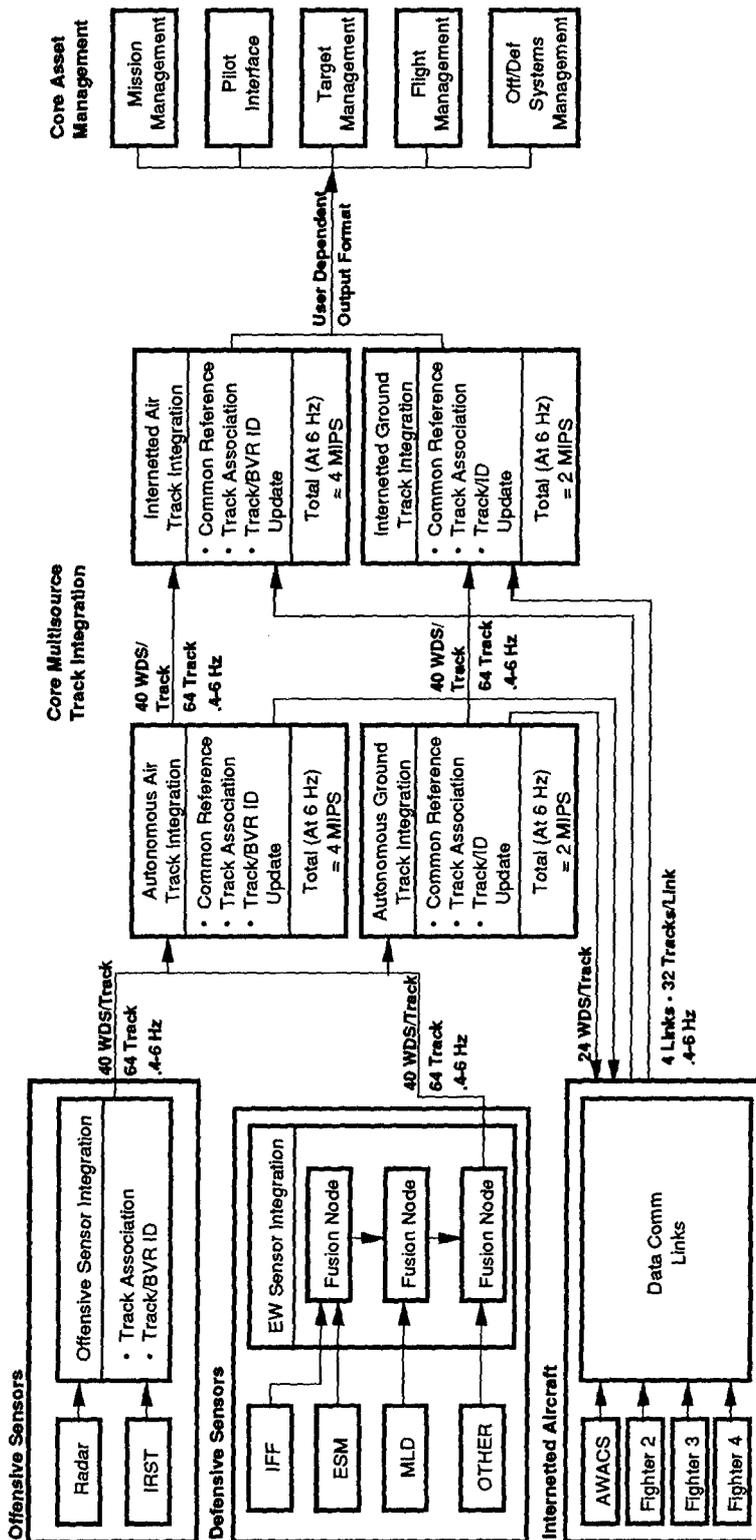
Fig. 3.21 Representative internetted and autonomous fusion architecture.[13]

issues. The overlapping data between the shared GTF and on-board LTF is used to associate corresponding target tracks. After association, the data from each is used to generate improved on-board GTF ID/track estimates. This on-board version of the global track file, being the best surveillance picture available to each aircraft, is available for pilot display and mission management.

Bowman and Snashall note two problems with the GTF-sharing design:

1. Duplicate GTF tracks arise because of the sharing of new detections by many platforms (e.g., many fighters will detect and report an emitter turn-on) within the link delay window. Other discrepancies between locally generated GTF files occur because of misassociation. Thus, the on-board GTF must continually be compared with shared GTF files to recognize discrepancies and report them, if necessary.
2. Correlated errors in similar sensors are reinforced.

The second problem need not be a factor in the LTF-passing alternative, because the individual sensor contributors can be reported. The association and ID/track updates are then solved each time using each platform's local data directly. By sharing significantly improving LTF and their associations, the LTF architecture can approach the efficiency of the GRF-passing architecture and with reduced errors. Message sizes and message rates are somewhat larger in the former design than in the latter. This may be burdensome in specific applications.

Because of the several second latencies expected in JTIDS networks, Bowman and Snashall find that the principal payoff for internetting in this implementation is to provide aircraft with target or measurement data not available via their on-board sensors (e.g., due to jamming, terrain blockage or equipment failure), and not improved tracking or ID accuracy.

3.7.2   Multiplatform Measurements

Several important applications exist in which spatially distributed sensors combine in obtaining measurements of types or with accuracies not attainable by any of the contributing sensors individually.

3.7.2.1   Stereoscopic Imaging.   Earth observation either from aircraft or from satellites has benefited from pixel registration of images taken from different angles. Parallax processing permits accurate measurement of terrain elevation by such means.

Techniques that have been employed in registering *Landsat* and *Spot* satellite imagery for three-dimensional mapping include perspective transformation and polynomial warping, as described in Sec. 3.4.2.

3.7.2.2   Passive Ranging.   Location and track kinematics of point-source targets can be obtained from spatially separated sensors in a number of ways, each of which places particular constraints on sensor design and multisensor processing. Prominent techniques exploit the difference in angle, time, or Doppler frequency as received by spatially separated platforms. In each of these techniques, there is a need to employ processes that will efficiently filter track states as passive data is received over time from separate sensor platforms.

Angle Difference Location.   This technique exploits the angular parallax Θ_2 − Θ_1 of a point source target as viewed from two sensor platforms, much as in stereoscopic imagery. Target range is given by

$$r_1 = \frac{d \sin(\Theta_2)}{\sin(\Theta_2 - \Theta_1)} ,$$ (3.63)

where

Θ_i = target line-of-sight angle from sensor platform i (in a common coordinate system)
d = distance between the sensor platforms
r_1 = target range from sensor platform 1.

Range accuracy is a function of the parallax angle and angular and distance measurement accuracies:

$$\delta R = \frac{\delta d \sin(\Theta_2)}{\sin(\Theta_2 - \Theta_1)} + \frac{d \cos(\Theta_2)\delta\Theta_2 + R\delta[\sin(\Theta_2 - \Theta_1)]}{\sin(\Theta_2 - \Theta_1)} .$$ (3.64)

Passive ranging using sensor platforms that track targets in a single plane is subject to ambiguous track associations. Pairs of rf direction-finding systems (and others) that track only in azimuth can generate "ghost" tracks resulting from misassociation. Figures 3.22 and 3.23 show that, as the true targets move, so do the ghosts. Under many conditions (e.g., with relatively closely spaced targets moving in formation, as in the figure), ghost tracks pass kinematic gates as plausible tracks for the given class of target.

Other single-phenomenology passive location techniques—such as the time-difference or frequency-difference techniques discussed later—are likewise
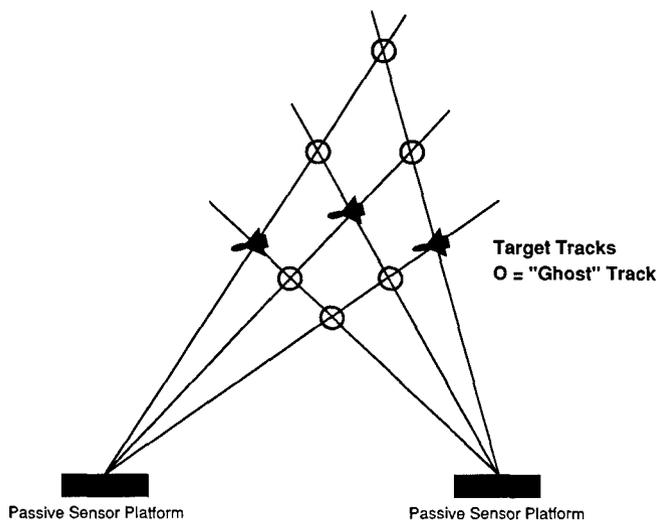
Fig. 3.22   The problem of ghost tracks in passive multiplatform data fusion.
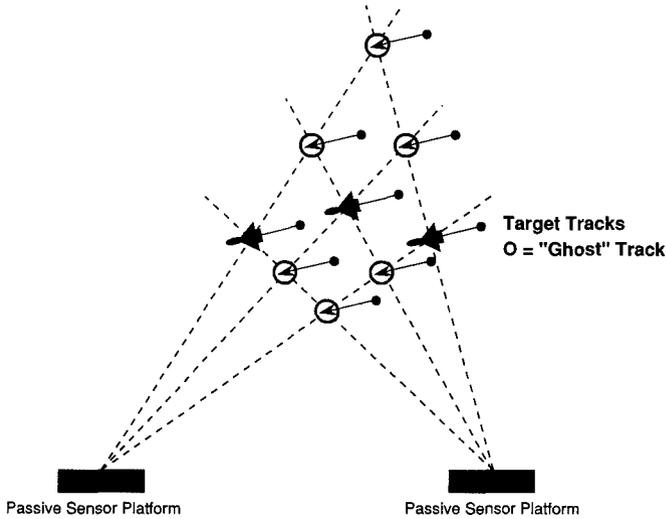
Fig. 3.23 Ghosts can present realistic track kinematics.

ambiguous when only two platforms are employed. In each case, ambiguity resolution can be attained with (1) measurements from a third (out-of-plane) sensor platform, (2) resolving track states using multiple measurements from moving sensor platforms, (3) attribute association; or (4) a hybrid tracking/ location technique including multiple phenomenology.

Time Difference Location.   The location of an emitter can be determined if the difference in propagation time to multiple receiving platforms can be established:

$$\tau = (r_2 - r_1)/c \ , \tag{3.65}$$

where

τ = time difference of arrival (TDOA)
r_1, r_2 = distances from emitter location q to receivers 1 and 2
c = speed of propagation.

For any value of τ, this equation represents a hyperboloid of revolution on which the emitter lies. If the emitter is known to lie on the surface of the earth, then the intersection of this hyperboloid with the surface describes a locus on which the emitter lies.

Thus, at least one additional measurement is needed to estimate the emitter's location. A third receiver location (out of the plane of 1, 2, q) allows resolution of emitter location by the intersection of TDOA surface traces.

As a measure of the contribution of a TDOA measurement to location error, the one-sigma width of the locus of constant TDOA is derived by Chestnut[77] as a function of measurement factors **M** and geometrical factors **G**:

$$\sigma = \mathbf{MG} \ , \tag{3.66}$$

$$\mathbf{M} = (\sigma_{\text{TDOA}}^2 + 2\sigma_p^2/C^2)^{1/2} \ , \tag{3.67}$$

$$\mathbf{G} = c[4 \, \sin^2(\theta/2) - (\cos\phi_2 - = \cos\phi_1)^2]^{-1/2} \ , \tag{3.68}$$

where

σ_τ = one-sigma TDOA uncertainty
σ_p = rms uncertainty in receiver location
θ = angle subtended at the emitter by the two receivers
ϕ_i = angle between the unit vector from the center of the earth in the direction of the emitter and the unit vector from the emitter location to the i'th receiver.

In particular, the ability to measure relative signal time of arrival and, therefore, σ_τ, is affected by such factors as signal modulation bandwidth, receiver bandwidth, time standard accuracy, and emitter/receiver relative motion.

Also, because it is necessary to measure time from the same signal epoch, periodic signal modulation (e.g., constant radar pulse repetition interval) causes ambiguities in the form of concentric surface traces. The technique is inapplicable to unmodulated emissions, for which no time-epoch association is possible.

Frequency Difference Location. A similar multiplatform passive location technique exploits the difference in frequency (usually radio frequency) in a signal received in two or more moving sensor platforms. The technique is also referred to as differential Doppler, integrated Doppler, and TDOA-dot (for time derivative of TDOA).

If a signal of the form $A \, \cos(2\pi f_c t)$ is transmitted, the average rf received by sensor platform 1 over a time interval $T = t_2 - t_1$ is

$$f_1 = f_c - (r_{12} - r_{11})/\lambda T \ , \tag{3.69}$$

where r_{1j} is the distance of receiver 1 from the emitter at time t_j and λ is the transmission wavelength. For two spatially separated receivers, the FDOA is

$$f_d = f_2 - f_1 = (1/\lambda T)(r_{22} - r_{21} - r_{12} + r_{11}) \ . \tag{3.70}$$

Note that, because it is the difference in received rf's that is used in FDOA location, it is not necessary that the transmitted rf be known. However, the technique depends on precise synchronization of the measurements, a high degree of receiver stability, and accurate knowledge of the receivers' relative location and motion.

A single FDOA measurement (i.e., taken at one time by two sensor platforms) generates a hyperboloid of possible locations, but similar to but distinct from that for TDOA. Often the earth surface loci of constant TDOA and frequency difference of arrival (FDOA) will intersect so that emitter geolocation can be resolved by a single coordinated TDOA/FDOA intercept by two sensor platforms, if signal and collection characteristics are appropriate.

FDOA location is most suitable to collections against signals with sufficiently high duty cycle that phase can be tracked over many cycles and compared in the receiving sensors.

The one-sigma width of the FDOA surface locus similarly involves measurement and geometry factors as in Eq. (3.66). Chestnut gives these factors for FDOA as

$$\mathbf{M} = [\sigma_{\mathrm{FDOA}}^2 + 2\sigma^2/\lambda^2 + (\alpha_1^2 + \alpha_2^{\frac{1}{2}})\sigma_p^2/(2\lambda^2)]^{\frac{1}{2}} \ , \tag{3.71}$$

$$\mathbf{G} = \lambda[\alpha_2^2 - 2\alpha_1\alpha_2 \cos\gamma + \alpha_1^2 - (\alpha_2 \cos\beta_2 - \alpha_1 \cos\beta_1)^2]^{-\frac{1}{2}} \ , \tag{3.72}$$

where

α_1 = angle subtended at the emitter between receiver 1 at times t_1 and t_2 (analogously for α_2)

σ_p = receiver positional uncertainty (assumed to be identical for the two receivers)

γ = angle between e_1 and e_2

β_i = angle between r and e_i

e_i = unit vector in the direction of the component of velocity of receiver i as observed from the emitter.

Semiactive Location. This involves the exploitation by a passive sensor of target-reflected energy. Generally target illumination is by a cooperatively participating actively radiating platform; examples include semiactive surface-to-air or air-to-air missile seekers being directed by an active illuminator (often associated with the launch platform, but not always). Also, passive fighter aircraft targeting techniques have been developed using illumination provided by an interlinked AWACS aircraft. There has also been work on the passive opportunistic exploitation of noncooperating illuminators.

Three-space target tracking in such cases implies a knowledge of the illuminator's location (the missile seeker usually relies on angle-only tracking for proportional guidance).

Researchers at Litton-Amecom have reported successful development of a system whereby a moving passive sensor platform infers the location of a noncooperating illuminator, and therefore the target state, by angle tracking of both illuminator and target.[78,79]

3.8 INTEGRATED COLLECTION MANAGEMENT

Considerable enhancement to the effectiveness of multisensor systems is generally achievable by the interactive management of the collection and estimation process; hence, the application of the term *level-four data fusion.*

Adaptive control of the collection process has been shown to afford early target detection, improved tracking performance, and robustness under natural or man-made environmental variations. Additionally, an integrated, flexible control scheme provides a means to manage data flow, thereby managing processing and communications loads on system elements.[80]

Managing the data collection process presumes some scheme for assigning value to different types of information and a method for predicting the likely information value achievable by collection actions possible to the system.

Historically, automated sensor allocation has generally been limited to scheduling resource allocations to provide a sufficient rate for performing re-

quired tasks. In most automated systems, task prioritization is fixed pre-mission or with gross selection capabilities; e.g., allowing a fighter pilot to select the mission type as ground attack, close air support, air superiority, etc., with corresponding prioritization among gross target categories. Sensors and related processing assets are allocated to maintain an update rate sufficient for maintaining continuity on each target track and to ensure the desired probability of detection of new targets in various search windows. Contentions for assets are typically resolved by preemption or blanking based on the static task prioritization.

A generally more effective approach is to perform allocation dynamically, as an integral element of the overall mission management. For example, rather than adhering to a predetermined track update rate, the desirability of updating each track can be regularly computed based on current mission goals and priorities. The costs for obtaining this information is estimated in terms of (1) the demand on sensor and processing assets required to achieve the desired update rate and measurement accuracy and (2) secondary costs (e.g., mission degradations resulting from interference with other sensors or increased susceptibility to detection by hostile systems). Utility and costs are compared with those for tasks that would compete for the same assets or subject to interference. Then, short-term objectives are formed and frequently updated.

This section provides a theoretical framework for an integrated information acquisition/response system and considers several implementations in which such integration is accomplished in varying degrees. This decision-theoretic model for information acquisition (or collection management) is adapted from prior work.[76,81]

3.8.1 Model for Integrated Sensor/Response Management

The utility of information can be evaluated on the basis of its contribution to system mission goals. Key factors in planning and executing any practical mission involve the unavoidable problems of situational uncertainty, contentions for finite system assets, and unexpected side effects of system actions.

Assuming perfect knowledge of current and future world states, a system could define a schedule of actions defined that would be optimal in terms of maximizing a mission objective function Ω^*, given the system's available repertoire of actions.

Unfortunately, real-world systems must generate and maintain action plans based on the error-prone estimates provided by realistic sensors and associated processing and control, together with erroneous, incomplete, and uncertain *a priori* knowledge.

The goal of information acquisition in a system responding to its environment, then, is to provide resolution of that environment sufficient to support response decisions. We follow Moore and Whinston,[82] who model the information acquisition problem as that of achieving a partition among possible world states such that the final partition corresponds to exactly one member of the system's repertoire of responses (i.e., effecting the selection of a specific response action).

The goal of collection management is modeled as that of choosing a sequence of information acquisition actions—i.e., a strategy α and a decision function $\delta : \mathbf{B} \to D$—to maximize the expected net payoff Ω^*:

$$\Omega*(\alpha,\mathbf{B},\delta) = \sum_{B \in \mathbf{B}} \sum_{x \in B} \Phi(x)\omega*[x,\delta(B),C(B)] \; ; \tag{3.73}$$

where

$\mathbf{B}$	$= \{B_1, B_2, \ldots, B_q\}$ = an exhaustive partitioning of the set of possible world states X; $B_i \cap B_j = \phi$ for $i \neq j$ and $\cup_{i=1}^{q} B_i = X$
$\Phi{:}X \to [0,1]$	= probability density function, with probability distribution $\pi{:}2^x \to [0,1]$ on the power set of X, defined by

$$\pi(Y) = \sum_{x \in Y} \Phi(x)$$

for $Y \subseteq X$

D	= set of available response decisions among which the decision function δ must select
$C{:}\mathbf{B} \to \mathbf{R}$	= cost function

$$C(B) = \sum_{t=1}^{r} c[a(t,B)] \; ,$$

where $c[a(t,B)]$ is the cost of the t'th action in the sequence leading to partition B and $\mathbf{R}$ is the set of possible utility results

$\omega*{:}X \cdot D \cdot \mathbf{R} \to \mathbf{R}$	= payoff function.

Each information acquisition action $a \in A$ (e.g., sensor assignment, parameter measurement, signal or track association, signal, or target ID task) yields an information set (or *signal set*) Y_a. In the ideal case of noiseless information, we can represent the information results of such actions by means of a function $n_a{:}X \to Y_a$. Per Shafer,[52] the belief function corresponding to n_a, $h_a{:}X \cdot Y \to [0,1]$, is "discounted" at a rate proportional to the *a priori* reliability of the sensor information. Moore and Whinston[82] argue that the general case with noisy sensor information is reducible to a noise-free model.

An information acquisition action $\mathbf{a}$ induces an information structure $\mathbf{M}_a$ on X:

$$\mathbf{M}_a = \{M | \exists y[M = n_a^{-1}(\{y\})]\} \; . \tag{3.74}$$

A sequence of information acquisition actions creates a sequence of partitions on X, each a refinement on the previous partition in the sequence (i.e., each is at least as fine grained as its predecessor). The refinement of an information structure $\mathbf{B} = \{B_1, B_2, ..., B_k\}$ by an action function $\alpha{:}\mathbf{B} \to A$ is defined as

$$R(\mathbf{B},\alpha) = \{r | \exists M[M \in \mathbf{M}_a \; \& \; \exists j(B_j \in B \; \& \; r = B_j \cap M)]\} \; . \tag{3.75}$$

A *feasible* information gathering strategy σ for D is a sequence of $\mathbf{r}$ partition/ action pairs, which culminate in a partition $\mathbf{B}_{r+1}$ which will support the response decision $\delta{:}\mathbf{B}_{r+1} \to D$:

$$\sigma = \langle (\mathbf{B}_1,\alpha_1), (\mathbf{B}_2,\alpha_2), \ldots, (\mathbf{B}_r,\alpha_r), (\mathbf{B}_{r+1},\delta) \rangle \; , \qquad (3.76)$$

satisfying the following conditions:

$\mathbf{B}_1 = \{X\}$

$\alpha_t{:}\mathbf{B}_t \to A, \, t = 1, \ldots, r$

$\mathbf{B}_{t+1} = R(\mathbf{B}_t,\alpha_t), \, t = 1, \ldots, r$

$\delta{:}\mathbf{B}_{r+1} \to D.$

It can be shown that this formulation of feasible strategy meets Bellman's condition for obtaining an optimal (maximum net payoff) solution via dynamic programming.[81,82] Unfortunately, there is no general effective procedure to work backward to develop an optimal strategy from a feasible set $B{\subseteq}X$ using induction. Therefore, the planner is generally obliged to use heuristics that will be suboptimal in resolving world states in support of a response decision.

If costs of actions leading to a decision are linearly separable from the decision payoff, the net payoff of a decision strategy σ is the difference between the gross payoff and the cost of the actions:

$$\Omega^*(\sigma) = \sum_{B \in B_{r+1}} \sum_{x \in B} \Phi(x)\omega^*[x,\delta(B)] - \sum_{B \in B_{r+1}} \pi(B)C(B) \; , \qquad (3.77)$$

where $C(B)$ is the cost of all actions yielding partition B:

$$C(B) = \sum_{t=1}^{r} c\{\alpha_t[\beta_t(B)]\} \; , \quad t = 1, 2, \ldots, q - 1 \; , \qquad (3.78)$$

where $\beta_t(B)$ is the predecessor partition to B at step t; i.e., $\beta_t(B)$ is that $B' \in \mathbf{B}$ such that $B{\cap}B' \neq \phi$.

The given model can be applied to a particular system by the definition of specific functional relations between the system's repertoire of information acquisition actions and the specific decision factors involved in selection among the system's repertoire of responses. This requires establishing

1. the system's response decision function $\delta{:}\mathbf{B} \to D$
2. the system's payoff function $\omega^*{:}X \cdot D \cdot \mathbf{R} \to \mathbf{R}$ in terms of specific response decisions $d{\in}D$ and world states $x{\in}X$
3. the likelihood of the planner selecting a feasible information acquisition strategy σ in situation x leading to a partition supporting the response decision δ per Eq. (3.76)
4. the costs in terms of Eq. (3.78) of implementing strategy σ in situation x.

3.8.2   Track-Level Cueing

Many applications involve using the data from one sensor to *cue* another. At the track support level, sensor cueing involves preselective steering toward a specific target. A request is made of the cued sensor to supplement existing target reports on a specific track.

Cueing can be invoked to take advantage of specific sensor capabilities. This might include overcoming terrain masking via geographically separated sen-

sors, ensuring tracking continuity during conditions of high target dynamics, or compensating for degradations caused by environmental changes (e.g., cueing an IRST to assume tracking in place of a jammed radar).[80]

Figure 3.24 depicts potential sensor and response cueing sequences applicable to automatic target recognition (ATR) in a fighter aircraft with an advanced multispectral sensor suite. These include hand-offs from WFOV warning sensors to precision tracking sensors and allocating active radiating sensors (such as radar or LADAR) to a track obtained by a passive sensor (e.g., IRST or IIR) to avoid the penalties of excessive radiation.[100]

Cueing can benefit target tracking in such cases as these, in which the cued sensor provides data for additional track states; e.g., the addition of range data to an IRST azimuth/elevation track. Conversely, an IRST could be cued to provide elevation angle tracking data to an existing 2-D search radar azimuth/range track.[80]

Other cueing examples include those illustrated in Fig. 3.2, such as hand-offs from one radar to another in an air traffic control net and hand-offs from early warning to tracking radar to missile seeker in an air defense net.

Figure 3.25 indicates two techniques for handing off tracks from a sensor with relatively large spatial resolution to one with a narrow field of view (or between sensors with large alignment uncertainty). In each case, the cued sensor must search the hand-off *error basket* until the target is acquired. In
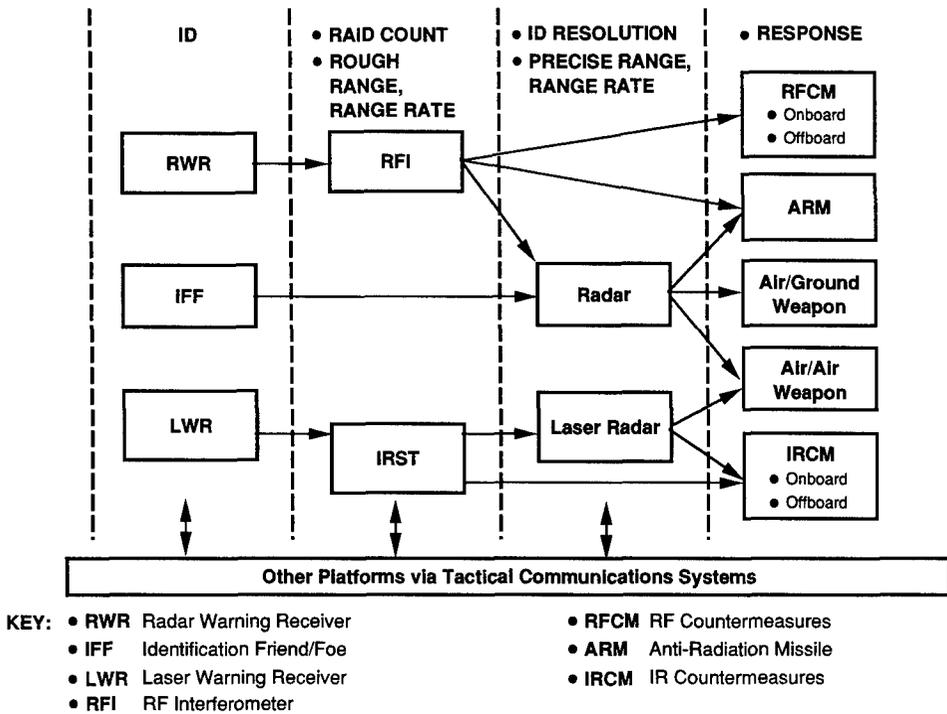
Fig. 3.24 Representative sensor cueing chains in fighter aircraft.
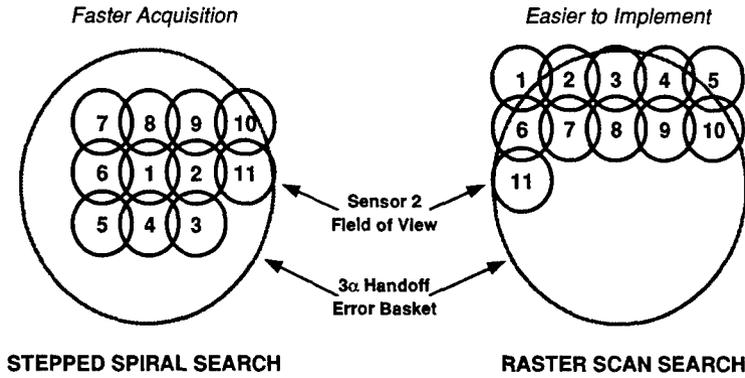
Fig. 3.25   Sensor hand-off techniques.[3]

one case, the cued sensor starts at the center of the basket (i.e., the expected target position) and searches the basket in a stepped outward spiral. Assuming a normal error distribution, such an approach should achieve the fastest acquisition. The raster scan search shown at the right in Fig. 3.25 is suboptimal (in any realistic error distribution) but is simple to implement.

The figure depicts hand-offs involving overlapping fields of view in angle. Analogous techniques are applicable in cueing sequences involving greater dimensionality or in which dimensionality differs between sensors. For example, in an air defense system, a target tracking radar is required to search a three-dimensional hand-off error basket received from an early warning radar. In fact, because of target motion, this basket ideally is a kinematic track state vector (e.g., with six degrees of freedom).

A track-level cueing process might operate as follows.[80] After selection of the target track and the sensor to be cued, the appropriate coordinate transform is calculated, based on the relative geometry and kinematics between cued and cueing sensors (or the MSI track data) and the target. The target state vector and its associated covariance matrix are transformed into a target state vector and covariance in the cued sensor coordinate frame. The target's present position and its predicted future position (allowing for hand-over latency) are compared with the cued sensor's coverage capability to determine if that sensor is capable of detecting the target. If so, the cued sensor is commanded to search for and acquire the target.

On target acquisition by the cued sensor, it may be necessary (in dense target or clutter environments) to verify that the true target is being tracked, employing hypothesis evaluation techniques as described. On such verification, the cued sensor's target measurement data is provided to the MSI data association and tracking algorithms.

3.8.3   Detection-Level Cueing

In more sophisticated systems, sensor cueing may involve not only sensor pointing but lowering of detection thresholds and parametric acceptance gates matched to the expected target characteristics. For example, in cueing between

radar systems, range and Doppler gates in the recipient sensor may be set to correspond to the expected target position and kinematics.

Directed search using predetection filters matched to the indicated target signature can provide optimal detection and discrimination, if the required prior data is available.

Detection at the pixel-level has been enhanced by the use of *interest images*.[35,36] Pixel-registered sensor images are transformed into interest images, having high pixel values at those locations containing some feature known to be associated with the target of interest.

This is a simplification of more general implementations, not dependent on cueing. These techniques involve constructing composite interest maps by pixel-wise combination of interest images for multiple targets of interest. Effectively, such a map is a more or less matched spatial filter for the indicated target set. Such techniques have been reported to have significant benefit in target extraction from noisy backgrounds.[35,36]

3.8.4 Sensor/Response Integration

The fully integrated sensor/response system is one that integrates estimation and control functions to create a full spectrum awareness and mission adaptation to the observable environment. Full integration of the collection management process into the overall mission management takes full advantage of each sensor's uniqueness, spectral diversity, and flexibility to respond to ever-changing conditions. Sensor integration at this level requires specific, detailed control over each sensor to react to variations in the environment and in mission priorities.[80]

As a relatively straightforward example, consider the information collection and exploitation problems that a fighter aircraft's self-protection system must solve in conducting its mission. It must

1. plan assignment of sensors, countermeasures, and evasive/avoidance maneuvers, given the uncertainties in threat assessments and in response effectiveness
2. schedule actions within time constraints for measurement, countermeasure, and processing
3. coordinate use of sensors to reduce uncertainties as required for response planning and cueing
4. resolve conflicting demands on assets so as to maximize survivability against the ensemble of threats to be encountered in conducting the assigned mission (i.e., allocating assets to ensure survival against one threat should not unduly jeopardize the chance of survival against other threats encountered concurrently or later in the mission or otherwise inhibit the likelihood of meeting global mission objectives).

A design for an aircraft self-defense mission planner (henceforth called PLANNER) to address these requirements is presented in Ref. 81. This section summarizes the concept for integrating collection management with mission management and response functions in this system.

The PLANNER system is designed to operate in a simulated fighter aircraft multisensor environment.

The PLANNER response management function (1) develops candidate responses to reported world states; (2) estimates the effects of candidate actions on survival and on mission objectives; (3) identifies conflicts for resource uses or detrimental side effects of candidate actions; and (4) resolves conflicts to assemble composite asset assignment plans, based on the estimated net impact on mission attainment.[83]

An information acquisition function supports response management, performing the level 1 to 3 fusion functions of (1) estimating threat identity, lethal potential, and intent on the basis of available sensor and intelligence data; (2) predicting the time to critical events in threat engagements (e.g., target acquisition, tracking, weapon launch, impact); (3) providing parametric threat data (e.g., threat radar parameters) for exploitation by specific countermeasure and weapons targeting techniques; and (4) monitoring the effectiveness of these responses to support subsequent response control and planning.

Collection management—the allocation of sensors and information acquisition processes (including level 1 to 3 fusion processes)—is an integral element of response management, as described in the following.

Measures of effectiveness for defensive weapons and electronic warfare countermeasures are readily defined in terms of their impact on the platform's probability of survival and mission objectives.[84]

In contrast, the utility of sensors and information acquisition in self-defensive systems can be related only indirectly to survivability and mission effectiveness factors. What counts is the ability to recognize a particular threat situation with sufficient accuracy to be able to select an appropriate response from the system's repertoire and to do so in enough time for the response to have a high likelihood of success.

Accordingly, it is the role of information acquisition in the self-defense application to (1) recognize particular world states that may lead to lethal states; (2) estimate the conditional likelihood of transitions to other particular states, given candidate action sequences; and (3) estimate the time of such state transition.

Cost factors involved in information acquisition in an aircraft self-protection system include

1. increased exposure to threats caused by flight path changes in support of information acquisition; e.g., pop-up maneuvers for threat/target detection or for azimuth/elevation passive location or delaying response decision until a sufficient base line has been flown for triangulation location of a given accuracy
2. use of actively radiating information sources (e.g., radar, ladar, data link), incurring an increased likelihood of acquisition, tracking, and lock-on by threats
3. reduced countermeasures effectiveness because of reduced duty factor, power level, or spectral coverage to accommodate sensor look-through
4. reduced ability to assess the threat environment because preemption of sensors by cueing or directed search, which may adversely affect the above three factors.

The PLANNER implementation to address these requirements features (1) a hierarchical, object-oriented architecture, which permits simultaneous rea-

soning at multiple levels of depth, and (2) an activation frame object paradigm, which facilitates reasoning within specific time constraints and focusing of the planning process as new external events force revision of priorities.

The PLANNER architecture distributes the problem solution into multiple parallel processes, then uses message passing among the processes to accumulate incremental evidence for or against each plausible solution. At any given time, therefore, PLANNER provides the best plan of action for achieving current goals, consistent with the available situational and procedural knowledge and the available planning and reaction times.

A mission plan is represented as a hierarchical data structure as shown in Fig. 3.26. Mission goals and constraints flow down from the mission level to allocations over successively finer mission phases. At each level, a mission phase constitutes a grouping of situations (e.g., threat encounters in the design application) for which responses are coordinated.

This procedure is recursive: subordinate planning nodes are activated to develop more detailed candidate plan segments when a higher level node determines that more detailed planning is both feasible (in terms of estimated response time and resource cost) and beneficial (in terms of improved likelihood of attaining the assigned goal).

In response to the requirements/constraints flowdown, there is an upward flow, consisting of candidate plan segments; i.e., proposed schedules of asset allocations and controls over time plus estimates of the likelihood that proposed plan segments will achieve their assigned goals. A higher level planning node

Fig. 3.26   Hierarchical mission plan example.

evaluates cost versus benefit (in terms of its own higher level goal) of the received proposals and may repartition its higher goal into a new set of subgoals if the initial allocations are not achievable. For example, if a required probability of aircraft survival against a particular threat grouping cannot be met, a new threat grouping must be developed.

Planning is a process analogous to data association (discussed in Sec. 3.5.1). Functions corresponding to hypothesis generation, evaluation, and selection are involved: (1) plan generation involves searching over a number of possible actions for assembly into candidate plan segments, which are passed to (2) plan evaluation and (3) plan selection.

As with association hypothesis generation, plan generation involves potentially massive searches, which must be constrained in practical systems. Ultimately, this could involve evaluating over all possible actions by all system assets against all defined goals over all future times (Table 3.6). Analogous to the selection of fusion trees to constrain the search for association hypotheses, a planning tree can be selected that searches only over assets and goals and implementation times that have high *a priori* payoff (Ω^*).

Figure 3.27 shows the structure of planning nodes in PLANNER. Candidate plans—i.e., schedules of tasks for system assets—are assembled by recursion on the PLAN procedure. The level of planning is adapted on the basis of (1) the assessed utility relative to current mission goals of the given plan segment as currently developed and (2) the time available for further planning.

By the first basis, near-term plan segments tend to be constructed in greater detail than are far-term ones (for which the expenditure in planning resources may outweigh the confidence that the plan will still be appropriate to the circumstances extant at the time for implementation).

Templates for constructing default task sequences are defined *a priori* appropriate to each generic type of goal defined for the system (and, therefore, for most far-term plan segments). If a given default plan is estimated not to be cost-effective and if there is time available for deeper planning, then planning proceeds at the next lower level.

Table 3.6   Design Alternatives for Data Fusion and Planning Trees

```
        Fusion Tree Alternatives

        Selected Sensors . . . . .All Sensors
        Selected Reports . . . . .All Reports
        Only Current Times . . . .All (past)
                                       Times

        Planning Tree Alternatives

        Selected Assets. . . . . .All Assets
        Selected Goals . . . . . .All Goals
        Only Current Times . . . .All (future)
                                       Times

     Min Performance,        Max Performance,
     Complexity & Cost       Complexity & Cost
     ◄─────────────────────────────────────►
```

Fig. 3.27   Planning node structure.

In this way, an adaptive planning hierarchy is effected recursively, in which each node is an instantiation of the PLAN procedure.

Deeper planning is accomplished by recursively partitioning a goal into candidate sets of subgoals (the PARTITION procedure) and combining them into a composite higher level plan (MERGE). At each level, candidate plans are evaluated as to their effectiveness in achieving assigned goals, the global value of each respective goal, and the cost of implementing each candidate plan (EVAL). By evaluating these cost/payoff factors to global mission utility, the need for deeper planning or for selection of alternate candidate plans is determined.

Contentions for assigning available assets are resolved by prioritized re-scheduling on the basis of time sensitivity and predicted utility of contending allocations.

As each planning node presents a candidate plan segment for higher level evaluation, it sets flags indicating potential side effects of the plan; e.g., de-tectable signature changes, flight path changes, or emissions in band of another sensor. Higher level nodes respond to such flags by estimating the impact of such effects on their respective higher level goals. In this way, plans responsive to global mission goals are assembled in a hierarchical fashion.

As argued, situational uncertainty is of interest to a planning system only as it affects planning uncertainty; an ambiguity between two situations that would elicit the same response is not worth resolving. When situational un-certainty does affect response selection, PLANNER employs the principle of least-commitment planning: attempting to maintain a plan consistent with competing hypotheses while pursuing information that would resolve the ambiguity.[6,7]

Collection management is an integral part of this response-planning process. Sensor allocation rules are used to determine whether the current situational information is adequate to support effective response selection. In addition, sensor allocation rules determine which sensor and processing controls are appropriate to obtain necessary information to resolve situational ambiguities within timing and resource allocation constraints.[6,60]

In general, collection management functions allocate sensors and processing resources to acquire information for one of the following tasks:

1. Search for information that might affect plan segmentation (i.e., prediction-based prioritized search).
2. Resolve selection among candidate plan segments (e.g., resolving threat types as necessary for countermeasure selection and resolving threat locations/tracks as necessary to schedule countermeasure action).
3. Monitor plan execution to determine the need for possible revision (e.g., scheduling "end-game" countermeasures if it is seen that the existing countermeasure plan is not succeeding in misdirecting an attacking missile).
4. Monitor plan execution for control (e.g., managing phenomenological countermeasures).

For example, if it cannot be determined whether an enemy aircraft is attacking with rf or IR missiles (requiring different countermeasure responses), PLANNER assesses candidate sensor techniques for a means to resolve the ambiguity within time and utility constraints (perhaps by monitoring the missiles' response to countermeasures of one type or the other). Meanwhile, the PLANNER attempts to maintain tactics consistent with either hypothesis. Once a decision has to be made between conflicting action plans, a minimum penalty solution is used.

3.8.5   Sensor Interoperability

An important aspect of sensor management involves coordinating the use of multiple sensors and other assets so as to avoid mutual interference.

Traditional techniques for ensuring the interoperability of sensors in close proximity—as on a single ship or aircraft—include various forms of spatial, spectral, and temporal allocation. As assets contend for the use of the electromagnetic spectrum, operations research, optimal control, and knowledge-based techniques have been employed to implement integrated, adaptive energy management.

By way of illustration, let us examine the problems of dynamic spectral management in the context of a version of the U.S. Navy EA-6B tactical jamming aircraft depicted in Fig. 3.28.

A system such as the EA-6B may employ a variety of techniques to coordinate the required access to the rf spectrum of its diverse transmitting and receiving assets. These various "look" techniques involve the spatial, spectral, and temporal allocation of transmitting and receiving assets.

Spatial Allocation—"look around."   Spatial allocation involves locating receiving apertures on a platform so as to be outside the beam and antenna scan patterns of radiating apertures. This is the only practical implementation for sensor systems that must operate uninterruptedly over the bands used by onboard emitters.

For example, in the version of the EA-6B aircraft shown, transmitting antennas are located outboard from in-band receiving antennas, affording as much blockage as possible by aircraft structure. Thus, the receiving antennas for the AN/ALQ-99 tactical jamming system are located in the tailcap (1 to 8,
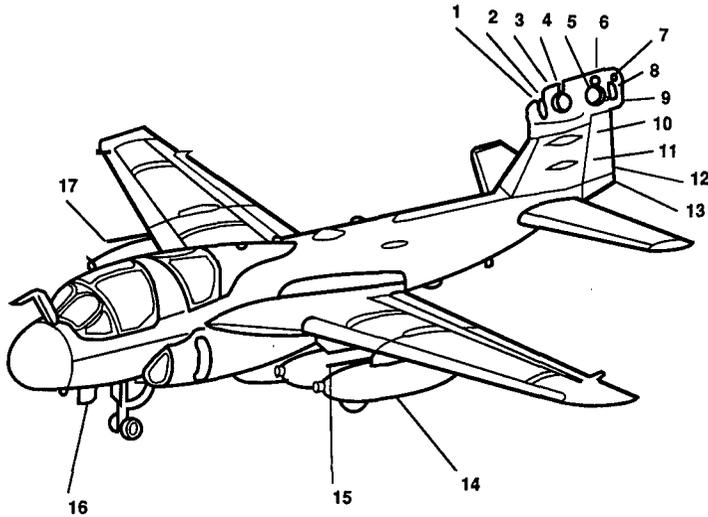
Fig. 3.28   The rf antenna locations in the U.S. Navy EA-6B aircraft.

10, and 11 in the figure) for maximum isolation from the in-band jammer emitter antennas, which point forward and aft from each of five pods (14).

The association transmitter antennas (9 and 17) and receiving antennas (12, 13, and 15) of the AN/ALQ-126 radar defense jamming system [planned to be replaced by the AN/ALQ-165 airborne self-protection jammer (ASPJ)] are similarly spatially isolated. Ventral blade antennas (16) are time shared for transmit and receive functions of the AN/ALQ-92 communications jammer (16) (now replaced by the AN/ALQ-149). Not labeled are rf antennas for the nose radar and (in later versions) JTIDS and GPS.

Increasingly, system engineers are recognizing the importance of designing sensor systems and their apertures integral with the overall system. Among considerations in aperture locations are (1) coordination between antenna patterns to reduce unwanted illumination, while allowing desired common fields of view (e.g., for jammer cueing or effectiveness monitoring); (2) element spacing [e.g., for interferometric phasing either for measuring signal direction of arrival (DOA) or for "cross-eye" countermeasures, or for time difference of arrival DOA measurement]; and (3) mechanical interference considerations.

Spectral Allocation—"look between."   Spectral allocation involves employing a frequency plan to prevent interference by in-band emission, harmonics, or cross-products.

Dynamic Range Allocation—"look above."   In some cases, a receiver operates on signals that are well above the peak power of on-board radiators at the receiver aperture, therefore requiring no particular controls other than appropriate thresholding and attenuation.

Temporal Allocation—"look through."   In other cases in which receiver assets must share a portion of the spectrum with transmitters, it is necessary to time-share the use of each. Blanking has traditionally been a brute-force process:

either the transmitter or the receiver is given precedence and its use always interrupts that of the other. More balanced resource schedulers have been developed in several applications, as in the ASPJ and EA-6B advanced capability (ADVCAP) systems, and in the AEGIS shipboard emissions control system.

Temporal allocation of diverse assets should, of course, be performed in ways that respect the time constraints imposed by each assets function. Some sensing or response techniques require that they be performed in more or less strict relationship to external events. Examples are weapon control and repeater jamming.

Some techniques may be most effective if synchronized with other system internal events (e.g., coordinating receiver dwells with jammer blanking). Other techniques have no synchronicity constraints, but require certain update rates. For example, an aircraft IRST or ESM system estimating target locations by processing successive bearings.

Yet other techniques have constraints on dwell lengths (e.g., to effect increased probability of detection by noise averaging over long dwells in a given channel/bearing). Finally, there are techniques that are temporally constrained only in terms of duty factors (e.g., to achieve an effective range resolution, a spread spectrum radar needs to maintain a sufficiently high time/bandwidth product).

Feldman[85] reports a marked improvement in both rf jammer and receiver performance by use of a rapid time sharing, or *chopping,* between the two systems. The system involved is the F-15 Tactical Electronic Warfare Suite (TEWS), which includes the AN/ALQ-135 Electronics Countermeasure (ECM) and AN/ALR-56C Radar Warning Receiver (RWR).

Such a coordinated approach, of course, can be most effective with low duty factor jamming techniques, such as various pulse repeater and pull-off techniques. In such cases, the low duty factor of the target radar signal and the propagation delays between the on-board jammer and receiver antennas can be exploited to allow the RWR a generous access to the signal environment in the jamming band, both for monitoring jammer effectiveness and for new signal acquisition and analysis.

3.9 PERFORMANCE MEASURES

3.9.1 Measures of Effectiveness

This chapter has given some indication of the great diversity of approaches being developed and applied relating to multisensor fusion. Unfortunately there has not to date been comparable effort given to developing generic, objective measures for assessing and comparing such techniques.[86]

Llinas[6] has undertaken to define concepts for comparing fusion systems' performance or their contribution to mission success. Suggested measures of performance (MOPs) with wide applicability to fusion systems are

1. detection probability
2. mean time to detect
3. false alarm rate
4. ID probabilities

5. ID range (i.e., the distance at which the sensing system's probability of correctly identifying a target exceeds a given threshold)
6. location estimation accuracy.

Measures of effectiveness (MOEs) quantify the contribution of a fusion system in the successful execution of an operational mission. MOEs for fusion systems may include

1. target nomination rate (i.e., the rate at which the system identifies/ assesses targets sufficiently for the system to respond with appropriate action)
2. timeliness of information (to support response decisions)
3. target leakage (percentage of objects or targets against which no adequate response is taken)
4. robustness (immunity to performance or effectiveness degradation in adverse environments; e.g., with noisy, ambiguous, missing or incomplete data, or with high data rates).

Identification performance in a sensor or in a fusion node can be evaluated in information-theoretic terms. Each node receives a set of noisy measurements, which it converts into a classification feature. For each target type there is an associated probability distribution for this feature.

The identification performance of the node can be modeled as the improvement from prior beliefs to posterior beliefs concerning the correct target ID. That is, the node effects a decrease in the uncertainty about the true target ID and this change in uncertainty is a measure of the node's identification performance. When the target by target uncertainty (called *local* uncertainty by Holmes[87]) is averaged over the entire target distribution, a global measure of system performance is obtained relative to a specified target set (universe of discourse). The pertinent performance measures, as given by Holmes, are as follows.

Entropic Confidence Measure.   Assume that the probability of various target classes are known, there being M target classes in all, each target class comprising a fraction π_i of the total target set. Then the uncertainty about the correct classification of a target prior to any observation can be quantified by the entropy H_0 of the values $\{\pi_i\}$:

$$H_0 = E_M(\pi_i, \ldots, \pi_M) = \sum_{i-1}^{M} \pi_i \log\left(\frac{1}{\pi_i}\right) . \tag{3.79}$$

An analogous measure H_1 of ID uncertainty is obtained after sensor observation. The degree of improvement, or entropic confidence measure, is

$$M_E = H_0 - H_1 . \tag{3.80}$$

This *entropic confidence measure M_E* is a natural analytic means for quantifying this residual uncertainty, being inversely proportional to the target uncertainty after the measurement.

Local Information Measure.   A second measure of information from measurements acknowledges the contribution of specific features in the measure-

ment. The sets of prior target probabilities $\{\pi_i\}$ are compared with posterior target probabilities $\{\rho_i(x)\}$, available after a target feature x is observed. For each feature x and target class i, $\rho_i(x)$ is proportional to the product of π_i and the sensor response density $p_i(x)$, describing the distribution of observations over that class. The effect of the sensor operation yielding the target observation x is to alter the probability distribution from $\{\pi_i\}$ to $\{\rho_i(x)\}$. The desired measure of the information content of this probability exchange is the cross-entropy between the prior and posterior probabilities:

$$M_L(x) = \sum_{i=1}^{M} \rho_i(x) \log\left[\frac{\rho_i(x)}{\pi_i}\right] , \qquad (3.81)$$

where M_L is termed the *local information measure*. Large values in M_L indicate a considerable change of probabilities and hence a high degree of information transfer.

Global Information Measure.   Both M_E and M_L can be generalized over the total target distribution of feature space to obtain the expectations (mean values) of observations. These can be given as respective *global information measures*:

$$M_{GE} = \int_{\Omega} M_E(x) \sum_i \pi_i \rho_i(x) \, dx , \qquad (3.82)$$

$$M_{GL} = \int_{\Omega} M_L(x) \sum_i \pi_i \rho_i(x) \, dx , \qquad (3.83)$$

where $p = \Sigma \pi_i \rho_i$ is the unconditional target density over the feature space.

It can be shown,[87] that $G_{ME} = G_{ML}$, and that this common value is exactly the classical Shannon expression for the average mutual information between the target distribution given by the density p and the prenode state of knowledge as determined by prior probabilities $\{\pi_i\}$. This value (which can be simply labeled M_G) is therefore an intrinsic numerical attribute of any generic classification system that adjusts prior information about target type on the basis of received information. The value M_G measures the average information transfer of the sensor of fusion node and is entirely defined in terms of the prior probabilities and node response densities.

Evaluation of measures of effectiveness for sensor fusion systems can employ the model presented in Sec. 3.8.1, in which the utility of an information acquisition process is defined in terms of the contribution to resolving action decisions in pursuit of system goals.

3.9.2 Fusion Testbeds

There have been several developments in recent years of facilities that provide environments for testing data fusion systems. Many of these have been developed or sponsored by the U.S. Department of Defense to support the developmental or operational testing and evaluation of specific systems. Table 3.7 lists some of these current or planned testbeds.

Table 3.7   Representative Multisensor Data Fusion Testbeds (adapted from Ref. 13)

Testbed	General Purpose	Characteristics
Multisensor, multitarget data fusion testbed (U.S.A.F. Rome Lab) Ref. 88	Compare the performance of various level 1 fusion algorithms (association, tracking, ID)	Three major components: scenario generator, platform simulator, data analysis and display; Monte Carlo capability
Advanced sensor exploitation testbed (ASET) (U.S.A.F. Rome Lab) Ref. 89	Large testbed to study level 2 fusion algorithms primarily for order of battle estimation	Scenario generation, sensor simulation, C^3I simulation, fusion, timing and control, evaluation and control, evaluation; man-in-loop; parametric flexibility
National test bed (SDIO) Ref. 90	Evaluate candidate fusion concepts for SDI level 1 tracking and discrimination functions. Study effects of interelement communication links	Large-scale, multi-Cray (plus other large computers) type environment to simulate broad range of functions in SDI problem; purposeful man-in-the-loop capability
Surveillance test bed (U.S. Army Strategic Defense Command) Ref. 91	High fidelity background and sensor (radar, optical) simulator for a broad range of SDI level 1 fusion algorithm testing	Algorithm-level test and replacement design; "framework" and "driver" concept that separates simulation and analysis activities
NATO data fusion demonstrator (NATO) Ref. 92	Initial configuration is to study levels 1 to 3 fusion process for AirLand Battle (initially Army) applications	"Client-server" design concept to permit algorithm-level test and replacement
All-source systems evaluation testbed (U. VA) Ref. 93	To evaluate data associations and correlation algorithms for ASAS-type applications	Algorithm-level test and replacement design; standardization metrics; connected to army command and control (C^2) software
AWACS fusion evaluation testbed (FET) (MITRE) Ref. 94	Provide analysis and evaluation support of multisensor integration (MSI) algorithms for AWACS level 1 applications	Algorithm-level test and replacement design; permits live and simulated data as drivers; standardized MOEs part of air defense C^2
MultiSIM (U.S. Army Labcom) Ref. 95	Analysis of multisensor systems with real-time data fusion. Evaluation of detection, track, ID and cueing	Engineering models of most military sensors; TRADOC-approved scenario terrain models; expert system base
Sensor data fusion analyst test bed (Boeing) Ref. 96	Development and evaluation environment; simulate, track, ID, display/analysis	Distributed data base; prototyping system configuration; open architecture
Modular fusion test bed— MOFUT (George Mason University) Ref. 97	Compare different multisensor fusion approaches in common environment. Develop sensor fusion and management algorithm	Scenario generation; data fusion levels 1 to 4; performance evaluation; analyst interface

Llinas[6] notes that many current testbeds are designed more for experimentation at the parametric level than at the algorithm level. That is to say, although there are means for varying the test conditions (and particularly the simulated operational scenarios), they do not necessarily allow easy replacement and comparison of alternative algorithms or fusion architectures.

On the other hand, a relatively high degree of flexibility is incorporated in several of the listed systems.

The first listed system is a multispectral, multisensor testbed established at the U.S. Air Force Rome Laboratory for development and testing of fusion processing and control concepts. This system is interfaced to a sensor suite consisting of two geographically separated North American Air Defense (NORAD) surveillance radars, along with a dedicated suite of collocated rf and IR sensors. These include an L-band search radar, an S-band monopulse radar and a C-band phased array radar, an IR staring sensor, an ESM sensor, and a passive bistatic rf sensor.

The Rome Laboratory fusion testbed has the flexibility to develop and test a diversity of system implementations to include evaluations in various sensor suite mixes, fusion algorithm designs, higher level fusion processing (e.g., situation assessment and event prediction) and C^3 architectures.

The surveillance testbed has been developed for high detail simulation of sensor, processing, and control systems for application as part of the Strategic Defense Initiative architecture. The testbed includes emulation or simulation of a diversity of scanning and staring IR, visible band, and UV sensors, plus active radar and ladar. These provide measurement-level reports based on detailed, dynamic threat, signatures, and environment models. Test articles to be evaluated include algorithms for bulk filtering (i.e., target hypothesis generation), tracking, discrimination (ID), and sensor tasking. The latter will include both single sensor and multisensor processes.

3.9.3   Performance Potential

This chapter has summarized techniques and design factors involved in the coordinated use of multiple sensors. In particular, the applicability and implementation of various methods for the combination of data from multiple sensor sources (sensor data fusion) have been addressed.

Some of the specific benefits achievable through sensor fusion are

1. Improved target detection, recognition, and tracking. In many applications, a variety of measurement techniques involving several types of sensors must be employed to resolve objects, events and situations with sufficient confidence to support decision making.
2. Reduced cost, size, and weight of multisensor systems, resulting from shared resources, and possibly reduced performance requirements for individual sensors.
3. Reduced workload for human operators, due to the automation of many information integration and assessment functions.
4. Improved operational availability achievable by the use of sensors with complementary performance under conditions in which the performance of individual sensors is diminished; e.g., by precipitation or low ambient light.

5. Reduced mission degradation from the failure of individual sensor assets.
6. Accomplishment of mission goals when the use of particular preferred sensors is restricted by mission considerations; e.g., limiting the use of active emissions to avoid detection or attack, expanding the role of passive sensors to include those that otherwise would be performed by active sensors.

Features 4 to 6 are obtained from the fact that the integration of data and control allows multisensor systems to be *opportunistic:* adapting so as to make maximal use of information as available, often able to operate adequately, although with degraded performance, with a reduced set of information sources.

References

1. Franklin E. White, Jr., *Data Fusion Lexicon,* prepared for Joint Directors of Laboratories, Technical Panel for C^3, Data Fusion Sub-Panel, Naval Ocean Systems Center, San Diego (1987).

2. Otto Kessler et al., *Functional Description of the Data Fusion Process,* prepared for the Office of Naval Technology Data Fusion Development Strategy (Nov. 1991).

3. Dennis Richman, "Multisensor data fusion and multisensor techniques," Course material, Center for Professional Development at the University of Maryland University College (Sep. 1989).

4. Franklin E. White, Jr., "A model for data fusion," in *Proceedings of the First National Symposium on Sensor Fusion,* Vol. 2, pp. 143–187 (1988).

5. *Department of Defense Critical Technology Plan for the Committee on Armed Services,* United States Congress, AD-A234 900 (1 May 1991).

6. J. Llinas, "Assessing the performance of multisensor fusion systems," in *Sensor Fusion, Proceedings of the SPIE* **1611** (1991).

7. Edward Waltz and James Llinas, *Multisensor Data Fusion,* Artech House, Boston (1990).

8. Y. Bar-Shalom and T. E. Fortman, *Tracking and Data Association,* Academic Press, San Diego (1988).

9. I. R. Goodman, "A general theory for the fusion of data," in *Proceedings of the Tri-Service Data Fusion Symposium,* Vol. 1, pp. 252–275 (1987).

10. I. R. Goodman, "Application of a conditional event algebra to data fusion," in *Proceedings of the Tri-Service Data Fusion Symposium,* Vol. 1, pp. 179–188 (1988).

11. I. R. Goodman and H. T. Nguyen, *Uncertainty Models for Knowledge-Based Systems,* North-Holland Co., Amsterdam (1985).

12. I. R. Goodman, "A probabilistic/possibilistic approach to modeling C^3 systems," in *Proceedings of the 9th MIT/ONR Workshop on C^3 Systems,* pp. 161–166 (1986).

13. Christopher L. Bowman and Gleason Snashall, "Internetted fighter sensor data association and state estimation approach comparison," in *Proceedings of the Second National Symposium on Sensor Fusion,* Vol. 2, pp. 297–320 (1989).

14. Robert C. Harney, "Practical issues in multisensor target recognition," in *Proceedings of the Third National Symposium on Sensor Fusion,* pp. 105–114 (1990).

15. A. Omar Aboutalib, "Multisensor data fusion and multisensor techniques," Course material, Center for Professional Development at the University of Maryland University College (Sep. 1989).

16. Guner S. Robinson, "ATR systems: a Northrop perspective," *Proceedings of the SPIE* **750,** 216–224 (1987).

17. David L. Hall, *Mathematical Techniques in Multisensor Data Fusion,* Artech House, Norwood, MA (1992).

18. H. L. Van Trees, *Detection, Estimation and Modulation Theory,* John Wiley and Sons, New York (1968).

19. A. Omar Aboutalib, "Multisensor fusion for target detection, tracking and recognition," in *Proceedings of the Fourth National Symposium on Sensor Fusion,* Vol. 1, pp. 99–118 (1991).

20. Guner S. Robinson and A. Omar Aboutalib, "Trade-off analysis of multisensor fusion levels," in *Proceedings of the Second National Symposium on Sensor Fusion,* Vol. 2, pp. 21–34 (1989).

21. A.Omar Aboutalib and Thanh Luu, "An efficient target extraction technique for laser radar imagery," in *Aerospace Sensing, Proceedings of the SPIE* **1096,** 23–33 (1989).

22. A. Omar Aboutalib, L. Tran and C.-Y. Hu, "Fusion of passive imaging sensors for target acquisition," in *Proceedings of the Fifth National Symposium on Sensor Fusion,* Vol. 1, pp. 151–162 (1992).

23. D. M. Buede and E. L. Waltz, "Benefits of soft sensors and probabilistic fusion," in *Signal and Data Processing of Small Targets, Proceedings of the SPIE* **1096,** 309–320 (1989).

24. Jeff Brandstadt, "Applied probabilistic data association," in *Proceedings of the Tri-Service Data Fusion Symposium,* Vol. 1, pp. 220–230 (1988).

25. Christopher L. Bowman and Ronald K. Munzer, *Integrated Fire Control Surveillance, Final Technical Report,* Vol. 1, AFWAL-TR-80-1151, VERAC, Inc., San Diego (Oct. 1980).

26. Oliver E. Drummond and Samuel S. Blackman, "Multiple sensor, multiple target tracking

challenges of the Strategic Defense Initiative," in *Proceedings of the First National Symposium on Sensor Fusion,* Vol. 2, pp. 503–514 (1988).

27. Julie A. Rosen and Peter S. Schoenfeld, "Parallel processing applications to multi-source correlation and tracking," in *Proceedings of the Tri-Service Data Fusion Symposium,* Vol. 1, pp. 202–214 (1987).

28. Christopher L. Bowman, "Artificial neural network adaptive systems applied to multi-sensor ID," in *Proceedings of the Tri-Service Fusion Symposium,* Vol. 1, pp. 161–172 (1988).

29. Wendy Amai, "A weighting/threshold approach to sensor fusion," in *Proceedings of the First National Symposium on Sensor Fusion,* Vol. 2, pp. 455–468 (1988).

30. Christopher L. Bowman, "Multisensor track file correlation using attributes and kinematics," in *Proceedings of the National Aerospace and Electronics Conference,* pp. 1181–1186 (1980).

31. D. Becker and A. M. Haimovich, "Sensor fusion for the F-14D," in *Proceedings of the Fourth National Symposium on Sensor Fusion,* Vol. 2, pp. 7–24 (1991).

32. *GPALS Navigation Standards,* SDIO report SDI-P-SD-92-000002 (July 1992).

33. Carl W. Clawson, "On the choice of coordinate systems for fusion of passive tracking data," in *Proceedings of the Tri-Service Data Fusion Symposium,* pp. 247–262 (1990).

34. Jeff Brandstadt and Lisa A. Maston, "Correlation of dissimilar sensor reports—an application of multiple hypotheses to the attack helicopter mission," in *Proceedings of the Fifth National Symposium on Sensor Fusion,* Vol. 1, pp. 241–260 (1992).

35. R. Delanoy, J. Verly, and D. Dudgeon, "Pixel-level fusion using "interest" images," in *Proceedings of the Fourth National Symposium on Sensor Fusion,* Vol. 1, pp. 29–49 (1991).

36. Daryl T. Lawton and Michael Callahan, "A simple tunable interest operator and some applications," in *Applications of Artificial Intelligence II, Proceedings of the SPIE* **548,** 186–193 (1985).

37. Christopher L. Bowman, "Max likelihood track correlation," in *Proceedings of the 18th IEEE Decision and Control Conference* (1979).

38. H. Siu and A. Pue, "Sensor data fusion for terminal homing missiles," in *Proceedings of the Fourth National Symposium on Sensor Fusion,* Vol. 1, pp. 191–210 (1991).

39. A. J. Maren, M. D. Lothers, R. M. Pap, and R. Akita, "Neural network sensor data fusion methods for naval air traffic control," in *Proceedings of the Fifth National Symposium on Sensor Fusion,* Vol. 1, pp. 285–294 (1992).

40. D. B. Reid, "An algorithm for tracking multiple targets," *IEEE Transactions on Automatic Control* **AC-24,** 843–854 (Dec. 1979).

41. Jay B. Jordan and Howon Choe, "A comparative analysis of statistical, fuzzy and artificial neural pattern recognition techniques," in *Signal Processing, Sensor Fusion, and Target Recognition, Proceedings of the SPIE* **1699,** 166–176 (1992).

42. T. D. Garvey, "Survey of AI approaches to the integration of information," in *Infrared Sensors and Sensor Fusion, Proceedings of the SPIE* **782,** 68–82 (1987).

43. Richard O. Duda and Peter E. Hart, *Pattern Classification and Scene Analysis,* Wiley-Interscience, New York (1973).

44. Belur Dasarathy, *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques,* IEEE Computer Society Press, Los Alamitos, CA (1991).

45. David E. Rumelhart and James L. McClelland, Eds., *Parallel Distributed Processing,* pp. 45–76, MIT Press, Cambridge, MA (1986).

46. Judea Pearl, *Probabilistic Reasoning in Intelligent Systems,* Morgan Kaufman, New York (1988).

47. S. C. A. Thomoupoulos and L. Zhang, "Networking delay and channel errors in distributed data fusion," in *Sensor Fusion I, Proceedings of the SPIE* **931,** 154–160 (1988).

48. S. C. A. Thomoupoulos, D. K. Bougoulias, and L. Zhang, "Optimal and suboptimal distributed decision fusion," in *Sensor Fusion I, Proceedings of the SPIE* **931,** 26–30 (1988).

49. S. C. A. Thomoupoulos and N. N. Okello, "Distributed detection with consulting sensors and communication cost," in *Sensor Fusion I, Proceedings of the SPIE* **931,** 31–40 (1988).

50. Dale Fixsen and Ronald Mahler, "A Dempster-Shafer approach to Bayesian classification," in *Proceedings of the Fifth National Symposium on Sensor Fusion,* Vol. 1, pp. 213–231 (1992).

51. E. L. Waltz and D. M. Buede, "Data fusion and decision support for command and control," *IEEE Transactions on Systems, Man and Cybernetics* **SMC-16**(16), 865–879 (1986).

52. Glen Shafer, *A Mathematical Theory of Evidence,* Princeton University Press, Princeton (1976).

53. John D. Lowrence and Thomas D. Garvey, *Evidential Reasoning: an Implementation for Multisensor Integration*, SRI International Technical Report 307 (1983).

54. John D. Lowrence, Thomas D. Garvey, and Thomas M. Strat, "A framework for evidential-reasoning systems," in *Proceedings of the National Conference on Artificial Intelligence*, pp. 896–903 (1986).

55. Douglas J. Hunt, Jack L. Sedwick, and William P. Zachar, "Determination of Dempster-Shafer mass vectors on k-nearest neighbor classifiers for multisensor fusion," in *Proceedings of the Fifth National Symposium on Sensor Fusion*, Vol. 1, pp. 233–240 (1992).

56. J. Gordon and E. H. Shortliffe, "Evidential reasoning in a hierarchical hypothesis space," *AI* **26**, 323–357 (1985).

57. F. I. Dretsky, *Knowledge and the Flow of Information*, MIT Press, Cambridge, MA (1982).

58. B. A. Kelley and W. R. Simpson, "The use of information theory in propositional calculus," in *Proceedings of the Tri-Service Data Fusion Symposium*, Vol. 1, pp. 125–141 (1987).

59. L. A. Zadeh, "Review of *A Mathematical Theory of Evidence*, by Glen Shafer," *Artificial Intelligence* **5**(3), 81 (1984).

60. L. A. Zadeh,"Possibility theory and soft data analysis," in *Mathematical Frontier of the Social and Policy Sciences*, L. Cobb and R. M. Thrall, Eds., pp. 69–129 (1981).

61. Richard O. Duda, Peter E. Hart, and Rene Reboh, "A rule-based consultation program for mineral exploration," in *Proceedings of the Lawrence Symposium Systems and Decision Sciences*, pp. 306–309 (1977).

62. M. S. Aldenderfer and R. K. Blashfield, "Cluster analysis," in *Quantitative Applications in the Social Sciences Series*, Paper 07–044, Sage Publications, London (1984).

63. P. Sneath and R. Sokal, *Numerical Taxonomy*, W. H. Freeman Publisher, San Francisco (1973).

64. A. Lippmann, "An introduction to computing neural nets," *AI Magazine* **3**(4) (1987).

65. Angel DeCegama and Jeff Smith, "Neural networks and genetic algorithms for combinatorial optimization of sensor data fusion," in *Signal Processing, Sensor Fusion, and Target Recognition, Proceedings of the SPIE* **1699**, 108–115 (1992).

66. C. Priebe and D. Marchette, "An application of neural networks to a data fusion problem," in *Proceedings of the Tri-Service Data Fusion Symposium*, Vol. 1, pp. 226–236 (1987).

67. B. Widrow and R. Winter, "Neural nets for adaptive filtering and adaptive pattern recognition," *IEEE Computer* (Mar. 1988).

68. R. North, "Neurocomputing: its impact on the future of defense systems," *Defense Computing* (Jan. 1988).

69. Thomas L. Dean, "Planning and temporal reasoning under uncertainty," in *Proceedings of the IEEE Workshop on Principles of Knowledge-Based Systems*, pp. 131–137 (1984).

70. Richard Alterman, "An adaptive planner," in *Proceedings of the American Association for Artificial Intelligence AAAI-86*, pp. 65–69 (1986).

71. F. L. Wright, "The fusion of multisensor data," *Signal* (Oct. 1989).

72. M. A. Griesel, "Cross-correlation statistics, templating and doctrine," *U.S. Army Intelligence Center and School*, NTIS AD-A155-624 (February 1984).

73. R. Turner, *Logics for Artificial Intelligence*, Halstead-Wiley (1985).

74. Edward E. Shortliffe, *Computer-Based Medical Consultations: MYCIN*, Elsevier, New York (1976).

75. Martin E. Liggins, II, William P. Berry, and Jack A. Mineo, "Multispectral sensor fusion testbed performance under 'noisy' sensor operations," in *Proceedings of the Third National Symposium on Sensor Fusion*, Vol. 1, pp. 27–45 (1990).

76. Alan N. Steinberg, "Sensor/response coordination in a tactical self-protection system," in *Sensor Fusion I, Proceedings of the SPIE* **931**, 115–122 (1988).

77. Paul C. Chestnut, "Emitter location accuracy using TDOA and differential doppler," *IEEE Transactions Aerospace and Electronic Systems* **AES-18**(2), 214–218 (1982).

78. Robert Berezdivin, Ray Rose, John McKee, and Jay Silverman, "Passive RF tracking technology in multisensor systems," in *Proceedings of the Second National Symposium on Sensor Fusion*, Vol. 2, pp. 479–516 (1989).

79. Richard M. Aronson, "Electronic support measures and bistatic radar sensor integration," in *Proceedings of the Fourth National Symposium on Sensor Fusion*, Vol. 2, pp. 77–96 (1991).

80. Martin E. Liggins, II, William P. Berry, Edwin L. Post, and Lisa A. Osadciw, "Real time cueing methodologies and performance for multispectral sensor fusion," in *Proceedings of the Tri-Service Data Fusion Symposium,* Vol. 1, pp. 81–104 (1990).

81. Alan N. Steinberg and Douglas P. Glasson, "A real-time tactical mission planner," in *Proceedings of the Second National Symposium on Data Fusion,* Vol. 2, pp. 461–476 (1989).

82. James C. Moore and Andrew B. Whinston, "A model of decision-making with sequential information acquisition," *Decision Support Systems* **2,** 285–307 (1986); **3,** 47–72 (1987).

83. Alan N. Steinberg, "A planner for threat assessment and response," in *Proceedings of the National Aerospace and Electronics Conference, NAECON-88,* pp. 1121–1127 (1988).

84. D. R. Drew, A. A. Trani, and T. K. Tran, "Aircraft survivability and lethality trade-off model (ASALT)," *Aircraft Survivability* **10**(2), 13–18 (1986).

85. Mark Feldman, "Achieving compatibility," *Journal of Electronic Defense,* 42–45 (Dec. 1991).

86. Richard Holmes and Sol Rocklin, "Performance modeling of multisensor networks," in *Sensor Fusion, Proceedings of the SPIE* **1306,** 136–146 (1990).

87. Richard B. Holmes, *Applications of Information Theory to Sensor Network Modeling,* Massachusetts Institute of Technology Lincoln Laboratory Technical Report 829 (June 1992).

88. M. Alford and D. Wynne, "A multitarget, multisensor tracking simulation for comparison of fusion algorithm performance," in *Proceedings of the Third National Symposium on Sensor Fusion,* Vol. 1, pp. 395–406 (1990).

89. J. F. LoSecco et al., *Target Recognition for Electronic Combat, Final Technical Report,* RADC Contract F30602-87-C-0150 (Dec. 1989).

90. N. W. Sojourner, "Sensor fusion at the National Test Facility," in *Proceedings of the Third National Symposium on Sensor Fusion,* Vol. 1, pp. 83–114 (1980).

91. *Surveillance Test Bed Design Documentation,* GE Aerospace Report, SDIO Contract SDIO84-C-0020 (May 1991).

92. *Data Fusion Demonstrator Defining Document,* NATO AC/243, Panel 3 Report (Feb. 1987).

93. D. Brown et al., "ASSET: all-source systems evaluation test bed," paper presented at the Tri-Service Data Fusion Symposium (Oct. 1991).

94. D. D. Freedman, "Interface control document for the fusion evaluation test bed," MITRE Working Paper WP 28812-1-1 (July 1991).

95. P. Gottschalk, R. Freeling, E. T. Buxton, P. Seegers, and F. Smith, "Development and application of MultiSIM: a multiple sensor data fusion simulation model," in *Proceedings of the Fifth National Symposium on Sensor Fusion,* Vol. 1, pp. 41–60 (1992).

96. K. F. Lavering, E. G. Muehle, K. Y. Ung, and R. Burkhart, "Sensor data fusion analyst testbed," in *Proceedings of the Fifth National Symposium on Sensor Fusion,* Vol. 1, pp. 75–80 (1992).

97. D. Buede, K. Lasky and C. Stewart, "Development of and research with the modular fusion testbed (MOFUT)," in *Proceedings of the Fifth National Symposium on Sensor Fusion,* Vol. 1, pp. 61–74 (1992).

98. R. Mahler, "Using *a priori* evidence to customize Dempster-Shafer theory," in *Proceedings of the Sixth National Symposium on Sensor Fusion* (in press).

99. D. E. Maurer, K. L. Dates, and A. K. Chrysostomou, "A dissimilarity metric for distributed multiple-hypothesis correlation and tracking," in *Proceedings of the Sixth National Syposium on Sensor Fusion* (in press).

100. Myron Greenbaum and Alan Steinberg, "Integrated electronic warfare systems and technology," paper presented at the Association of Old Crows Northeastern Regional Symposium (24 July 1985).

Bibliography

Aboutalib, A. Omar and Thanh Luu, "An efficient target extraction technique for laser radar imagery," in *Aerospace Sensing, Proceedings of the SPIE* **1096,** 23–33 (1989).

Aboutalib, A. Omar, L. Tran, and C.-Y. Hu, "Fusion of passive imaging sensors for target acquisition," in *Proceedings of the Fifth National Symposium on Sensor Fusion,* Vol. 1, pp. 151–162 (1992).

Aboutalib, A. Omar, "Multisensor fusion for target detection, tracking and recognition," in *Proceedings of the Fourth National Symposium on Sensor Fusion,* Vol. 1, pp. 99–118 (1991).

Aboutalib, A. Omar, "Multisensor data fusion and multisensor techniques," Course material, Center for Professional Development at the University of Maryland University College (Sep. 1989).

Aldenderfer, M. S. and R. K. Blashfield, "Cluster analysis," *Quantitative Applications in the Social Sciences Series,* Paper 07-044, Sage Publications, London (1984).

Alford, M. and D. Wynne, "A multitarget, multisensor tracking simulation for comparison of fusion algorithm performance," in *Proceedings of the Third National Symposium on Sensor Fusion* (1990).

Alterman, Richard, "An adaptive planner," in *Proceedings of the American Association for Artificial Intelligence AAAI-86,* pp. 65–69 (1986).

Amai, Wendy, "A weighting/threshold approach to sensor fusion," in *Proceedings of the First National Symposium on Sensor Fusion,* Vol. 2, pp. 455–468 (1988).

Aronson, Richard M., "Electronic support measures and bistatic radar sensor integration," in *Proceedings of the Fourth National Symposium on Sensor Fusion,* Vol. 2, pp. 77–96 (1991).

Bachelor, J., "Combined mode countermeasures," in *Proceedings of the Fourth National Symposium on Sensor Fusion,* Vol. 2, pp. 133–152 (1991).

Bar-Shalom, Y. and T. E. Fortman, *Tracking and Data Association,* Academic Press, San Diego (1988).

Barr, John K., "An application of Computerized Axial Tomography (CAT) technology to mass raid tracking," in *Proceedings of the Tri-Service Data Fusion Symposium,* Vol. 1, pp. 155–160 (1988).

Becker, D. and A. M. Haimovich, "Sensor fusion for the F-14D," in *Proceedings of the Fourth National Symposium on Sensor Fusion,* Vol. 2, pp. 7–24 (1991).

Beltz, John W., "Belief representation for fusion using evidence theory," in *Proceedings of the Tri-Service Data Fusion Symposium,* Vol. 1 (1988).

Berezdivin, Robert, Ray Rose, John McKee, and Jay Silverman, "Passive RF tracking technology in multisensor systems," in *Proceedings of the Second National Symposium on Sensor Fusion,* Vol. 2, pp. 479–516 (1989).

Borky, John M., "Modular integrated avionics—strategies and challenges," *Journal of Electronic Defense* 15(5), 36–43 (May 1992).

Bowman, Christopher L., "Artificial neural network adaptive systems applied to multi-sensor ID," in *Proceedings of the Tri-Service Data Fusion Symposium,* Vol. 1, pp. 161–172 (1988).

Bowman, Christopher L. and Ronald K. Munzer, *Integrated Fire Control Surveillance, Final Technical Report,* Vol. 1, AFWAL-TR-80-1151 VERAC, Inc., San Diego (October 1980).

Bowman, Christopher L., "Multisensor track file correlation using attributes and kinematics," in *Proceedings of the National Aerospace and Electronics Conference,* pp. 1181–1186 (1980).

Bowman, Christopher L. and Gleason Snashall, "Internetted fighter sensor data association and state estimation approach comparison," in *Proceedings of the Second National Symposium on Sensor Fusion,* Vol. 2, pp. 297–320 (1989).

Bowman, Christopher L., "Max likelihood track correlation," in *Proceedings of the 18th IEEE Decision and Control Conference* (1979).

Bowman, Christopher L., "Data integration (fusion) tree paradigm," in *Signal and Data Processing of Small Targets, Proceedings of the SPIE* **1698,** 372–381 (1992).

Brandstadt, Jeff, "Applied probabilistic data association," in *Proceedings of the Tri-State Data Fusion Symposium,* Vol. 1, pp. 220–230 (1988).

Brandstadt, Jeff and Lisa A. Maston, "Correlation of dissimilar sensor reports–an application of multiple hypotheses to the attack helicopter mission," in *Proceedings of the Fifth National Symposium on Sensor Fusion,* Vol. 1, pp. 241–260 (1992).

Brown, D., et al., "ASSET: All-Source Systems Evaluation Test Bed," paper presented at the Tri-Service Data Fusion Symposium (Oct. 1991).

Buede, D., K. Lasky, and C. Stewart, "Development of and research with the modular fusion testbed (MOFUT)," in *Proceedings of the Fifth National Symposium on Sensor Fusion,* Vol. 1, pp. 61–74 (1992).

Buede, D. and E. L. Waltz, "Benefits of soft sensors and probabilistic fusion," in *Signal and Data Processing of Small Targets, Proceedings of the SPIE* **1096,** 309–320 (1989).

Camana, Peter, "Integrated avionics," in *Proceedings of the First National Symposium on Sensor Fusion,* Vol. 2, pp. 409–421 (1988).

Clawson, Carl W., "On the choice of coordinate systems for fusion of passive tracking data," in

Proceedings of the Tri-Service Data Fusion Symposium, pp. 247–262 (1990).

Comparato, Vito G., "Fusion—the key to tactical mission success," in *Proceedings of the First National Symposium on Sensor Fusion*, Vol. 2, pp. 1–8 (1988).

Cox, Ben and Douglas G. Bowen, "Tactical communications to support intelligence," in *Proceedings of the First National Symposium on Sensor Fusion*, Vol. 2, pp. 183–192 (1988).

Dasarathy, Belur, *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, IEEE Computer Society Press, Los Alamitos, CA (1991).

Data Fusion Demonstrator Defining Document, NATO AC/243, Panel 3 Report (Feb. 1987).

Dean, Thomas L., "Planning and temporal reasoning under uncertainty," in *Proceedings of the IEEE Workshop on Principles of Knowledge-Based Systems*, pp. 131–137 (1984).

Delanoy, R., J. Verly, and D. Dudgeon, "Pixel-level fusion using "interest" images," in *Proceedings of the Fourth National Symposium on Sensor Fusion*, Vol. 1, pp. 29–42 (1991).

Dempster, Arthur P., "A generalization of Bayesian inference," *Journal of the Royal Statistical Society*, Series B (30), 205–247 (1968).

Dempster, Arthur P., "Upper and lower probabilities induced by a multi-valued mapping," *Annals of Mathematical Statistics* no. 38 (1967).

Department of Defense Critical Technology Plan for the Committee on Armed Services, United States Congress, AD-A234 900 (1 May 1991).

DeCegama, Angel and Jeff Smith, "Neural networks and genetic algorithms for combinatorial optimization of sensor data fusion," in *Signal Processing, Sensor Fusion, and Target Recognition, Proceedings of the SPIE* **1699,** 108–115 (1992).

Doyle, R. J., D. J. Atkinson, and R. S. Doshi, "Generating perception requests and expectations to verify the execution of plans," in *Proceedings of the American Association for Artificial Intelligence AAAI-86*, pp. 81–88 (1986).

Dretsky, F. I., *Knowledge and the Flow of Information*, MIT Press, Cambridge, MA (1982).

Drew, D. R., A. A. Trani, and T. K. Tran, "Aircraft survivability and lethality trade-off model (ASALT)," *Aircraft Survivability* **10**(2), 13–18 (1986).

Drummond, Oliver E. and Samuel S. Blackman, "Multiple sensor, multiple target tracking challenges of the Strategic Defense Initiative," in *Proceedings of the First National Symposium on Sensor Fusion*, Vol. 2, pp. 503–514 (1988).

Drummond, Oliver E., S. S. Blackman, and K. D. Hell, "Multiple sensor tracking of clusters of extended objects," in *Proceedings of the Tri-Service Data Fusion Symposium*, Vol. 1, pp. 231–238 (1988).

Duda, Richard O. and Peter E. Hart, *Pattern Classification and Scene Analysis*, Wiley-Interscience, New York (1973).

Duda, Richard O., Peter E. Hart, and Rene Reboh, "A rule-based consultation program for mineral exploration," in *Proceedings of the Lawrence Symposium Systems and Decision Sciences*, pp. 306–309 (1977).

Feldman, Mark, "Achieving compatibility," *Journal of Electronic Defense*, 42–45 (Dec. 1991).

Firby, James R., "An investigation into reactive planning in complex domains," in *Proceedings of the American Association for Artificial Intelligence AAAI-87*, pp. 202–206 (1987).

Fixsen, Dale and Ronald Mahler, "A Dempster-Shafer approach to Bayesian classification," in *Proceedings of the Fifth National Symposium on Sensor Fusion*, Vol. 1, pp. 213–231 (1992).

Freedman, D. D., "Interface control document for the fusion evaluation test bed," MITRE working paper WP 28812-1-1 (July 1991).

Fucik, George and Stephen J. Gottron, "Development of efficient and effective multisensor processing systems," in *Proceedings of the First National Symposium on Sensor Fusion*, Vol. 2, pp. 275–318 (1988).

Gamble, S. O., W. R. Ditzler, and A. G. Sutton, "Multisource integration for a Navy tactical aircraft," in *Proceedings of the Fifth National Symposium on Sensor Fusion*, Vol. 1, pp. 117–134 (1992).

Garvey, T. D., J. D. Lowrence, and M. A. Fischler, "An inference technique for integrating knowledge from disparate sources," in *Proceedings of the Seventh International Joint Conference on Artificial Intelligence*, pp. 319–325 (1981).

Garvey, T. D., "Survey of AI approaches to the integration of information," in *Infrared Sensors and Sensor Fusion, Proceedings of the SPIE* **782,** 68–82 (1987).

Goodman, I. R., "A general theory for the fusion of data," in *Proceedings of the Tri-Service Data*

Fusion Symposium, Vol. 1, pp. 252–275 (1987).

Goodman, I. R., "Application of a conditional event algebra to data fusion," in *Proceedings of the Tri-Service Data Fusion Symposium,* Vol. 1, pp. 179–188 (1988).

Goodman, I. R. and H. T. Nguyen, *Uncertainty Models for Knowledge-Based Systems,* North-Holland Co., Amsterdam (1985).

Goodman, I. R., "A probabilistic/possibilistic approach to modeling C^3 systems," in *Proceedings of the Ninth MIT/ONR Workshop on C^3 Systems,* pp. 161–166 (1986).

Gordon, J. and E. H. Shortliffe, "Evidential reasoning in a hierarchical hypothesis space," *AI* **26,** 323–357 (1985).

GPALS Navigation Standards, SDIO Report SDI-P-SD-92-000002 (July 1992).

Gottschalk, P., R. Freeling, E. T. Buxton, P. Seegers, and F. Smith, "Development and application of MultiSIM: a multiple sensor data fusion simulation model," in *Proceedings of the Fifth National Symposium on Sensor Fusion,* Vol. 1, pp. 41–60 (1992).

Greenbaum, Myron and Alan Steinberg, "Integrated electronic warfare systems and technology," paper presented at AOC Northeastern Regional Symposium (24 July 1985).

Griesel, M. A., "Cross-correlation statistics, templating and doctrine," *U.S. Army Intelligence Center and School,* NTIS AD-A155-624 (Feb. 1984).

Hall, David L., *Mathematical Techniques in Multisensor Data Fusion,* Artech House, Boston (1992).

Hansen, Allan R., Edward M. Riesman, and Thomas D. Williams, "Sensor and information fusion from knowledge based constraints," in *Proceedings of the Tri-Service Data Fusion Symposium,* pp. 113–122 (1988).

Harney, Robert C., "Implementation of multiple sensors in weapon systems," Course material, Center for Professional Development at the University of Maryland University College (September 1989).

Harney, Robert C., "Practical issues in multisensor target recognition," in *Proceedings of the Third National Symposium on Sensor Fusion,* pp. 105–114 (1990).

Harris, S., J. Owens, C. Barrett and M. Parisi, "Sensor fusion and situation assessment: a perspective," in *Proceedings of the Fourth National Symposium on Sensor Fusion,* Vol. 2, pp. 195–220 (1991).

Hart, Thomas M., "An ESM data fusion process for tracking mass raids," in *Proceedings of the Tri-Service Data Fusion Symposium,* Vol. 1, pp. 149–154 (1988).

Hitzelberger, R. and V. Jonkoff, "Systems integration: an Air Force perspective," *Defense Electronics,* 95–102 (Nov. 1986).

Holmes, Richard B., "Applications of Information Theory to Sensor Network Modeling," Massachusetts Institute of Technology Lincoln Laboratory Technical Report 829 (June 1992).

Holmes, Richard and Sol Rocklin, "Performance modeling of multisensor networks," in *Sensor Fusion, Proceedings of the SPIE* **1306,** 136–146 (1990).

Hoppenstand, Gregory S., "A real-time fusion capability applied to drug interdiction: project Classic Gangster," in *Proceedings of the Tri-Service Data Fusion Symposium,* Vol. 1 (1990).

Hunt, Douglas J., Jack L. Sedwick, and William P. Zachar, "Determination of Dempster-Shafer mass vectors on k-nearest neighbor classifiers for multisensor fusion," in *Proceedings of the Fifth National Symposium on Sensor Fusion,* Vol. 1, pp. 233–240 (1992).

"INEWS sensor data fusion final report," Verac, Inc., Report R-071-85 (1985).

Jernigan, Gregg, "Dynamic profiling of enemy operations: fusing multisource data with terrain and doctrine to predict enemy action," in *Proceedings of the Tri-Service Data Fusion Symposium,* Vol. 1 (1990).

Jordan, Jay B. and Howon Choe, "A comparative analysis of statistical, fuzzy and artificial neural pattern recognition techniques," in *Signal Processing, Sensor Fusion, and Target Recognition, Proceedings of the SPIE* **1699,** 166–176 (1992).

Kaun, Carl F., "Technologies for integrated fusion/battle management," in *Proceedings of the Tri-Service Data Fusion Symposium,* Vol. 1, pp. 303–314 (1990).

Kelley, B. A. and W. R. Simpson, "The use of information theory in propositional calculus," in *Proceedings of the Tri-Service Data Fusion Symposium,* Vol. 1, pp. 125–141 (1987).

Kenyon, Stephen C., "Fusion of target features and target dynamics to enhance multitarget tracking," in *Proceedings of the First National Symposium on Sensor Fusion,* Vol. 2, pp. 423–431 (1988).

Kessler, Otto, et al., *Functional Description of the Data Fusion Process,* prepared for the Office of

Naval Technology Data Fusion Development Strategy (November 1991).

Krueger, Charles H., Jr., "Technology development in integrated avionics," *Journal of Electronic Defense* **15**(5), 44–49 (1992).

Lavering, K. F., E. G. Muehle, K. Y. Ung and R. Burkhart, "Sensor data fusion analyst testbed," in *Proceedings of the Fifth National Symposium on Sensor Fusion,* Vol. 1, pp. 75–80 (1992).

Lawson, Joel S., Jr., "Command and control as a process," in *IEEE Symposium Proceedings* (1980).

Lawton, Daryl T. and Michael Callahan, "A simple tunable interest operator and some applications," in *Applications of Artificial Intelligence II, Proceedings of the SPIE* **548,** 186–193 (1985).

Liggins, Martin E., II, William P. Berry, and Jack A. Mineo, "Multispectral sensor fusion testbed performance under "noisy" sensor operations," in *Proceedings of the Third National Symposium on Sensor Fusion,* Vol. 1, pp. 27–45 (1990).

Liggins, Martin E., II, William P. Berry, Edwin L. Post, and Lisa A. Osadciw, "Real time cueing methodologies and performance for multispectral sensor fusion," in *Proceedings of the Tri-Service Data Fusion Symposium,* Vol. 1, pp. 81–104 (1990).

Linn, R. J., D. L. Hall, and J. Llinas, "A survey of multisensor data fusion systems," in *Sensor Fusion, Proceedings of the SPIE* **1470,** 13–29 (1991).

Lippmann, A., "An introduction to computing neural nets," *AI Magazine* **3**(4) (1987).

Llinas, J., "Assessing the performance of multisensor fusion systems," in *Sensor Fusion, Proceedings of the SPIE* **1611** (1991).

LoSecco, J. F., et al., *Target Recognition for Electronic Combat, Final Technical Report,* RADC Contract F30602-87-C-0150 (Dec. 1989).

Lowrence, John D. and Thomas D. Garvey, *Evidential Reasoning: an Implementation for Multisensor Integration,* SRI International Technical Report 307 (1983).

Lowrence, John D., Thomas D. Garvey, and Thomas M. Strat, "A framework for evidential-reasoning systems," in *Proceedings of the National Conference on Artificial Intelligence,* pp. 896–903 (1986).

Maren, A. J., M. D. Lothers, R. M. Pap and R. Akita, "Neural network sensor data fusion methods for naval air traffic control," in *Proceedings of the Fifth National Symposium on Sensor Fusion,* Vol. 1, pp. 285–294 (1992).

McFawn, L., "Advanced concepts for avionics system architectures," paper presented at the Second National Symposium on Sensor Fusion, Orlando, FL (1989).

Menon, M. and P. Kolodzy, "Active/passive IR scene enhancement by Markov random field sensor fusion," in *Proceedings of the Fourth National Symposium on Sensor Fusion,* Vol. 1, pp. 155–172 (1991).

Moore, James C. and Andrew B. Whinston, "A model of decision-making with sequential information acquisition," *Decision Support Systems,* **2,** 285–307 (1986); **3,** 47–72 (1987).

Morgan, R., "PAVE PACE: systems avionics for the 21st century," in *Proceedings of the National Aerospace and Electronics Conference,* pp. 146–157 (1988).

Morgenstern, Leora, "Knowledge preconditions for actions and plans," in *Proceedings of the International Joint Conference on Artificial Intelligence IJCAI-87,* pp. 867–874 (1987).

Nagy, P. G. and P. L. Rothman, "Internetted airborne sensor management systems," in *Proceedings of the Tri-Service Data Sensor Symposium,* Vol. 1, pp. 193–206 (1990).

Noble, David, "Managing temporal uncertainty in situation assessment," in *Proceedings of the Tri-Service Data Sensor Symposium,* Vol. 1, pp. 112–117 (1988).

North, R., "Neurocomputing: its impact on the future of defense systems," *Defense Computing* (Jan. 1988).

O'Neil, S. and M. Bridgland, "Fast algorithms for multi-target data association," in *Proceedings of the Fourth National Symposium on Sensor Fusion,* Vol. 1, pp. 173–190 (1991).

Ogrodnik, R. F., "Adaptive multispectral bistatic/ESM fusion for target detection, track and ID," in *Proceedings of the Fifth National Symposium on Sensor Fusion,* Vol. 1, pp. 163–174 (1992).

Osche, G. R., D. S. Young, W. Tanaka, J. Nettleton, and W. Harmon, "Tri-service active/passive CO_2 laser radar," in *Proceedings of the First National Symposium on Sensor Fusion,* Vol. 2, pp. 491–502 (1988).

Pearl, Judea, *Probabilistic Reasoning in Intelligent Systems,* Morgan Kaufman, New York (1988).

Price, G. E. and D. J. Bielecki, "Choke point ocean surveillance," in *Proceedings of the Fifth National Symposium on Sensor Fusion,* Vol. 1 (1992).

Priebe, C. and D. Marchette, "An application of neural networks to a data fusion problem," in

Proceedings of the Tri-Service Data Fusion Symposium, Vol. 1, pp. 226–236 (1987).

Reid, D. B., "An algorithm for tracking multiple targets," *IEEE Transactions on Automatic Control* **AC-24,** 843–854 (Dec. 1979).

Richman, Dennis, "Multisensor data fusion and multisensor techniques," Course material, Center for Professional Development at the University of Maryland University College (September 1989).

Robinson, Guner S. and A. Omar Aboutalib, "Trade-off analysis of multisensor fusion levels," in *Proceedings of the Second National Symposium on Sensor Fusion,* Vol. 2, pp. 21–34 (1989).

Robinson, Guner S., "ATR systems: a Northrop perspective," *Proceedings of the SPIE* **750,** 216–224 (1987).

Romie, R. E., R. C. Beeching, A. S. Brothers, K. H. Irk, and T. P. Zahm, "MATCHWELL: an integrated data fusion program," in *Proceedings of the Fifth National Symposium on Sensor Fusion,* Vol. 1 (1992).

Rosen, Julie A. and Peter S. Schoenfeld, "Parallel processing applications to multi-source correlation and tracking," in *Proceedings of the Tri-Service Data Fusion Symposium,* Vol. 1, pp. 202–214 (1987).

Rumelhart, D. E., G. E. Hinton, and J. L. McClelland, "A general framework for parallel distributed processing," in *Parallel Distributed Processing,* David E. Rumelhart and James L. McClelland, Eds., pp. 45–76, MIT Press, Cambridge, MA (1986).

Schleher, D. Curtiss, *Introduction to Electronic Warfare,* Artech House, Norwood, MA (1986).

Schneider, Bernard A., "Mission/flight systems integration study," in *Proceedings of the Second National Symposium on Sensor Fusion,* pp. 683–692 (1989).

Shafer, Glen, *A Mathematical Theory of Evidence,* Princeton University Press, Princeton (1976).

Shortliffe, Edward E., *Computer-Based Medical Consultations: MYCIN,* Elsevier, New York (1976).

Silk, J. D., J. D. Nicoll, and D. A. Sparrow, "Modeling the performance of fused sensor ATRs," in *Proceedings of the Fourth National Symposium on Sensor Fusion,* Vol. 1, pp. 323–336 (1991).

Sindoris, Arthur R. and Norman J. Berg, "The Army's investment strategy for automatic target recognition technology," in *Army Research, Development and Acquisition Bulletin,* pp. 4–8 (Nov.–Dec. 1990).

Siu, H. and A. Pue, "Sensor data fusion for terminal homing missiles," in *Proceedings of the Fourth National Symposium on Sensor Fusion,* Vol. 1, pp. 191–210 (1991).

Sneath, P. and R. Sokal, *Numerical Taxonomy,* W. H. Freeman, San Francisco (1973).

Sojourner, N. W., "Sensor fusion at the National Test Facility," in *Proceedings of the Third National Symposium on Sensor Fusion* (1980).

Steinberg, Alan N., "Sensor/response coordination in a tactical self-protection system," in *Sensor Fusion I, Proceedings of the SPIE* **931,** 115–122 (1988).

Steinberg, Alan N., "A closed-loop simulator for tactical aircraft systems," in *Proceedings of the National Aerospace and Electronics Conference NAECON-87,* pp. 1034–1041 (1987).

Steinberg, Alan N., "High-level issues in multisensor data fusion," Course material, Center for Professional Development at the University of Maryland University College (Sep. 1989).

Steinberg, Alan N. and Douglas P. Glasson, "A real-time tactical mission planner," in *Proceedings of the Second National Symposium on Sensor Fusion,* Vol. 2, pp. 461–476 (1989).

Steinberg, Alan N., "A planner for threat assessment and response," in *Proceedings of the National Aerospace and Electronics Conference NAECON-88,* pp. 1121–1127 (1988).

Steinberg, Alan N., "An expert system for multispectral threat assessment and response," in *Applications of Artificial Intelligence V, Proceedings of the SPIE* **786,** 52–62 (1987).

Steinberg, Alan N., "Threat management system for combat aircraft," in *Proceedings of the Tri-Service Data Fusion Symposium,* Vol. 1, pp. 528–542 (1987).

Surveillance Test Bed Design Documentation, GE Aerospace Report, SDIO Contract SDIO84-C-0020 (May 1991).

Tenney, R. R. and N. R. Sandell, Jr., "Detection with distributed sensors," *IEEE Transactions on Aerospace and Electronics Systems* **AES-17,** 501–512 (July 1981).

Terrance, J., III, "Integrated avionics—are we ready?" *Journal of Electronic Defense,* 47–51 (March 1989).

Thomoupoulos, S. C. A. and L. Zhang, "Networking delay and channel errors in distributed data fusion," in *Sensor Fusion I, Proceedings of the SPIE* **931,** 154–160 (1988).

Thomoupoulos, S. C. A., D. K. Bougoulias, and L. Zhang, "Optimal and suboptimal distributed decision fusion," in *Sensor Fusion I, Proceedings of the SPIE* **931,** 26–30 (1988).

Thomoupoulos, S. C. A. and N. N. Okello, "Distributed detection with consulting sensors and communication cost," in *Sensor Fusion I, Proceedings of the SPIE* **931,** 31–40 (1988).

Turner, R., *Logics for Artificial Intelligence,* Halstead-Wiley (1985).

Vannicola, Vincent, C., "Algorithm development for sensor fusion," in *Proceedings of the First National Symposium on Sensor Fusion,* Vol. 2, pp. 469–472 (1988).

Van Trees, H. L., *Detection, Estimation and Modulation Theory,* John Wiley & Sons, New York (1968).

Waltz, E. L. and D. M. Buede, "Data fusion and decision support for command and control," *IEEE Transactions on Systems, Man and Cybernetics* **SM-16**(16), 865–879 (Nov.–Dec. 1986).

Waltz, Edward and James Llinas, *Multisensor Data Fusion,* Artech House, Boston (1990).

White, Franklin E., Jr., *Data Fusion Lexicon,* prepared for Joint Directors of Laboratories, Technical Panel for C^3, Data Fusion Sub-Panel, Naval Ocean Systems Center, San Diego (1987).

White, Franklin E., Jr., "A model for data fusion," in *Proceedings of the First National Symposium on Sensor Fusion,* Vol. 2, pp. 143–187 (1988).

Widrow, B. and R. Winter, "Neural nets for adaptive filtering and adaptive pattern recognition," *IEEE Computer* (March 1988).

Wise, Carl and Jim Armitage, "Electronic combat and weapon system integration," *Journal of Electronic Defense* **14**(7), 37–39 (July 1991).

Wright, F. L., "The fusion of multisensor data," *Signal* (Oct. 1989).

Zadeh, L. A., "Review of *A Mathematical Theory of Evidence,* by Glen Shafer," *AI* **5**(3), 81 (1984).

Zadeh, L. A., "Possibility theory and soft data analysis," in *Mathematical Frontier of the Social and Policy Sciences,* L. Cobb and R. M. Thrall, Eds., pp. 69–129 (1981).

Zadeh, L. A., "A simple view of the Dempster-Shafer theory of evidence and its implication for the rule of combination," *AI* **7**(2), 85–90 (1986).

Automatic Target Recognition Systems

James W. Sherman
David N. Spector
C. W. "Ron" Swonger
Environmental Research Institute of Michigan
Ann Arbor, Michigan

Lloyd G. Clark
Edmund G. Zelnio
U.S. Air Force Wright Laboratory
Wright-Patterson Air Force Base, Ohio

Martin J. Lahart
Terry L. Jones
U.S. Army Night Vision and Electronic Sensors Directorate
Fort Belvoir, Virginia

CONTENTS

4.1   INTRODUCTION

This chapter reviews a very complex and emerging collection of technologies and subsystems that, when taken together, have the potential for partial or complete automation of an essential process in warfare. This is the process of searching for, detecting, recognizing, more precisely identifying, and accurately locating targets, and directing weapons so as to neutralize those components of the enemy forces.

These emerging systems are now most commonly referred to as *automatic target recognizers*. Systems that continue to rely on human decision making to facilitate the process to some degree are often called *automatic target cueing systems*. A number of other terms have been widely applied to describe the intended functions of such systems and their prospective components over the course of their development. References to automatic target selection, automatic target identification, automatic target correlators, automatic target seekers, and aided target recognition systems have all appeared in the literature and in common use. Only automatic target recognition (ATR), automatic target cueing (ATC), or ATR/ATC are used to describe these technologies and systems in the remainder of this chapter.

The operational implementation, degree of autonomy, and many of the other functional attributes of an ATR can vary dramatically from application to application. Each postulated system generally represents a balance or compromise drawn between a set of requirements derived from the military mission and the enemy threat perspectives, on one hand, and on what technology is believed to be capable of providing, on the other.

The physical description of a particular engagement scenario defines the major variables that describe the systems performance criteria for an ATR/ATC mission. The possible variables for air-to-air, air-to-surface, surface-to-surface, surface-to-air, and other engagement scenarios span an extremely wide range. Characteristics of the weapons systems delivery platform, the sensors, and the physical observables for the targets of interest all impose significant constraints on the development of a balanced solution for an ATR application. The development of practical ATR solutions is also driven by political and economic factors, as well as all of the above, and a number of other classes of variables.

The Department of Defense Working Group on ATR Technology defines[1] ATR as a function that follows the collection of target signature data by a sensor, or by multiple sensors, in which signal processing algorithms and software and image or signal processing hardware is employed to classify, recognize, identify, interpret, and display the significance of the target data for the purpose of initiating actions in real time, or in near real time, or to provide options for actions to an external system operator. This definition provides a useful framework for a discussion of the five major ATR system elements: (1) sensors, (2) signal processing hardware, (3) algorithms, (4) systems software, and (5) operator and control systems interfaces.

All ATRs are designed around a sensor suite consisting of one or more sensors. For the most part, ATR research and development has generally involved sensors that produce sequences of either two-dimensional, three-dimensional, or higher dimensional images. Imaging sensors provide data that

are essentially views or mappings of spatial areas or volumes of the world. Images are often called scenes, frames, and so forth. Each image may, in principle, contain: (1) either none, one, or more targets of interest; (2) either none, one, or more intentional false targets; (3) a lesser or greater amount of man-made or cultural clutter; and (4) varying degrees of natural clutter representing any of the possible views of the sky, land, or water surfaces encountered on the face of the earth.

The processor contained in each ATR is the hardware subsystem that physically implements or executes the processes for arriving at target recognition or cueing decisions, measurements such as location, or other estimates for delivery to either an automatic system or to a human operator for suitable subsequent decisions and/or actions. In theory, processors may be of any physical form, but almost all modern developments are focused on either electronic digital processors, optical processors, or hybrid processors combining functional elements of both of these technologies.

An algorithm suite that contains the instructions for the execution of the automated recognition or cueing processes by the processor hardware subsystem is an essential part of every ATR system. The development of the algorithm suite has historically been the essential or core problem area addressed by ATR research. It constitutes the logic and calculations by which targets are discriminated from everything else that may appear in each image processed by the ATR. It is inherently the case that there is no universal ATR/ATC algorithm, because no two target classes, sensor suite classes, mission conditions, and so forth, are the same. To further complicate the matter, an ATR system is inherently aimed at neutralizing an intelligent adversary who may attempt to defeat the process. For these and other reasons, ATR algorithms are very difficult to develop and evaluate, their implementation is a very complex process, and they are often subject to a variety of countermeasures. Algorithms are manifested in computer software and/or hard-wired logic in digital processors. In optical processors, the algorithms are reflected in physical and/or programmable arrangements of components such as filters, lenses, and modulators.

A software operating system and utility software are also critical elements of the digital signal processing activities in ATR systems. This software does not perform the recognition or cueing functions; however, it is essential to the efficient and adequate coordination and control of the complex system of processor, sensor, and the referred to algorithm elements above. Finally, a human operator interface and the other system control interface elements are also always necessary to effectively accept direction from, and deliver results to, other systems in the overall weapon complex or platform that the ATR supports.

4.2 HISTORY

Research in the field of ATR science began in the early 1960s with the emergence of the basic concepts for automatic pattern recognition, early theoretical and experimental work on neural networks and self-organizing systems, and the first wave of artificial intelligence (AI) research. The principal body of ATR research at that time was concerned with what would today be considered

component algorithms for such functions as detection, segmentation, and classification of targets. These early efforts were primarily focused on the manipulation and exploitation of photographic imagery taken in the visible region of the electromagnetic spectrum. Much of the research at that time involved statistical pattern recognition methods, nonparametric classifiers, neural network technologies, and heuristic techniques for extracting specific attributes from images that were believed to be relevant to finding targets. Artificial intelligence research of this era was primarily limited to the search for generalized learning mechanisms by experimentation on more abstract logic problems with little, if any, application to real data representative of target images.

Pattern recognition research was dealt a significant blow by the publication and aggressive advocacy of a book[2] by Minsky and Papert of the Massachusetts Institute of Technology (MIT) in the late 1960s. The authors asserted that there were significant limitations on the practical utility of both statistical pattern recognition methods and neural networks. Government agencies that traditionally sponsored ATR research backed away from the field for most of the next decade. The Advanced Research Projects Agency (ARPA) provided the principal support for a second wave of AI research during that same period of time. The majority of these efforts were typified by the formulation of complex software systems structures intended to accomplish automated knowledge-independent learning in specific problem domains. The MIT AI Lab Building Blocks World was a part of this second wave of AI research.

The implementation, complexity, and limited learning potential of these approaches became widely perceived in the mid-1970s, and several new trends began to develop in ATR research as a result. Earlier critiques of statistical methods, nonparametric techniques, and neural network approaches were reexamined and earlier opinions were moderated. New trends were born out of the emergence of modern microprocessors and very large scale integrated circuit technologies, which opened the door for the practical implementation of techniques that were formerly considered infeasible.

There were four major trends in ATR algorithm research from the mid-1970s through the 1980s: (1) statistical pattern recognition research was combined with more capable heuristic feature extraction algorithms; (2) mathematical morphology and image algebra emerged as a mathematics of image processing, permitting effective definitions of broad classes of multidimensional, nonlinear, target-specific image transformations; (3) there was a third wave of AI research incorporating more pragmatic concepts such as knowledge-based systems, world models, reasoning under uncertainty, and other topics; and (4) there was a reincarnation of neural network research and development including a wider variety of techniques for structuring and training networks by adjusting their parameters.

Literally hundreds of companies began developing technologies and products for automatic image processing or machine vision with the advent of new computer device, architecture, and software technologies, making the implementation of these research results economically practical. There were at least 50 companies active, to one degree or another, in the more specific fields of ATR technologies. Many of the new and old organizational teams began to find solutions to specific ATR and other machine vision applications that combined the use of two or more of the previously alienated schools of statistical, non-

parametric, morphological, neural network, and artificial intelligence techniques. Systematic methodologies for characterizing ATR algorithm components and subsystems began to emerge from the research sponsored by DARPA and others at the same time. The 1980s saw the passing of ATR algorithm techniques from "childhood" at least into "puberty" as a science and engineering discipline ready to start performing significant work in the real world of defense systems.

The evolution of computers for ATR applications has followed along with the long and tortuous evolution of ATR algorithms. In hindsight, and throughout most of the cited history, the amount of computing power required for the execution of sufficiently capable ATR algorithms has been grossly underestimated by many of the workers in the field. The algorithmic complexity required for the development of a capable ATR was not well understood. Even today, that complexity is the subject of fierce debate. However, many experienced investigators would now estimate the required computing capacity for typically difficult ATR problems to be in the range of 10^{10} to 10^{14} general-purpose computer operations per second, assuming a computer with a moderately rich repertoire of instructions. The wide range of these capacity estimates reflects uncertainty about both what capable algorithms must include and the significant variation in difficulty of different ATR applications. Nevertheless, this range corresponds to the throughput of approximately 10^2 to 10^6 of the most powerful general-purpose computers available commercially in 1990.

The need to realize such extraordinary computer power in the small packages necessary for military systems applications led to the emergence of various special-purpose architectures in the 1970s and 1980s, which seek to achieve amplification of computer capacity by architectural specializations such as pipelining, parallelization, special instruction sets, vision-specialized subsystems, special-purpose semiconductor components, high-bandwidth internal data paths, and various combinations of these measures. Paradoxically, hardware architecture concepts and physical implementations have both led and impeded the development of complete operationally successful ATR algorithms. High-performance hardware is needed by developers in order for them to know of the weaknesses in their component algorithms when very high recognition accuracies are required.

4.3   ATR PERFORMANCE MEASURES

Target recognition tasks are defined in terms of the levels of discrimination achieved by the process. The Department of Defense Working Group on ATR Technology has also defined[1] a useful hierarchical system of target discrimination levels: (1) target detection, (2) target classification, (3) target recognition, (4) target identification, and (5) target characterization.

Target detection, classification, and recognition are the first three levels of ATR discrimination performance. Target detection is defined as the process of discovering the existence of a potential target from sensor data. Target classification is defined as the process of distinguishing target classes, such as combatant ships from commercial ships or tracked vehicles from wheeled vehicles. Target recognition distinguishes objects within a given class to an even higher level. Examples of recognition performance include the ability to dis-

tinguish between trucks, tanks, and armored personnel carriers within the general class of vehicle targets or the ability to distinguish a destroyer from a carrier within the general class of ship targets and a fighter from a bomber within the overall class of possible aircraft targets.

Identification and characterization are still higher levels of ATR target discrimination performance. Identification is the process of distinguishing a particular model of target from another of a similar type within an overall target class. Identification is the process of distinguishing among several models of fighter-type aircraft such as F-15s and F-16s. Characterization is the highest level of ATR target discrimination performance. Target characterization is defined as the process of clearly distinguishing the detailed physical characteristics of a particular target. This is the performance needed to assess the particular external ordnance load or other external equipment features on a target of interest.

The term *classification* is also used in a general way to describe any process of distinguishing targets from one another. Algorithms for detecting targets and classifying targets perform differently and often make use of different information. Algorithms for detecting targets generally measure size and local contrast, and they sometimes measure the shape and brightness of hot spots or of other easily measured target characteristics. There is a greater variety of algorithms for classifying targets. Some of these algorithms segment or compute the silhouette of an object and measure the features of this silhouette. Others compute the correlation function between an object and each member of a stored library of targets and use this correlation as the basis for classification. Classification algorithms also search for specific features that characterize the targets, such as guns or wheels.

Algorithms for detecting and classifying targets are often combined in an ATR system. For example, a classifier algorithm that is too complex to be applied to every part of a scene may be combined with a detection algorithm that defines a region of interest for it. Alternatively, a classification algorithm may be used to make the detection process more reliable. More detailed descriptions of ATR algorithms and algorithm suites are in Sec. 4.6. The measurements that are used to characterize their performance are discussed in the remainder of this section.

4.3.1   Target Detection Performance Measures

These definitions are general and must be applied in terms of specific test scenarios. For example, the false alarm rate is often reported as the number of false alarms per frame of imagery. Normalization by the number of false alarm opportunities permits comparison of results from different tests. The number of false alarm opportunities is defined as the number of times a target fits without overlapping in the image frame. To compute this, the target is often approximated as a rectangle.

Similarly, the precise nature of a given detection or false alarm must be defined for specific test conditions. *Ground truth,* in which the sizes and locations of targets are specified, can be used for this purpose. The number of detection errors is then computed by comparing the record of target reports in each frame with the ground truth imagery. The inference of pixels as target

pixels requires that the ground truth include silhouettes of each target as they would appear at the appropriate range and aspect. The set of pixels within the silhouette is then the target region and the silhouette itself constitutes the truthed target. *Area-based detection* identifies detection reports within each target region with the silhouetted target. *Norm-based criteria* identify each detection report within some distance of the centroid of the region with the target. A *false alarm* is then a detection that could not be identified with any target using these criteria.

Frequently used performance measures that characterize the detection process are probability of detection P_D, probability of a false alarm P_{FA}, and probability of multiple detections P_{MD}. These statistical performance measures are usually estimated from experimental data using the relative frequency definitions of probability:

$$P_D = \frac{\text{number of target detections}}{\text{number of actual targets}} \; ,$$

$$P_{FA} = \frac{\text{number of false alarms}}{\text{number of false alarm opportunities}} \; ,$$

$$P_{MD} = \frac{\text{number of targets detected more than once}}{\text{number of actual targets}} \; .$$

The measurements that are used in the detection process can, in principle, be combined to form a single number known as a *sufficient statistic s*. This is a quantity whose ability to discriminate between target and nontarget regions is optimal in some sense. The distributions of the sufficient statistic for targets and backgrounds must be significantly different if they are to be discriminated. Typical probability density functions of s for two targets, T_1 and T_2, and the background T_B are shown in Fig. 4.1. The vertical dashed line in this figure is a decision threshold. The *decision threshold* is defined such that a target is declared if s is greater than the threshold, and a nontarget is declared if s is below the threshold.

Two types of detection errors can occur. A type 1 error is a missed target. A type 1 error occurs when a target has a value of s that is below threshold. A type 2 error is a false alarm and occurs when a nontarget area has a value of s above the threshold. The Bayes threshold is set at the value of the sufficient statistic that minimizes both kinds of error given *a priori* probabilities of the events. The Neyman-Pearson threshold is set to maximize the detection probability consistent with a selected value of type 2 errors. This value is the integral of the nontarget probability density function of s, beginning at the threshold, so that the Neyman-Pearson threshold corresponds to a preset false alarm rate.

Trade-offs of performance measures are frequently used to design and evaluate the detection process and to optimize the performance of ATR systems. The trade-off between the probability of detection P_D and the probability of false alarm P_{FA} is expressed in an ATR operating curve. This is a graph of the cumulative probability distributions for P_D and P_{FA} for increasing values of
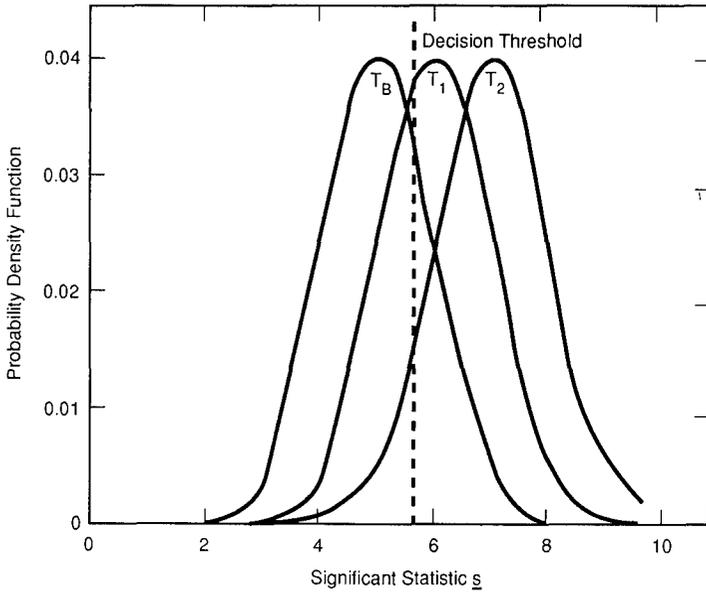
Fig. 4.1 Probability density functions of a sufficient statistic s for a background distribution T_B, and two possible target distributions, T_1 and T_2.

the threshold. Figure 4.2 is an example of two ATR operating curves. Each point on the curves in this figure corresponds to the fraction of detections and false alarms expected for all values of s at or above a threshold value. The regions of the curves for the lowest false alarm rates correspond to the highest thresholds. The slope of the operating curves is highest here because most detections in this region represent targets. The slope of the operating curves is lowest for values P_D and P_{FA} near 1. Lowering the threshold in this region increases the percentage of false alarms without significantly increasing the detection rate. Because the ATR operating curve is based on the two probabilities P_D and P_{FA}, all operating curves begin at the origin and end where the values of P_D and P_{FA} are equal to 1.

The two operating curves in Fig. 4.2 are based on the distributions of s in Fig. 4.1. The upper curve uses the distribution T_2 and the lower curve uses the distribution T_1. The detector represented by the upper curve is better than that represented by the lower curve because P_D is always higher for a given P_{FA}. This is because the separation between T_B and T_2 is greater than that between T_B and T_1. Figure 4.2 illustrates the comparison by defining regions where performance of one detector is better than the other. The operating curves also make possible performance tradeoffs between P_D and P_{FA} for individual detectors.
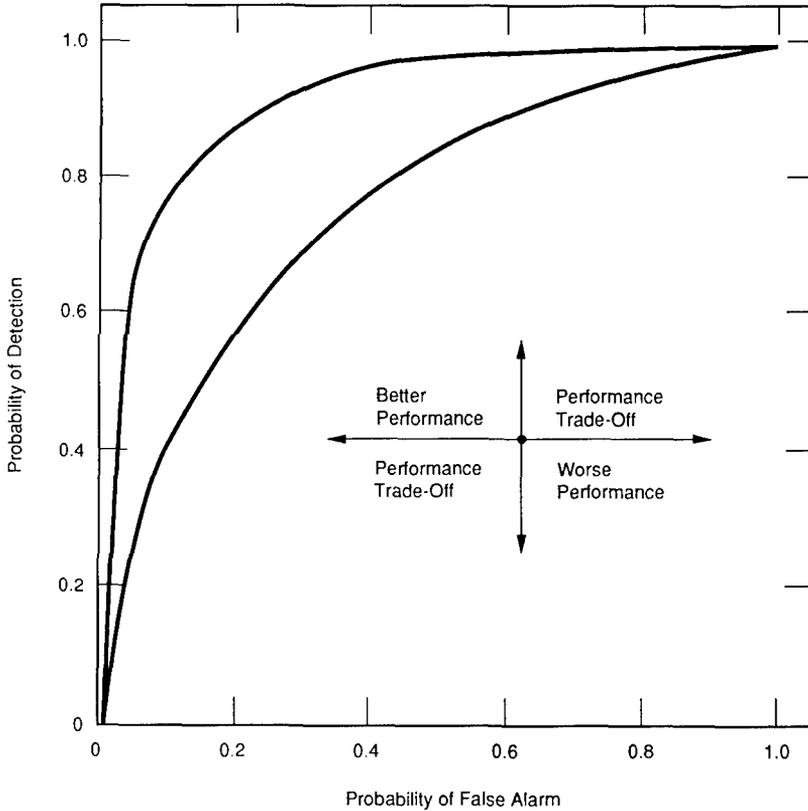
Fig. 4.2 Typical ATR operating curves.

4.3.2 Target Classification Performance Measures

Regardless of the level of discrimination performance, algorithms for target classification distinguish the target from other man-made objects and from the natural clutter in the scene. Performance results for classifiers are often reported in what is known as a *confusion matrix,* sometimes called a *discrimination event matrix.* A set of targets is classified and results are presented in this matrix form, which shows the percentage of times each possible target classification occurred for each truth class, as shown in Fig. 4.3. Confusion matrices are useful for comparing different classifiers if the measurements for each are made under the same conditions and with the same targets. Results are not easily extrapolated to different scenarios or even to different target sets because measurements are not made of the individual features that the classifier uses. Such measurements can serve as the basis for models of the confusion matrix performance measures, however. Feature measures are discussed in greater detail in Sec 4.6.2 under the heading Statistical Classifiers.

ATR Report

	M-60	M-1	M-113	M-2	M-35	HUMMV	M-730	M-163
M-60	54	7	16	13	4	2	2	2
M-1	11	45	17	18	16	0	1	2
M-113	1	16	47	17	7	1	1	10
M-2	6	6	25	48	1	1	2	11
M-35	5	12	8	17	52	1	1	4
HUMMV	1	22	25	23	9	9	0	11
M-730	5	7	8	14	4	0	54	1
M-163	3	9	30	24	3	2	1	28

Ground Truth (row label)

Fig. 4.3 Confusion matrix for an eight-class discrimination problem. The numbers in the matrix are numbers of the reports of each target type for a given vehicle when each target type has 100 samples.

4.4 VARIABLES AFFECTING ATR PERFORMANCE

The typical targeting scenario of an ATR includes different target arrangements and conditions as well as a variety of imaging geometries, intervening atmospheres, and background environments. In most cases, ATR subsystems are trained using imagery collected with the desired imaging sensor and scenario conditions. The question that naturally arises is how robust is the performance of the ATR for all scenario conditions of interest. There are two issues to be examined for answering this question. They involve (1) representativeness and (2) sufficiency. Representativeness is related to how well a set of training imagery or test flights represents the true mission characteristics of interest. Sufficiency is the question of whether or not the sample size is adequate to perform the appropriate hypothesis testing. To determine representativeness and sufficiency of data sets for training or testing an ATR, one must first define the input characteristics that affect ATR performance.

The overall chain of events in an operational ATR is depicted as a series of information transfer steps in Fig. 4.4. A complete ATR system information transfer process includes the scenes, the atmosphere, the sensor, the ATR itself, and subsequent trackers and weapon hand-off components. Table 4.1 summarizes the important parameters affecting the performance of an operational ATR system.
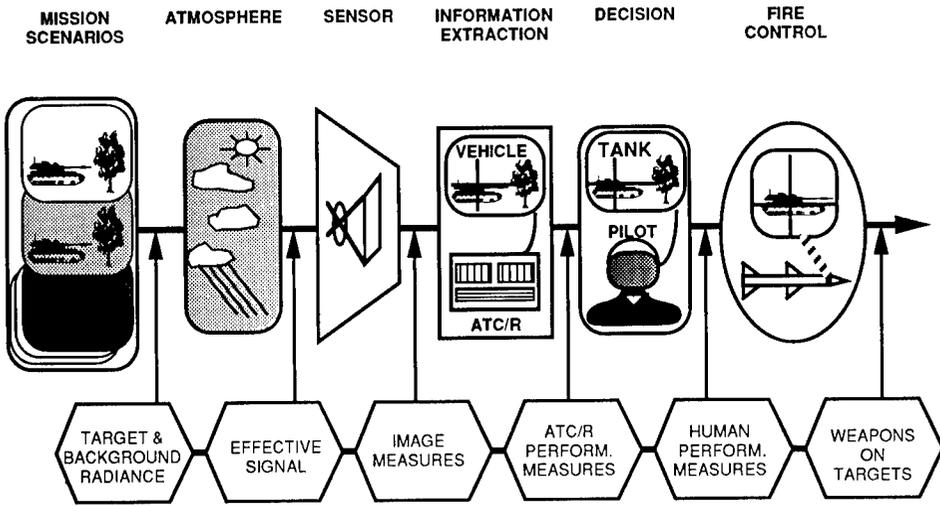
Fig. 4.4   ATR system information transfer process.

Table 4.1   Input Parameters Affecting ATR Performance

Background	Atmospherics
Location	Air temperature
General description	Relative humidity
Context present (roads, treelines, etc.)	Dewpoint
Surface types	Aerosols
Surface states (incl. moisture content)	Barometric pressure
Surface temperatures	Precipitation
Vegetation types	Wind velocity
Vegetation states (incl. moisture content)	Visibility
Man-made objects (non-target, incl. decoys)	Solar radiation
Sunrise/sunset times	Atmospheric transmission
Time of day	Cloud cover
	Sun angle
Targets	
Types	**Sensor**
History of operating conditions	Type
Surface temperatures	Scan procedure
Aspects	Gain, bias, and dynamic range settings
Formations	Drift rate
Camouflage	MRT
	MTF
Platform Dynamics	Sensor artifacts (incl. distortions, dropouts, etc.)
Pointing angles	Field of view/field of regard
Line of sight dynamics	Instantaneous field-of-view
Platform vibrations	

4.4.1 Scene Parameters

The important parameters for an ATR engagement scenario are summarized in Fig. 4.5. They include the predominant factors related to the targets and backgrounds. For the targets, they may include (1) surface structure; (2) thermal mass and history; (3) exhaust trails; (4) physical geometry; (5) surface interfaces, including texture, precipitation, and camouflage; and (6) target motion parameters. For the backgrounds, the parameters are mission specific. For ground backgrounds there are (1) surface types and the physical geometry; (2) vegetation types; and (3) the distribution of man-made clutter objects such as buildings, civilian vehicles, and target decoys.

Emissive and reflective components are both significant factors determining the appearance of the target and the background in the 3- 5-μm spectral band. In the 8- to 14-μm spectral band, the reflective component is not so pronounced, but there can be significant reflection of hot target portions off very smooth surfaces, such as standing water nearby. For imaging radar sensors, the arrangement of radar frequency scatterers, such as corners and edges, are significant, and motion may be significant if moving target indications are processed by the sensor.
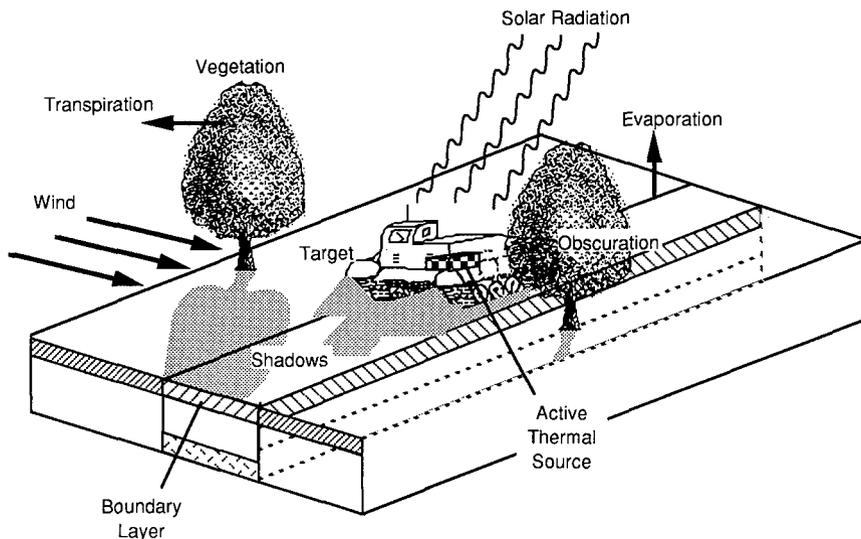
Fig. 4.5 Scene parameters of interest including target and background.
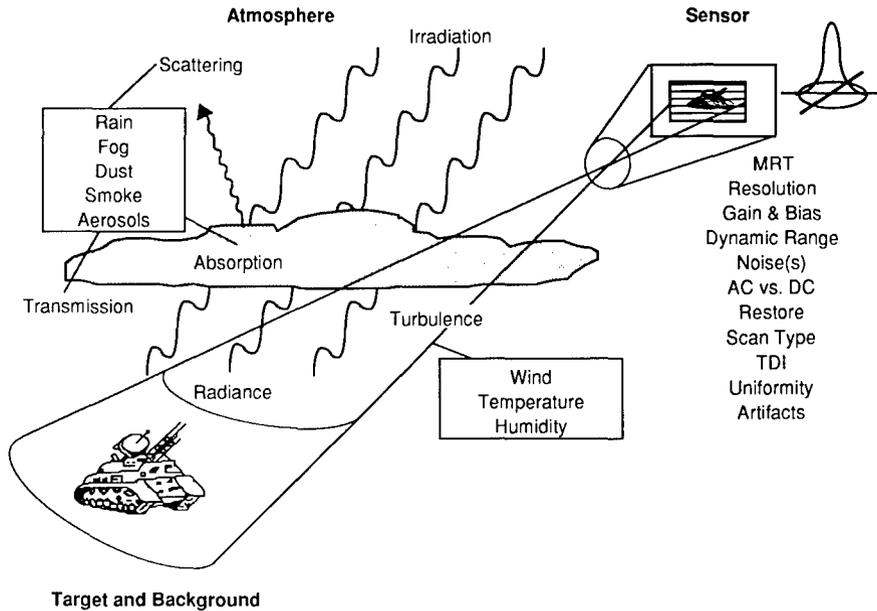
Fig. 4.6 Atmosphere and infrared sensor transfer components.

4.4.2   Atmospheric Parameters

Atmospheric parameters affecting the infrared spectrum are well documented in other volumes of this Handbook. They include an absorption component, a reflection or scattering component, and a distortion component. As indicated in Fig. 4.6, the primary absorption and scattering parameters are related to the contents of the air: (1) molecular composition, (2) particulate matter, and (3) water vapor and droplets. Distortion parameters are usually associated with turbulence in the intervening path or to significant nonuniformities in composition or temperature of the air within the field of view of the sensor. Reflection and scattering of solar energy occurs among the components of the air in the intervening path between scene and sensor, as well as from the scene objects themselves.

4.4.3   Sensor Parameters and Platform Dynamics

A sensor transfer model must take the spectral energy available at the input lens of the sensor and translate that to effective or apparent energy in terms of image intensity produced by the optical, mechanical, detector, and electronics subcomponents of the sensor. Each of these sensor subcomponents contributes to the transfer, in terms of optical spread, distortions, nonlinearities in mapping, or thermal and electronic noise, as well as the sensor artifacts associated with scanning and frame composition. Sensor transfer models and sensor artifacts are also well documented in other volumes of this Handbook. Figure 4.6 also illustrates some of the more pertinent parameters describing the sensor component.

Platform dynamics and sensor slewing controls have a significant impact on the imagery generated by the sensor. Vibration of the platform, sensor angular slewing for coverage of more field of regard than simply the field of view, and fast translation of the sensor with regard to the imaged scene can produce blurring, time delay and integration errors, or distortion in the imagery. It is important to establish and characterize these sources of image degradation prior to ATR system specification, data collection exercises, or system deployments and operations.

4.4.4   Image Measures Related to ATR

A useful parameter describing the sensor component of an ATR system is the SNR for the individual image pixels or for groups of pixels in an image. Minimum resolvable temperature, resolution, gain and bias, dynamic range, electronic noise, and other sensor system artifacts are useful descriptions for sensor quality; however, their effects on ATR performance have not yet been effectively modeled. System performance models are often used to specify infrared imaging system parameters such as sensitivity, aperture size, and magnification. These models are usually based on the Johnson criteria[3] for visual detection and recognition, developed out of early experimental investigations of human observer visual target recognition performance.

A more complex but quantitative approach to the characterization of ATR system inputs involves the description of an image set by the measured properties of the imagery itself. These measurements of the image properties are called *image metrics*. Image metrics are based on the gray-level intensity values and the spatial and temporal information available in an image or sequence of images and serve to parameterize the input to the ATR in a given scenario. It is possible to map the ATR scenario parameters described in Table 4.1 into image metrics. The image metrics can, in turn, be mapped into useful measures of ATR performance.

The motivation for the development and application of image metrics is based on their anticipated properties: (1) fewer image metrics than scenario parameters are required to characterize ATR performance, (2) image metrics are more directly related to ATR algorithms, (3) image metrics will span all potential scenario domains, (4) image metrics provide insight into ATR development and evaluation, and (5) image metrics can be used to model ATR performance.

The selection of image metrics[4,5] is also motivated by several other factors. These include the ability of the metric to characterize (1) its significance relative to ATR performance, (2) its monotonic relation to ATR difficulty in detecting targets, (3) its simplicity as opposed to complexity or peculiarity, and (4) its ability to relate more to the information content of target signatures as opposed to a specific ATR algorithm's feature extraction technique.[6]

There are four basic types of image metrics: (1) global image metrics, (2) global target prominence (GTP) metrics, (3) target-specific metrics, and (4) target-to-local-background image metrics. Global image metrics are mathematical descriptors for the entire image, such as mean or standard deviation of the image pixel intensity distribution. Global target prominence is a characterization of the entire image relative to a computation for one or more subregions of the

image (e.g., target subregions). Target-specific metrics are absolute measures for target-defined subregions. Target-to-local-background metrics are comparative measures for target- and proximity-area-defined subregions of the image. There may be other metric categories, but the preceding four have thus far served as an adequate set of metric categories. Image metrics are summarized in Table 4.2.

Global image metrics serve to characterize the structure and information content of the image as a whole. These metrics include everything from simple mean and variance of pixel intensity distribution to complex morphological descriptors for two-dimensional shape variations and fractal decomposition vectors. The global target prominence metrics are typically employed to determine what percentage of the entire image looks like the targets of interest

Table 4.2   Examples of Image Metrics Currently Used for FLIR Target, Background, and Clutter Descriptions

Global
Image intensity mean
Image intensity standard deviation
Image edge strength mean
Image edge strength standard deviation
Image distribution measures--nonparametric
Image average contour length
Texture classes
Spatial frequency content
Morphological structure classes

Global Target Prominence (GTP)
GTP of Target/Background Contrast
GTP of Target/Background Interference
GTP of each measure in target-to-background list

Target to Background
Target/Background Contrast (ratio)
Target/Background Interference (ratio)
Target/Background Average Edge Strength
Target/Background Average Contour Length (ratio)
Target-Background Entropy (difference)
Parametric test statistics for difference between T & B distributions
Non-parametric test statistics for difference in T & B distributions
T & B Texture class separation
T & B Spatial frequency content ratios or differences

Target
Number of pixels or resolution cells on target
Target length to width ratio
Target intensity mean
Target intensity standard deviation and higher order moments
Target edge strength mean
Target edge strength standard deviation and higher moments
Target intensity distribution measures--nonparametric
Target average contour length
Target convexity measures
Target spatial frequency content

with regard to some target or target-to-local-background metric. In this sense, they are a characterization for assessing potential of an ATR to produce false alarms resulting from similarity with target character. An emphasis is usually placed on the word *global,* in the sense that the image metric is generalized for a set of more than one image.

Target-specific and target-to-local-background metrics are expected to be correlated with the detectability of targets and the ability to discriminate targets from natural clutter in the scenes. Target-to-local-background metrics relate to the ATR's ability to initially detect the vehicles as regions of interest as well as its ability to extract targets from the background in the segmentation process. Target-specific metrics address the features of the vehicle that set it apart from other objects in the scene. Target-specific metrics can be used as features to discriminate between classes of vehicles.

4.4.5   Ancillary Data Sources

Ancillary data sources enable the ATR to integrate useful information into the decision-making process. They include mission status and timeliness and map information for registering the imagery to the ground plane or other points of reference. They also include intelligence on the types of targets to expect and on the likelihood of target visibility. Ancillary data sources can simply be the platform dynamics and timing information. Operator inputs for the prioritization of target types and decision threshold levels are often the most important ancillary inputs.

Information derived from a second sensor can be used to confirm single-sensor ATR decisions. Information from multiple sensors can also be used to support ATR decision processes at a target feature level. Sensor and data fusion is discussed in detail elsewhere in this Handbook.

4.5   ATR SYSTEMS ANALYSIS

4.5.1   Systems Analysis Methodology

To apply the systems analysis methodology to any system one must conceptually or analytically partition the system into a set of functional components. This is typically done using a hierarchical structure known as a *tree*. The analysis and synthesis of individual components is then carried out independently. Higher levels of the system are then analyzed or synthesized according to the tree structure. This facilitates an evaluation of the system performance and characteristics and sets the requirements for the specification of the overall system.

An ATR subjected to a systems analysis can be related to an overall weapons system in two ways. First, an ATR can be viewed as a component or a subsystem of some greater and larger weapons system. Second, the ATR itself can be viewed as a system composed of a variety of subsystems. Both of these viewpoints are used in the following discussion.

A simple example of a weapons system containing ATR would be an aircraft with an ATR that detects, identifies, and cues targets for the pilot or weapons officer. In analyzing the performance of this system, one would consider the
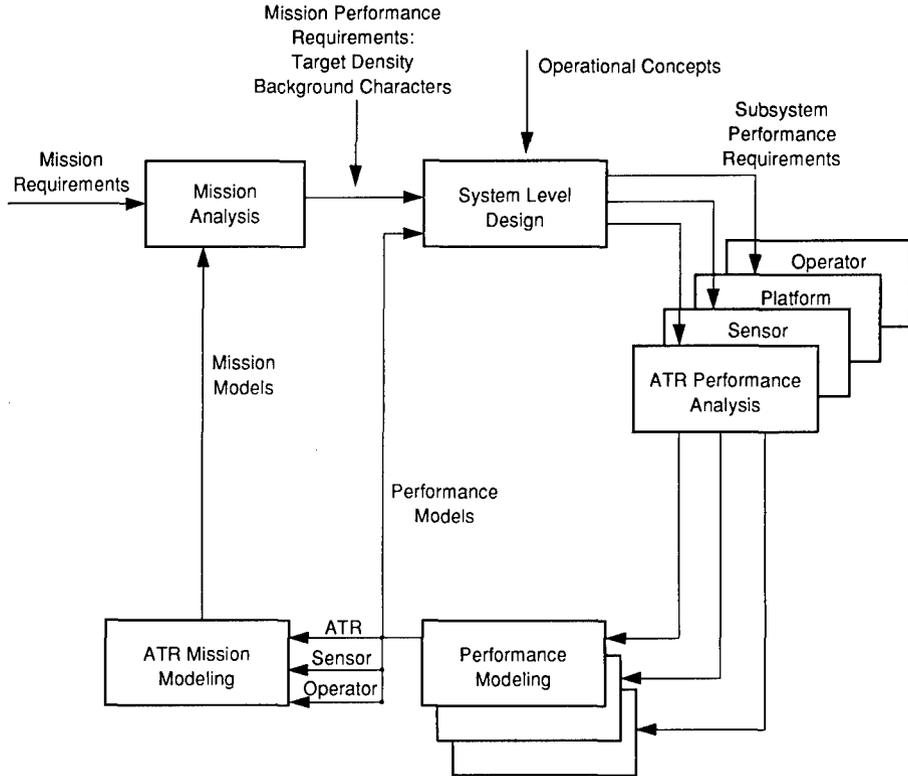
Fig. 4.7 System performance analysis.

mission that is to be carried out, the environment in which it is to be carried out, the performance of the platform, the weapons carried by the aircraft, and the performance characteristics of the ATR. These factors are then considered together to determine the performance of the overall system for the given mission. Figure 4.7 illustrates how the performance analysis of the system can be carried out. Five steps are shown in the figure: (1) mission analysis, (2) system-level design, (3) ATR performance analysis, (4) performance modeling, and (5) ATR mission modeling. Given the mission requirements, the mission analysis determines the mission performance requirements for a variety of situations in terms of target density, the character of the background, the character of the targets, and the character of all the components of the system. When the analysis is combined with an operational concept, a system-level design can be carried out to determine the performance requirements for each component of the overall system. The ATR components or subsystems are then analyzed in terms of their performance given the constraints placed on them by the mission and the system-level design based on the particular operational concept. Results of these analyses are then used to develop performance models for each component or subsystem of the ATR. Performance models are used to develop mission-level models given models for each subsystem: the sensors, the humans involved, the ATR, the weapons, and the platform. Thus, the subsystem analysis process provides feedback to the mission analysis process

concerning how well this mission can be carried out. There is also feedback from the performance modeling process to the system-level design as to whether the system-level design that has been created can be carried out with the components and subsystems specified. Systems analysis is widely and effectively employed in many system developments. The difficulty in applying it to ATR systems has been the lack of validated models of the automatic target recognition performance over a variety of mission conditions and system level designs so that higher level trade-offs can be performed. ATR modeling has been limited to date because ATR is a complex emerging technology for which the theory and empirical base of data concerning stimulus and response is still inadequate. Such data are needed to model the effects of target and background variability in realistically complex situations.

The analysis of ATRs as systems is thus of great interest. Figure 4.8 illustrates the steps involved currently in the performance analysis of ATRs as systems, which overlap with and support the processes of Fig. 4.7. Generally, an experimental method is used because the underlying environment, background, and target characteristics are sufficiently complex that statistical models on which to build a completely analytical analysis do not exist. The figure shows an experimental method involving six steps. They are (1) an analysis of the requirements, (2) an experimental design, (3) a design of data collection, (4) data collection, (5) experiment performance based on the data, and (6) analysis of performance based on that experiment. A higher level systems analysis gives the performance requirements for the ATR as a system in terms of performance criteria for the ATR (for example, the area to be covered and the required performance probabilities) and the various system parameters that
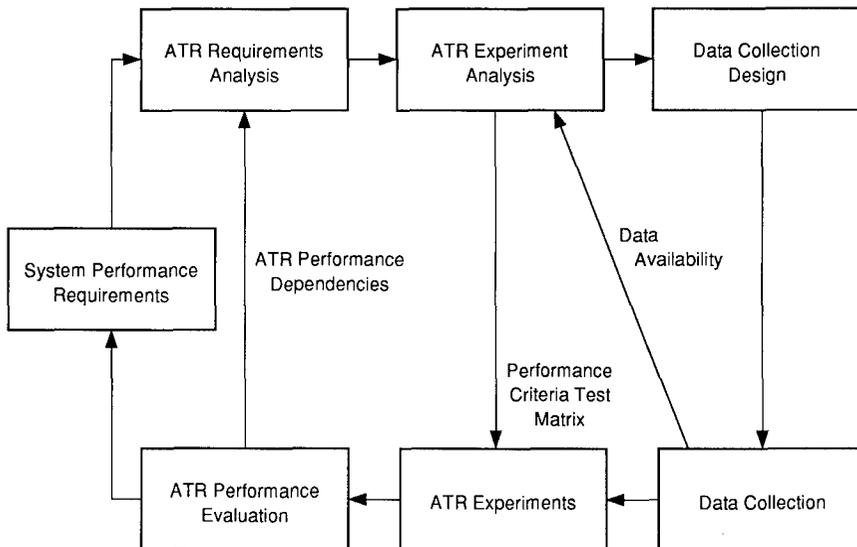
Fig. 4.8   ATR performance analysis.

confine the ATR as a system. These requirements must be analyzed in terms of the ATR components required for carrying out the mission or role of the ATR and their performance goals. Given this information, an experiment can be designed that covers the range of performance and the range of conditions that must be represented in the experiment to adequately describe the ATR for the mission and operational concept being considered in the overall system design. This experimental design is translated into a data collection design, which will depend on the experimental assets available and the range of environmental and background conditions that can be experienced at any given site. The design of data collection also depends on what data are already available. The feedback between the actual data collection and the experiment design illustrated in Fig. 4.8 limits the effort associated with the collection of data. Another output of the experiment design is a performance criteria test matrix that defines the expected performance of the ATR. Experiments can be performed with an ATR to determine what actual performance criteria values can be obtained by various methods. The experiments are evaluated and the performance characteristics of the overall ATR as a system is estimated and used as the input for higher level systems performance analyses. These ATR performance dependencies that also come out of the performance evaluation are fed back to the requirements analysis to understand the levels of performance that are possible with an automatic target recognizer.

4.5.2   Functional Components of Systems Containing ATRs

An ATR is often a component in a larger weapons system. Other components in the weapons system include the ordnance delivery system, the weapons system platform, the sensor, the operator, and/or the navigation and guidance system. The functional decomposition of this larger system depends on both the physical components and how they are to be analyzed. The analysis of the total system and the components that comprise it are interdependent. Figure 4.9 shows a functional flow between the analyses of various subsystems and components. It is expressed in a probabilistic view that is often taken. This view assumes a degree of independence between variables that interrelate the components. The example illustrates nine types of analyses that can be performed based on an operational concept and scenario describing the ATR mission. Feedback processes are indicated by the dashed lines in Fig. 4.9.

The first analysis involves the geometry of the system. Given the sensor platform and its characteristics, the type of sensor, a description of the targets, and the timeliness involved in the assumed operational concept and scenario, a variety of things can be determined. For example, the probability that the target is in the line of sight is a function of various input parameters. The percentage of the targets that would be in the line of sight can be determined as a function of range and other parameters. The time line can also tell us the geometric area that is required to be covered in the given time period. As another example, one can determine the minimum number of systems that is needed to obtain a level of coverage for the targets given the system characteristics.

An analysis of the sensor scan dynamics provides information about the probability that the target will be in the field of view and the sensor charac-
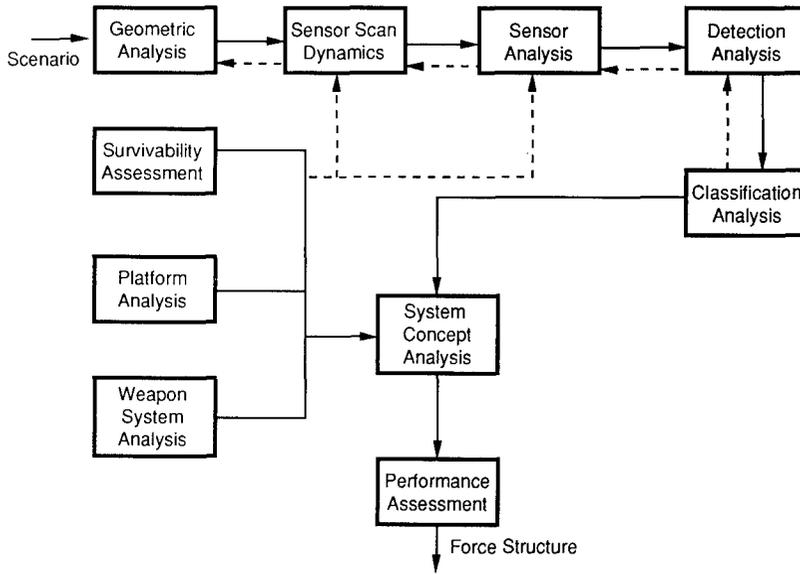
Fig. 4.9   Functional flow for ATR system performance analysis.

teristics of range, instantaneous field of view, resolutions, scan rate, and over-lap that are required to attain the desired performance. The data rate needed to support a given engagement scenario is also an important driving function for the definition of system components. All of the outputs of these analyses depend on a variety of parameters determined by other analyses. For example, they depend on the scan geometry and scan rates obtained from the geometric analysis and on the effective resolution determined from the subsequent analysis of ATR system performance.

An ATR system typically uses an imaging sensor and some form of prepro-cessing to compensate for the sensor artifacts. The analysis of the sensor and the preprocessing typically evaluate the raw characteristics of the image data. The types of threat information required for these analyses include target signatures, decoy signatures, any countermeasure threat characteristics, and the characteristics of the background and clutter. Given required resolution and SNR information derived from subsequent ATR system analysis, factors such as detector sensitivity; data rate; required preprocessing; and the power, size, and volume of the sensor can be determined.

The analysis of the ATR performance is typically broken down into the two generalized performance levels for detecting and classifying targets discussed in Sec. 4.3. A generalized detection analysis typically determines the proba-bility of detecting a target versus labeling something else falsely as a target as a function of a variety of parameters involving the target, the background clutter, and sensor characteristics. The analysis also yields the requirements for the processor in terms of speed, weight, size, and cost versus the input parameters. The generalized analysis for classification, on the other hand,

examines the likelihood of correctly classifying a detection be it a true detection or a false alarm, on further examination of the original imagery or additional imagery where collection is triggered by the detection. This type of analysis also determines the support requirements required to carry out the higher levels of target classification function.

Because the systems we are considering deal with an intelligent adversary, the survivability of the system must be assessed. Survivability depends on the characteristics of the system, and in particular, on those of the platform and the sensor that affect whether or not the system can be observed by an adversary. The speed of the platform and the tactics used to carry out a mission have vital impact on the survivability of the weapon system. Survivability can be analyzed as a function of these various input parameters, and a set of speeds, altitudes, and covertness that maximizes the overall survivability can be determined.

The analysis of the platform from the viewpoint of the ATR typically provides a variety of constraints based on analyses that are in much greater depth in the development of the overall system. These constraints specify the payload available, speed and endurance, altitude of the platform, and the constraints of the various other components that must be carried by the platform. Examples include constraints on the aperture of the sensor; the characteristics of the weapons delivery subsystems dispensing the ordnance load; the inertial navigation system errors; and the size, weight, and power constraints for each of the other components.

The weapons delivery or ordnance dispensing subsystem, if there is one, can be analyzed to determine the effectiveness of the overall system in carrying out its mission. Given that, the ATR identifies, or aids an operator in identifying, a target of interest and hands it off to the weapon subsystem. The success of the weapons delivery subsystem depends on not only the weapon characteristics but also the characteristics of the platform, the sensor, and the ATR. Given the characteristics of the target acquisition process, we can determine how many weapons are needed to kill a single target and what the chances of executing the expected mission are.

Finally, the overall system concept can be analyzed to estimate its performance based on the analyses of all the components of the system. This allows the probability of carrying out the mission to be determined for a variety of system concepts and mission timeliness. This information can be further processed to determine the structure of the force that would be required to carry out a specific description of the mission, and the trade-off of the cost of achieving a kill versus the cost of the system, and thus serves as the basis for deciding the overall usefulness of the concepts considered.

Each of these analyses that has been described can be applied to define the system and its performance or it can be applied in reverse to define the requirements for each of the subsystems.

4.5.3 Functional Components of ATRs as Systems

Figure 4.10 illustrates the input/output relationships between the functional components of an ATR for an imaging sensor. Each component operates on the output of the previous components to localize and recognize targets in sensor
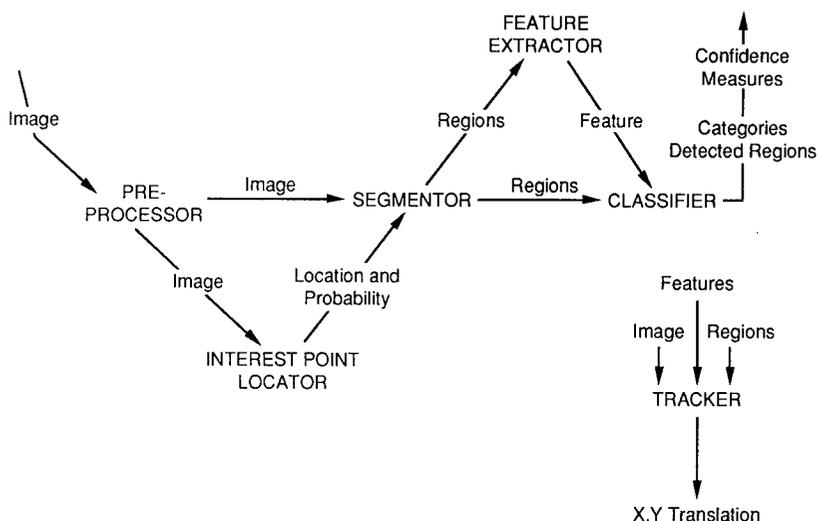
Fig. 4.10   ATR component structure input/output diagram.

output imagery. The component definitions used in this figure are (1) the pre-
processor, (2) the interest point locator, (3) the segmentor, (4) the feature ex-
tractor, (5) the classifier, and (6) the tracker. *Preprocessor* describes any pro-
cedure that enhances the utility of the raw image data. *Interest point locator*
describes a process that examines an image for specific object attributes and
reports interest points in the image. *Segmentor* refers to a process that par-
titions an image into regions, such as tank, wheel, tree, sky, forest, house,
roof, airfield, hangar, or road. *Feature extractor* refers to a process that deter-
mines specific attributes of image regions. *Classifier* refers to the process that
assigns categories with associated confidence measures to image regions, and
tracker refers to a process of using information from a sequence of images that
outputs the current location of objects in an image.

Figure 4.10 also shows the types of input and output of each process and
how the processes are interrelated. This structure fits straightforwardly into
the system analysis methodology and it covers most of the algorithmic concepts
described in the next section. Some ATR components are hard to confine to the
structure of Figure 4.10. A prime example is the model-based vision paradigm,
which controls low-level image processing based on feedback of knowledge
gained by a higher level of reasoning processing. Although the increased com-
plexity makes analysis of a system more difficult, the corresponding increase
in complexity of target and background model or description may be more
realistic.

4.6 ATR ALGORITHMIC CONCEPTS

4.6.1 Target Representation Schemes

The most common target representation schemes employed by ATR algorithms
include statistical, syntactic, relational, projective geometry, and sensor-physics-

based representations. Figure 4.11 illustrates these concepts. There is increasing structure in knowledge representation in the progression from statistical target models to geometrical and sensor physics models. There is more decision-making information available regarding the orientation, type, and condition of the targets in going to the more complicated representational schemes. The more complex representation schemes provide more control for algorithm matching schemes, and they offer greater accuracies in the achievable recognition performance. The drawbacks of more complicated representation schemes are that larger amounts of memory are generally required to store the target models and the time required to hypothesize and verify against very specific target models is generally increased.

Statistical models represent the target as a host of features, including abstract image features such as intensity moments or edge content and geometric features such as length-to-width ratio or shape moments. Syntactic models represent the target as a grammar where the parts and subparts of the target are listed in an attribute tree structure, which describes features, such as
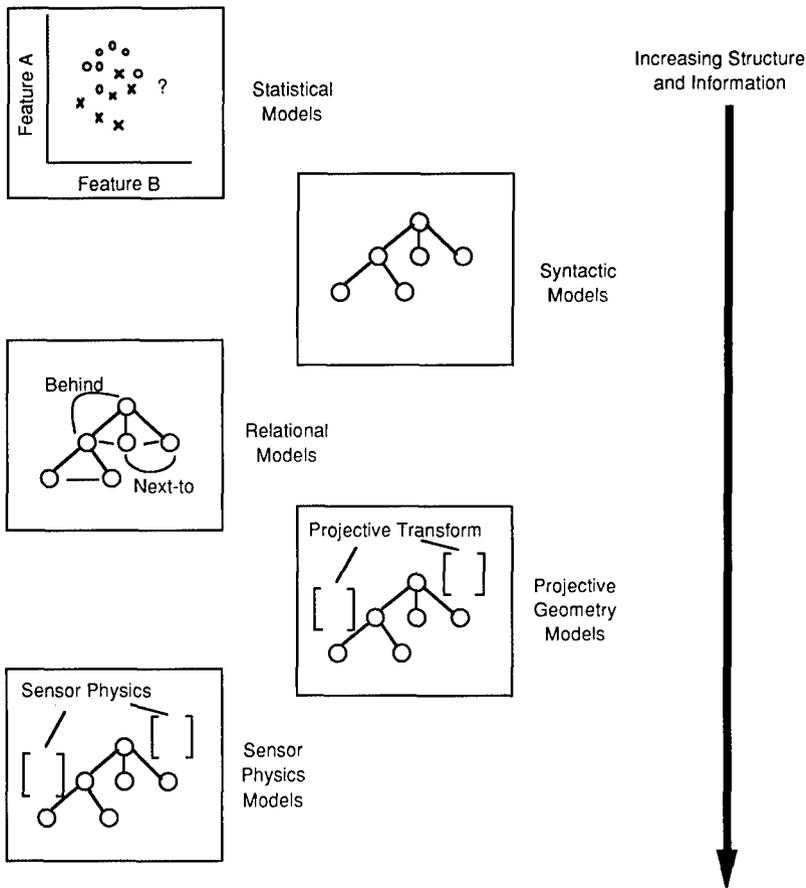
Fig. 4.11   Target representation schemes.

tanks have turrets, tracks, a gun barrel, and an engine; tracked vehicles have treads; and so forth.[7] Relational models consider the individual parts of targets and the rules regarding the relationships, such as direction and approximate distances between these parts. Projective geometry models include the explicit relationship of target parts in three-dimensional wire frame or facetized model representations and they transform projections into one-dimensional or two-dimensional sensor domains, as appropriate. Sensor physics models use information about phenomena in the environment and sensor domains to obtain probabilistic representations for the expected observance of each of the target parts. Sensor physics models can be employed with any of the syntactic, relational, or geometrical representation schemes, adding another dimension of information regarding expected observability.

4.6.2   Statistical Pattern Recognition Algorithms

Many claim that the source of current automatic target recognition algorithms is statistical pattern recognition. Statistical decision and estimation theories are the underlying basis of all target recognition algorithms. The goal is to extract areas of interest from the signal domain and attempt to match these candidates with some prestored representation of objects or targets of interest. Exactly how this goal is achieved often involves some combination of procedures that include techniques to extract signals from noise and some form of clustering or template matching process for classification. The introduction of noise and uncertainty in both the available target signature and the reference target signature necessitates the use of statistical methods and decision theory applied to the pattern recognition problem. The methods described herein are not exclusive, and in many senses there is much overlap between the various approaches.

Region of Interest Detection.   Each stage of the target recognition process is affected by different attributes of the target and background scene imagery. Clutter objects lead to false alarms in the detection process because their attributes, such as size and brightness, are often similar to those of targets. Clutter can be distinguished through the use of additional features and by careful adjustment of discriminant thresholds. Preprocessing to remove artifacts not associated with targets can also reduce clutter.[8]

   Shape information is often used as a discrimination factor in the detection process. This is done by measuring the brightness within boxlike regions placed at each candidate detection region in the image. Figure 4.12 illustrates the double-box operator that is the basis of many algorithms. The inner box is approximately the size and shape of the expected target. The outer box has the same area as the inner box. The region between the two boxes is a guard ring whose size is chosen so that the inner and outer boxes distinctly characterize a target and its background when a target is centered in the inner box. A statistic such as the mean image intensity is computed in the inner and outer boxes when the double-box operator is centered at each pixel in the image, and a target is declared when these quantities are significantly different in each region. In some cases, estimates of the statistic over the entire image are included in the decision process. The mean is an important statistic in infrared imagery because targets are usually brighter or darker than the
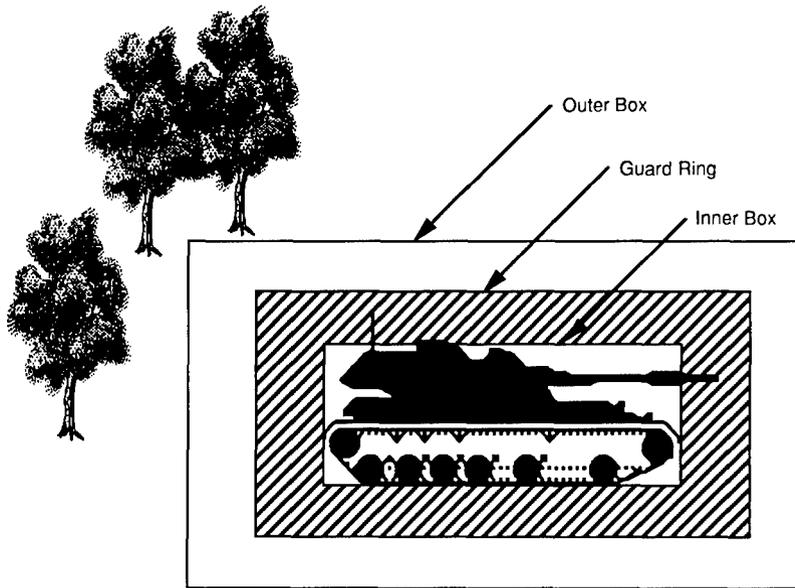
Fig. 4.12   Double-box operator used in many detection algorithms.

surrounding background in infrared imagery. For laser radar imagery, the standard deviation plays a more significant role because the effect of coherent laser speckle produces images with relatively high intensity variations.

The effectiveness of several discriminants has been investigated by Schachter[9] by computing ATR operating curves, using the discriminant as the sufficient statistic. Operating curves were used to evaluate the discriminants because curves that rise more steeply or reach higher probabilities of detection represent more effective discriminants. Three cases are discussed here. In the first, target and background have intensity distributions with unequal means and equal standard deviations. The sufficient statistic s is the quantity

$$s = \frac{\mu_t - \mu_b}{\sigma} \; , \tag{4.1}$$

where μ_t and μ_b are the means of the target and background regions and σ is the common standard deviation. The second case involves target and backgrounds having intensity distributions with unequal means and unequal standard deviations. Here, a sufficient statistic s corresponds to the quantity

$$s = \frac{\mu_t - \mu_b}{(\sigma_t^2 + \sigma_b^2)^{1/2}} \; , \tag{4.2}$$

where σ_t and σ_b are the standard deviations of the target and background regions. Both this case and the preceding case are variations of the statistical

t test for normally distributed intensity. The assumption for the third case is that targets and backgrounds have cumulative distributions that differ in one or more intensity values. In this case, sufficient statistics may be either

$$s = \left\{ \sum_i [P_t(I_i) - P_b(I_i)]^2 \right\}^{1/2}$$

(4.3)

or

$$s = \sum_i |P_t(I_i) - P_b(I_i)| \; ,$$

(4.4)

where $P_t(I)$ is the sample cumulative distribution of intensity up to the intensity I over the target region. This is known as the *Cramer-von Mises test*.

ATR operating curves for five different statistics are illustrated in Fig. 4.13. The curve labeled T is the *t* test with equal variances for target and background with variance calculated from the outer (background) box. Similarly, the curve labeled T_G is a *t* test with equal variances, but the variance is taken to be that of the entire image (global variance). The curve labeled T_U corresponds to the case where a *t* test with unequal variances is applied, and the curves labeled C and C_A are derived from the Cramer-von Mises test. The latter uses absolute values of the differences of the cumulative distributions. All of the curves are generally similar, but the statistic for the *t* test with unequal variances leads to the best curve. This has often been observed to be the case with a variety of other data sets.
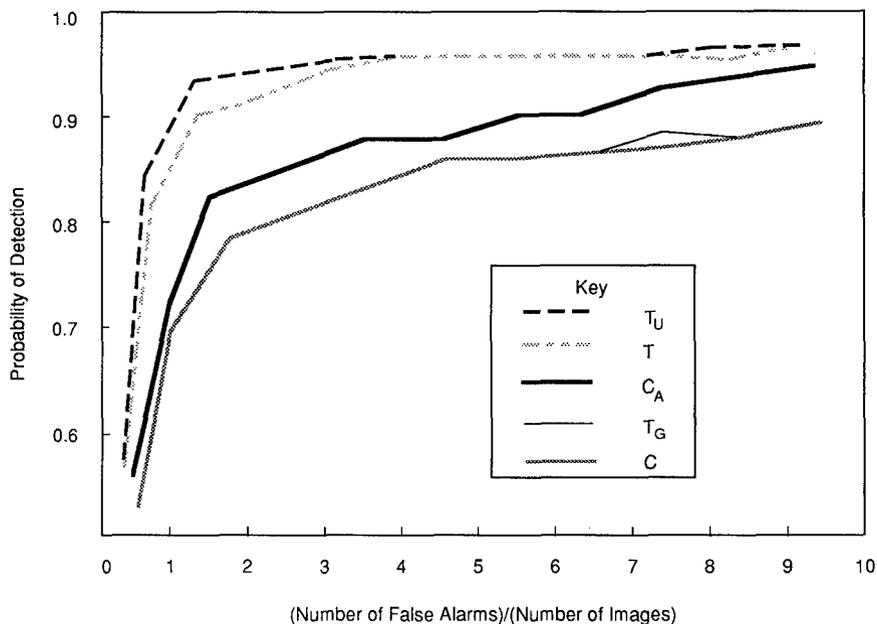
Fig. 4.13 ATR operating curves for five different statistics.

The presence of hot spots can also indicate a target.[9] Hot spots are collections of pixels whose intensities are significantly greater than would be expected from the intensity distribution of the background. If this distribution is known, order statistics can be used to determine how likely the highest measured intensities are. The intensity of the brightest pixel in a window consisting of n pixels is measured. Assuming pixel-to-pixel independence, the probability that the intensity will be greater than I is

$$g(I) = 1 - [1 - P_b(I)]^n \ , \tag{4.5}$$

where $P_b(I)$ is the cumulative distribution of intensities in the background. Similar formulas exist for the second largest, third largest, and other pixels, but an approximation to the probability that m pixels have intensities greater than I is $g^m(I)$. The measurement of m pixels whose intensity is above their probable value is considered a detection.

The expected target size usually enters into the computations in the selection of the inner box. To estimate the size of the target, the range of each image region must be known and assumptions must be made about the orientation of the target. It is sometimes possible to refine the initial estimate of target shape by using an edge detector to find possible borders. These can be located by processing the image with a differentiation operator or other edge detector and locating nearby pixels with the strongest response to this operator. Rates of detection have been improved about 10% by using, instead of a rectangle, an octagon whose sides are placed along brightness gradients that may correspond to edges.[9]

The performance of detection algorithms can often be improved by removing noise in a preprocessing step. Two types of noise that create false alarms can be reduced: (1) hot spots whose size is smaller than expected targets and (2) horizontal streaks associated with some of the early infrared imaging systems. Hot spots can be found by the previously described method. The size of the hot spot in terms of connected pixels assists in identifying it as target or clutter. Streaks can be identified by measuring the averages of rows or pixels and comparing that of each row with that of its neighbors. Rows that are significantly different can be replaced by averages of neighboring rows.

Segmentation Methods.   Segmentation can stabilize information in target images by finding the silhouette of each candidate object. As an example, consider the thermal image of a military vehicle. Its appearance is highly variable and depends on many factors, including operating state of the vehicle, time of day, weather, and other factors that affect the temperature of the vehicle's external surfaces. The computed silhouette is a function of the shape of the vehicle, and its features, especially its area, should be useful in identifying the vehicle. Segmentation algorithms utilize differences in intensity or other characteristics between the vehicle and the background and the degree of homogeneity in the appearance of the vehicle. Segmentation leads to an accurate silhouette when (1) the contrast of a target is high, (2) its boundaries are sharp, and (3) there is no targetlike background clutter near the object.

At least three segmentation methods have been used to find the silhouette of possible targets: (1) edge detection, (2) thresholding, and (3) region growing.
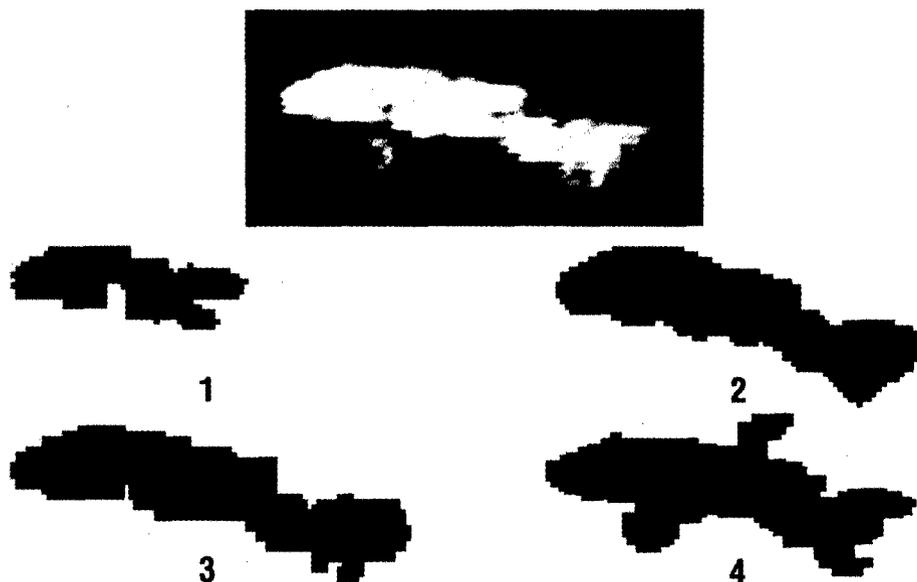
Fig. 4.14   Segmentations of an infrared image of a truck produced by four different algorithms.

Many segmentors use some combination of these three techniques. For example, edges may be found by applying a differentiation operator to the image and the average intensity of the edges used to determine a threshold. Region growing may be used after an edge operator has been applied or to combine regions that are separated spatially by a small distance.

Figure 4.14 is an example of segmentations of the infrared image of a truck produced by four different algorithms.[10] The boundaries of all of the segmentations are very dependent on the local temperature of the vehicle and the background. Computed shapes vary significantly, and numerical descriptors of the shape of the vehicle that are computed from these silhouettes are of limited use in identifying vehicles.

Edge-finding operators are local in nature. One of the most widely used is the Sobel operator,[11,12] which analyzes 3×3 arrays of pixels to find abrupt changes in intensity. The Laplacian of Gaussian operator is also often used because it not only detects edges but preserves hot spots.[13] Errors can be caused by anomalous high-intensity regions. Edge detection can be improved by using morphological operations, such as erosion and/or dilation operations.[14] Like edge detectors, they are also local operations: Erosion makes regions smaller by removing pixels from their edges, while dilation makes them larger by adding pixels. Erosion can remove isolated pixels not associated with edges. Dilation can be used to fill in gaps in edges by extending lines until they meet. Alternatively, gaps can be filled by using the edge operator to detect smaller changes in intensity in the neighborhood of an end of an edge. Recursive morphological filters can be used to fill or remove the larger extraneous artifacts.

Segmentors that are based on thresholding often compute the thresholds from the statistics of the image. This is sometimes done by computing a histogram of intensities and setting the threshold between peaks on the histogram. However, the histograms of many images do not have distinct peaks, and other methods must then be used. As mentioned, thresholds can be set at the intensities of edges that have been detected with an edge operator. Because much imagery is not uniform, local thresholds may be computed from the statistics in regions where detection algorithms have found targets. The application of information derived from the intensity histogram is more likely to be effective in these regions, because detection algorithms are designed to find regions that contain two significantly different intensity distributions.

Region growing is an iterative process in which thresholding is combined with computations based on relationships among pixels to segment an image into two or more regions. At each step of the iteration, a consistency function is computed that depends on the difference between each pixel and its neighbors. The value of the intensity of the pixel is changed in some way that depends on this function so that the intensity is closer to the intensities of its neighbors if the differences were already small. The effect of changing the intensities of pixels, based on the consistency function, is to make large regions larger and small regions smaller at each interaction. The *consistency function*, sometimes called the *cost function,* may depend on the averages or standard deviations of the targets and background regions; in this case, the image is thresholded before iteration so these quantities can be computed. The consistency function may include other information besides intensity differences of neighboring pixels. One function has been constructed[15] that favors the construction of edges and makes smaller changes at later iterations than at earlier ones.

Ideally the region growing process will reach a stable condition at which further applications of the procedure will not change the segmentation. This does not happen for every consistency function, and some test must be applied to stop the process at an appropriate time. One method for doing this, called the *stable structure method,*[16] assumes that the number of separate regions decreases rapidly at each interaction until an optimum segmentation is reached and decreases more slowly after that. The number of separate regions is counted after each iteration and the algorithm is stopped when the rate of change of this number becomes smaller.

Statistical Classifiers. Features that can be measured by the ATR sensor are the basis of target classification. These are measured or computed characteristics of a target. For example, if visible or infrared is used in the classification process, the perceived area of each object's visible or infrared signature in each image is often used as a feature. The perimeter and the lower order moments may also be features. A statistical classifier compares feature measurements with values that are expected for specific vehicle types and chooses the type that most closely matches those measurements. A correlation classifier computes the correlation function between a measured image and members of a library of images of possible targets and chooses the one with the greatest correlation. Correlation classifiers are equivalent to statistical classifiers with features approximately equal to spatial frequency spectral components. In

these cases, the values of individual spatial frequency spectral components are not computed explicitly.

A body of theory has been developed that provides procedures for classifying targets and predicts the accuracy of the processes. Classification accuracy is determined by the number and statistical distribution of the measured features of targets and by uncertainties in the knowledge of the feature distributions.

Detection can be considered a two-class classification problem. The target and background have different distributions of a statistic that is used to label each region as one of the two classes. A threshold, or discriminant surface, determines which class a measurement of the statistic indicates. The classification error rate is the sum of the integral of the part of the background distribution above the threshold and the integral of the part of the target distribution below the threshold.

This procedure can be extended to include many classes and many features. In a multiclass problem, different combinations of features may be the optimal discriminants for each pair of classes, and discriminant surfaces in a multi-dimensional feature space define the combinations of features that best identify each class.

A classifier can be evaluated in terms of the average and standard deviation of each independent feature when the features that are used in the classification process are known. For each class, the standard deviation of features is usually related to the accuracy or consistency of the measurement process and is properly termed the feature noise. When the feature noise is the same for each class of data, as is often the case, a correlation matrix can be defined for the feature set as a whole. The eigenvectors of the correlation matrix are independent feature combinations that can be used for classification. The number of eigenvectors whose eigenvalues are significantly different from zero is the dimensionality of the feature set.

The average value of a feature i in the class j can be characterized by the number μ_{ij}. A standard deviation σ_i of the μ_{ij} characterizes the distribution of class averages of the i'th features. The characterization inherently assumes that each feature of the objects is statistically distributed according to a normal, or Gaussian, distribution. The ability of the feature to distinguish classes can be characterized by the set of ratios σ_i/σ_{ni}, where σ_{ni} represents the standard deviation of the i'th feature within each class. If these ratios have similar values for each feature, the feature set can be characterized by the single approximate value σ/σ_n and the dimensionality of the feature set. This represents the minimal number of quantities that can characterize a feature set; that is, an SNR for each feature and the dimensionality or the number of significant features.

Feature data can be represented as points in a hyperspace whose axes represent the values of individual feature measurements. Figure 4.15 is a representation of a two-dimensional feature space, which may be a subspace of a larger feature space. The individual points represent the class averages of a training set. In the classification process, measurements are compared to these class averages, and the measurement is assigned to the class whose average is closest. The error rate is the probability that an incorrect class is closer to the measurement than the correct one. In Fig. 4.15, a measurement is represented by the open circle and the correct class average at point x_0 a distance
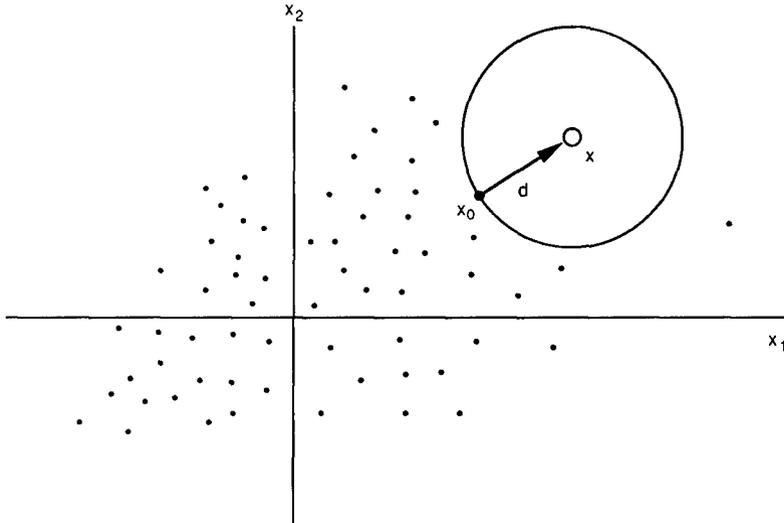
Fig. 4.15 Distribution of class averages in a two-dimensional feature space. The open circle represents a measurement, and d is the distance between the feature measurement and the class average, x_0, for the correct class.

d from the measurement. The error rate is the probability, averaged over all points x that are encountered, that an incorrect class average will be within the hypersphere of radius d centered at the measurement point x and with radius $d = |\bar{x}_0 - \bar{x}|$.

The error rate can be calculated if the probability distribution functions of class averages and of feature noise is known.[17] At each class average x_0 and each feature noise value d, the probability that an incorrect class is within the hypersphere of radius d is calculated. The average of this probability over the distribution of feature noise values and the class averages is the error rates. Figure 4.16 shows a set of approximate curves of error rate versus feature noise made under the assumption that the probability distribution of feature noise is normal with standard deviation σ_n, and the probability distribution of class averages is normal with standard deviation σ. In this example, there are 20 possible classes. Curves are for feature set dimensionalities of 5, 10, and 20.

The classification error that results from overlapping distributions of features is known as the *Bayes error*. If the distributions themselves must be estimated before classification can be done, additional errors result because the classification error is comprised of both the Bayes error and the error in the estimate of the probability distribution of the statistic or discriminant. Estimates of distributions of features are made from data similar to that which is to be classified, making use of any available *a priori* knowledge. There are at least three commonly used estimates of the probability distribution of a discriminant: (1) a parametric estimate, (2) a Parzen estimate, and (3) a nearest neighbor estimate.

A parametric estimate assumes that the functional form of the distribution is known and that only parameters of this form need to be estimated. For example, if it can be assumed that the distribution is Gaussian, the mean and standard deviation need to be estimated. Estimates of the errors of the parameters may also be made and used to determine the uncertainty in the probability distribution of the discriminant.

The Parzen estimate is constructed from a set of training data as the convolution of each data point with an appropriate kernel:

$$P_h(x) = N^{-1} \sum_{i=1}^{N} k_h(x - x_i) \ , \tag{4.6}$$

where N is the number of data points, k_h is a kernel function, and h is a parameter in the kernel. A kernel that is often used is a Gaussian of width h:

$$k_h(x) = (h\sqrt{2\pi})^{-1} \exp(-x^2/2h^2) \ . \tag{4.7}$$

The parameter h is chosen to minimize the deviation of the estimated distribution function from the actual in a least-squares sense.[18] This parameter is larger when the data is sparse.[19] The error in the Parzen distribution can be estimated by a resampling procedure known as the *bootstrap technique,* in which separate estimates are created with data points randomly removed and
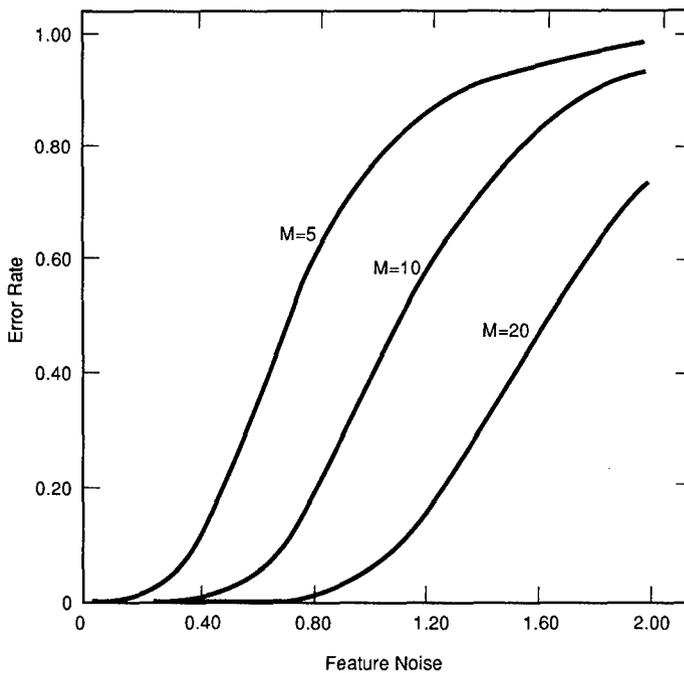
Fig. 4.16 Error rates versus feature noise for a set of 20 classes with 5, 10, and 20 features.

others duplicated so that the total number of data points, including the duplications, is always the same.[20] A collection of separate estimates gives a reasonable indication of the range of values that the distribution function might have.

The nearest neighbor method of classification assigns data points to the same class as the member of the training set that is closest to it in feature space.[21] It is a nonparametric method in that it does not require the explicit definition of a probability density function. However, it does require that a distance measure be defined for the feature space, such that the values of probability density functions of the feature measurement and training set member are nearly the same when the distance between them is small. The nearest neighbor rule has been generalized to the k-nearest neighbor rule by assigning a measurement to the class that is most frequently represented among k nearest neighbors. A lower bound on classification error by the nearest neighbor method can be estimated by classifying members of the training set using its distance to members of the same training set. An upper bound can be estimated by classifying members of a different data set from the training set. Both of these estimates approach each other when the number of points is large.[22] Foley[23] has applied this method of analysis to the two-class problem in which both classes are normally distributed and found that overall classification error approaches the Bayes estimate when ratio N/F is at least 3. Here N is the number of samples per class and F is the number of features. The result, known as *Foley's criterion,* is often used to estimate the sample size required for a nearest neighbor classifier.

Statistical classifiers are sometimes implemented as a decision tree. A data point is assigned on the basis of a single feature to a group of classes, then assigned on the basis of other single features to successfully smaller groups of classes until the most precise classification is reached. The method must be used when features applicable in one branch of a tree cannot be defined for other branches. When this is not the case, the method can still make computations faster. However, the process constrains the region of feature space in which a datum is located to one that can be defined through a series of successive dichotomizations. It may not be as accurate as a method that relies fully on the multidimensional distribution of features.

Moving Target Indication.   Target motion has been used in the detection process by comparing successive frames in sequences of images. To perform this analysis, geometric distortion resulting from changes in the orientation and position of the camera must be rectified. Successive rectified images are compared either to create an average image in which noise is reduced or to create a null image by subtraction, in which the only objects are those that are in motion.

Geometric distortion is measured by matching portions of successive frames with high-spatial-frequency content using the correlation process. As described in the last section, high spatial frequencies are associated with narrow correlation functions, which are needed for accurate positional matching. An optical flow field is computed consisting of vectors that represent the magnitude and direction of the image displacements. The effects of measurements can be reduced by parameterizing the optical flow and computing the parameters that

best fit the data in a least-squares sense. The optical flow field is then described by an affine transformation of the form

$$y_1 = a_{11}x_1 + a_{12}x_2 + a_{13} , \tag{4.8}$$

$$y_2 = a_{21}x_1 + a_{22}x_2 + a_{23} , \tag{4.9}$$

in which y_1 and y_2 are coordinates of a rectified image and x_1 and x_2 are measured coordinates. The parameters a_{ij} are computed to give the smallest sum of squares of error between measured and predicted coordinates (y_1, y_2).

If the platform moves large distances between frames, the effects of parallax must also be included in the rectification. Motion of the foreground is different from that of the background and must be corrected differently. A knowledge of the perspective is necessary for this.

4.6.3 Template Correlation Algorithms

Template matching is the application of very specific filters or templates to image segmentation, feature extraction, or classification processes. It can also be applied for all three of these processes simultaneously. When applied to image segmentation, the matching templates are simple, for example, edge matching templates. For feature extraction, the templates are portions of the target, such as the turret, wheels, or hull; scene features, such as roads and boundaries; or abstracted image features, such as tank-shaped silhouettes, engine-shaped blobs, and building-shaped areas. For classification, there are usually a number of target or scene templates representing a single target or scene under specific conditions and from a single viewing geometry. A template-matching algorithm may accomplish the entire recognition process by translating a number of complete binary or gray-shaded target or scene templates directly over the image, and finding maximum cross-correlation function responses. For any of these processes, the templates may be compared with the signal or image in the original domain or in a transform domain. Sobel, Hough, and Fourier transforms are often applied to this problem.

Matching would be a trivial task if we could expect an exact copy of the template to be found in the image or signal. Because the image or signal contains noise, including clutter that is not multiplicative or additive noise, and because it is infeasible to run all possible conditions of multiple targets over the image, we are forced to produce some measure of match between template and image or signal. In the digital processing domain, matching is usually accomplished by cross-correlation measures. We can measure the degree of match between two functions f and g over a region Ψ using one of the following distance measures[24]:

$$\max_\Psi |f - g|, \quad \text{or} \quad \iint_\Psi |f - g|, \quad \text{or} \quad \iint_\Psi (f - g)^2 . \tag{4.10}$$

If we use the latter expression as a measure of mismatch, we can derive a useful measure of match from it. Note that

$$\int\int (f - g)^2 = \int\int f^2 + \int\int g^2 - 2\int\int fg \; . \qquad (4.11)$$

Thus, if $\int\int f^2$ and $\int\int g^2$ are fixed, the mismatch measure $\int\int (f - g)^2$ is large if and only if $\int\int fg$ is small. In other words we can use $\int\int fg$ as a measure of match. The same conclusion can be reached using the Cauchy-Schwarz inequality, which in the digital case is

$$\sum_i \sum_j f(i,j)g(i,j) \leq \left\{ \sum_i \sum_j [f(i,j)]^2 \sum_i \sum_j [g(i,j)]^2 \right\}^{1/2} \; . \qquad (4.12)$$

If f is a template, g an image, and we want to find pieces of g that match f, we shift f into all possible positions relative to g and compute $\int\int fg$ for each shift $(x + u)$ and $(y + v)$. We then compute the normalized cross-correlation as a measure of match:

$$\frac{\sum_i \sum_j f(i,j)g(i + u, j + v)}{\left\{ \sum_i \sum_j [g(i + u, j + v)]^2 \right\}^{1/2}} \; . \qquad (4.13)$$

This quotient takes on the maximum possible value for displacements (u,v) at which $g = cf$, where c is a constant. Cross correlating f and g is a linear shift-invariant operation, because it is the same as convolving g with f rotated 180 deg.

Optical methods of matched filtering emerged from the field of optical information processing. Optical matched filtering can be accomplished in the spatial domain or the spatial frequency domain. The frequency domain approach is based on the fact that a lens produces a Fourier transform of any input signal. A linear shift-invariant filter is said to be matched to the particular signal $s(x,y)$ if its impulse response $h(x,y)$ is given by

$$h(x,y) = s^*(-x,-y) \; . \qquad (4.14)$$

In a coherent processing geometry, placing a frequency-plane mask between identical input and output lenses provides a simple method for matching an input with its frequency domain representation.[25] The optical arrangement for this operation is depicted in Fig. 4.17. If a filter (or frequency-plane mask) is to be matched to the input signal $s(x,y)$, it should have an amplitude transmittance proportional to S^*. Fourier transforming the impulse response $h(x,y)$ = $s(-x,-y)$ produces

$$H(f_x, f_y) = S^*(f_x, f_y) \; . \qquad (4.15)$$

Placing this filter S^* into the optical path as shown in Fig. 4.17 produces a transmitted field distribution of SS^*. This implies that the filter exactly cancels the curvature of the incident wavefront S, producing a plane wave incident on lens 3. The plane wave is brought into focus by lens 3 to produce a bright
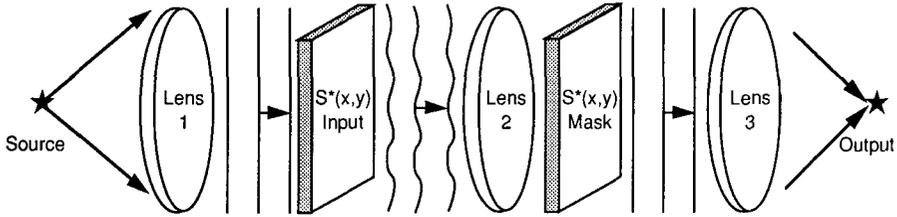
Fig. 4.17   Optical interpretation of the matched filtering operation.

spot on the output plane. When an input signal other than $s(x,y)$ is present, the wavefront curvature in general is not canceled by the filter, and the transmitted waveform is not brought to bright focus by the final transforming lens. Thus, the presence of signal s can be detected by measuring the intensity of light at the focal plane of the final transform lens. Note that if the input signal of interest is not centered at the origin, the bright point in the output plane is simply shifted by a proportional misregistration distance.

Even though the optical frequency-domain matched filter is linear shift-invariant, size and rotation present significant problems in implementing this technique for practical matching of targets in a two-dimensional image. It is possible to cycle a number of filters through the optical transform plane using high-speed digital spatial light modulators, thus removing the need to swap filters mechanically into and out of the optical path. However, the number of filters representing size, rotation, and three-dimensional to two-dimensional projection variations must be minimized to obtain reasonable matching performance while maintaining a practical processing rate for ATR applications.

The advantage of doing the matched filtering process in the spatial frequency domain instead of the spatial domain is that the filter does not need to be convolved by scanning across the entire input image. If a match to a target template occurs in the image, no matter where it is, it will appear as a bright spot in the output plane. The advantage of using the optical methods instead of the digital processing methods is that the speed of Fourier transform computations, even with fast Fourier transform computation algorithms, is still very time consuming for large input images.

Practical use of matched filter techniques must deal with the throughput limitations of only so many allowable filters for the processing time line. The question naturally arises as to how much mismatch is tolerable to maintain detectability of the targets. The degree of mismatch in scale, rotation, and projection must be determined and filters generated at intervals of these variations in order to maintain desired detection performance. Variants of template matching that allow for greater error tolerance are called either *rubber masking* or *spring tension fitting*. In these schemes, relatively few templates are stored along with equations describing the possible distortions, such as changes in scale, rotation, and projection, for each template. These approaches are possible when relatively few object types are being searched for, and some constraints can be levied regarding the expected distortion.

An example of the rubber masking approach would be to assign spring functions to each of several templates that assign increasing spring tension

with increasing change in each of several dimensions of scale, in-plane rotation, and normal-to-plane rotations. To match the templates with portions of the image, we must first find a first-level matching energy such as the cross-correlation functions described earlier, with no distortion for any template. At each of these candidate locations, a least-tension optimization process is pursued wherein the distortion in each dimension is adjusted and a tension measure is made. The path of minimum tension in the springs is pursued until a minimum energy measure is obtained. This procedure is repeated for each starting template and the best match chosen for declaration. This procedure can be computationally intensive. A useful shortcut is to first find good matches for subtemplates, and then try to build up a low-tension combination of these matches stepwise, using heuristic search techniques.

Similarities exist for the template-matching techniques described here for optical and digital approaches as compared with the neural network approaches. Neural network approaches are claimed by some to have the advantage of built-in tensioning functions within the network. Neural network methods and algorithms are discussed further in Sec. 4.6.5.

4.6.4   Model-Based Algorithms

Model-based approaches to ATR algorithms extend the base line of the matched filter and statistical pattern recognition approaches in several ways. Before discussing the differences, however, it is important to point out that model-based approaches, as described by their researchers, include model-based reasoning, model-based recognition, model-based vision, and model-based training. Although there is much overlap between these categories, there are subtle differences. The difference between model-based recognition and model-based vision is simply the source of the information to be extracted. Model-based vision is by name associated with imaging systems and is a subclass of model-based recognition processes that include nonimaging sources of input information. Model-based reasoning, on the other hand, is more comprehensive in that there are models for decision structures as well as for target models. Model-based reasoning could be said to include meta-models, or knowledge about reasoning with target models.

There is a significant difference between model-based vision and model-based training of statistical pattern recognition algorithms. Model-based training is simply the use of modeled target signatures to train a statistical pattern recognition algorithm. In model-based training, a target model is used to generate a synthetic image or signature. Target-like features are then extracted from the image and stored in the algorithms' multidimensional feature space.

Zelnio[26] presents a useful comparison of model-based, matched filter, and statistical pattern recognition algorithms. The key difference between model-based approaches and statistical pattern recognition approaches is the reference space or knowledge base employed for matching. With statistical pattern recognition, there is a set of training images from which targetlike patterns are extracted to serve as features in the multidimensional feature space. These features are usually abstracted features, such as convexity or shape moment, as opposed to the target geometrical and wavelength related phenomenological features, such as wheels, turrets, wings, superstructures, engines, or the hot

exhaust parts used in model-based reasoning. One of the advantages of true model-based approaches is that new targets and environments can be added without having to retrain the algorithms. A new model can be added directly to the knowledge base portion of the model-based algorithm. The difference between model-based algorithms and matched filter processes is the degree of iteration and adaptability in the hypothesis generation and matching process, as is clarified in the following.

Components of Model-Based Algorithms. The primary components of a model-based vision algorithm are shown in Fig. 4.18. The model-based recognition algorithm generally considers the components or processes as listed here. Certain of these processes may well be combined in specific algorithm implementations. These are the processes for (1) control mechanisms, (2) planning mechanisms, (3) hypothesis generation, (4) knowledge storage and knowledge retrieval, (5) knowledge extraction, (6) inference mechanisms, (7) uncertainty management, and (8) conflict resolution.

Control mechanisms are the processes that act as the manager for controlling the whole process of recognition by reasoning. Planning mechanisms optimize the recognition process and provide flexibility in dealing with different situations. Hypothesis generation is the process that predicts objects, parts, and features in the scene to be detected and verified. Knowledge storage or knowledge retrieval is the process that stores model knowledge. Knowledge extraction is the process that controls the information extraction from one or
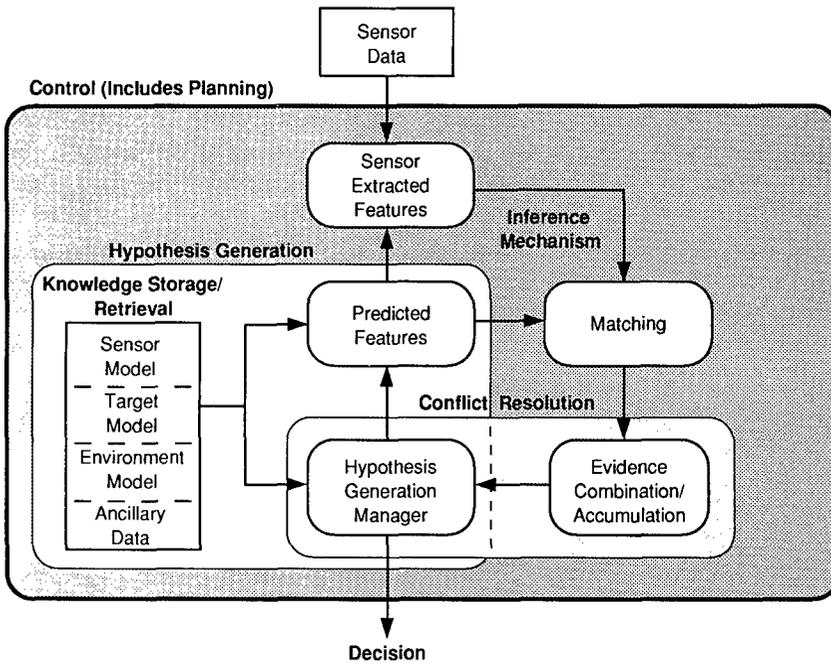
Fig. 4.18 Model-based vision algorithm structure.

more imaging or nonimaging sources. Inference mechanisms match predicted features against extracted features, and uncertainty management handles uncertain, unclear, or incomplete evidence. Finally, conflict resolution is the process that handles evidence that supports conflicting hypotheses.

The knowledge storage or retrieval processes contain reference models and source information regarding targets of interest and how or where these targets are deployed. The sensor extraction process is essentially the signal or image processing procedures used to find objects of interest. The inference or matching process analyzes candidate image objects and matches them with candidate model objects, at any of the various levels of target representation discussed previously. Given a large number of models, the hypothesis generation process includes an indexing procedure to determine which models should be compared with the candidate extracted image region or features. The control process is required to initiate the processing, search the references, maintain intermediate accumulations of evidence for several competing matches, resolve conflicts, and implement a decision threshold procedure for determining when to stop the iteration process and report decisions.

Model-based algorithms reason by generating hypotheses and attempting to verify candidate hypotheses. A top-level hypothesis, perhaps based on ancillary intelligence information, starts the feature and evidence extraction without a specific object hypothesis. Normally a few initial extracted features when combined with evidence will trigger the generation of an object hypothesis. In a top-down manner, the hypothesis manager generates subcomponent or secondary hypotheses from the object hypothesis, and they in turn predict certain features in the image. Then at all levels of abstraction, the inference mechanism attempts to verify these predicted features by matching them against subsequently extracted image features. This inferential evidence from a verified subcomponent hypothesis is used to update and modify the object hypothesis and then either start another round of subcomponent hypotheses and verification or declare a final verified object hypothesis. The search for another object in the image begins after an object is declared.

There are numerous levels of complexity at which to implement the model-based recognition approach. Less robust implementations of model-based vision use only a single pass at image extraction and matching with a host of target templates, to select the best match. This level of implementation is basically the matched filter approach mentioned earlier. Model-based reasoning generally includes execution of more than one pass of the image extraction, hypothesis generation, and matching processes, in an iterative optimization cycle. The power of this iterative reasoning process is similar to the game of asking 20 questions. Each question or hypothesis is progressively more directed and detailed until a very good match is obtained. The initial target candidate image subregions are extracted based on a series of coarse or invariant filters on the first pass, and through a series of iterations of signal extraction at different thresholds or with finer filters, and matching, evidence is accumulated to reach a high-confidence decision on the presence of targets and nontargets and on target type. It is also possible to employ time sequences of signals or images as inputs to obtain better signal to noise or different looks of the same candidate target region. The process is illustrated in Fig. 4.19.
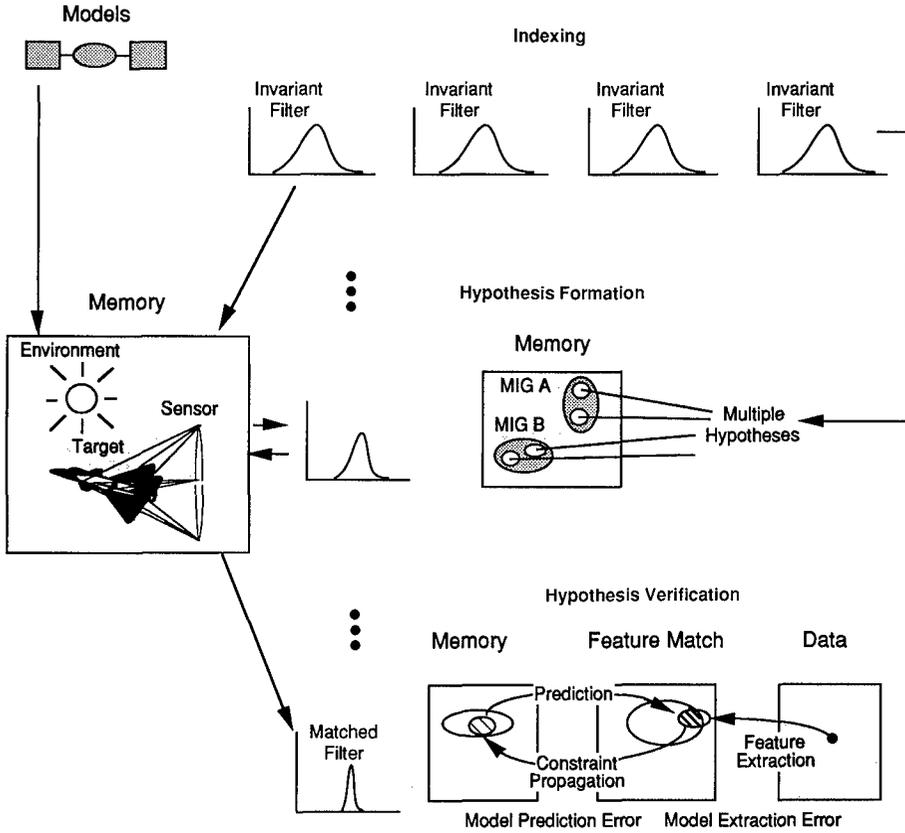
Fig. 4.19 Model-based reasoning illustrating hierarchical reasoning methodology.

An example of the full model-based reasoning process is as follows. A mission description is provided based on recent information that an armored cavalry regiment has been sighted moving on a road toward a forward engaged battle location. Based on this information, the algorithm knows that there should be a number of exercised hot tanks in column formation, moving, with their frontal aspects exposed and with minimal camouflage deployed. The algorithm may also have access to recent weather information in the expected encounter area. The algorithm hypothesizes several different tank aspects at a few different target configurations, such as surface temperatures or turret-to-hull relationships. On arriving in the target region, the sensor picks up several hot blobs in a line, segments them from the background imagery using coarse filters, and passes them to the matching routine. The matching routine finds several length-to-width ratios within certain error bounds of the hypothesized models. The hypothesis generator indicates the image extractor to look for engine and exhaust within the candidate regions. The matching process determines how well the image regions support engine and exhaust hypotheses. If the match is high when matching certain tolerances for engine and exhaust,

further tanklike features are sought from the image. If the matching process only supports the presence of an engine, then other candidate models are tried. Intermediate evidence for certain model matches is maintained for a camouflage or obscuration analysis. The hypotheses list is refined and reduced on subsequent passes until a decision confidence threshold or a processing time line threshold is exceeded and decisions are reported with confidences.

Referring again to Fig. 4.19, we see that the model-based reasoning process contains the knowledge storage and retrieval or memory process, the indexing or inference process, the hypothesis generation process, and the hypothesis verification process. In particular, note that the hypothesis verification process includes a representation of uncertainty in the three different domains of memory, feature space, and data. Feature data is extracted from the data with some uncertainty, represented by an area in the feature domain. The prediction or hypothesis instantiates a model from memory, and that model also has uncertainty associated with the target features originating from both inaccuracies in the target modeling process and the uncertainty of time of day or operating conditions. The uncertainties of the model and the data are compared in the feature domain, and an accounting is made of evidence to support the existence of such a feature conditioned on the data and the model overlap. The degree of match is used for two purposes. Initially, the degree of match is used to constrain the selection of further modeled features for the hypothesized target. Second, the degree of match, or mismatch, may drive further image processing or feature extraction in a directed way to confirm or deny additional features associated with a particular target model. The management of uncertainty and evidence accrual is the heart of the model-based reasoning approach and to a large part determines its success or failure.

Inference Mechanisms and Evidence Accrual.   A typical way of representing reasoning within hypotheses is to form networks of parameters and relationships. The parameters represent free variables corresponding to attributes of entities such as objects or parts of signal components, and the relationships correspond to the way these entities interact in the real world. The network representation provides structure for adding additional information or knowledge. In networks, the uncertain variables are connected by relations that specify local uncertainty measures. These can be conditional probabilities, joint probabilities, belief values, and symbolic or logical constraints on possible joint values. The configuration of the network at any given moment in the decision process is often called the *state,* and the overall set of possible states is called the *state-space.* What is desired for inferencing and hypothesis verification is a consistent method of representing the uncertainty of information in the network.

Use of the Bayes network and state-space representation for matching extracted target geometric features with some stored target representations is illustrated in Fig. 4.20. The network representation is for predominant modeled features for a particular aircraft in a tree structure. The higher nodes in the tree structure denote larger target part signatures, and each of these parts may have several subpart signatures. Underneath the lowest nodes of the tree structure are the image extracted features, which may correspond to the modeled features. Initially, a feature, such as part of the hot engine signature on
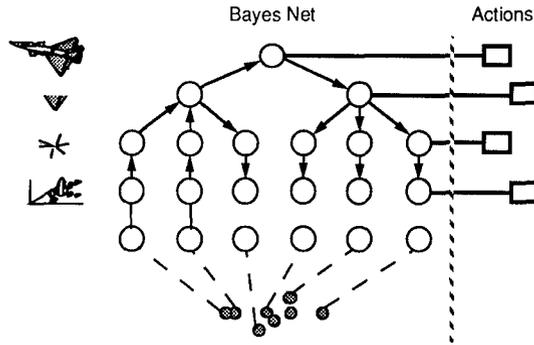
Fig. 4.20   Bayes net evidence accumulation.

an aircraft, is detected in an image, and this matches with the model feature for engine part signature as represented by the leftmost node in the tree structure. The hypothesis generation routine instantiates several possible aircraft models and attempts to verify by matching additional features in a model tree structure for each target. Within this particular model, the query is sent for another target engine part signature, as indicated by the down arrow on the second node of the leftmost tree branch. If this part signature is found, the evidence supporting this particular model increases in the state-space representation. The hypothesis generator then queries for other target parts, either associated with engine signature, or some other part, such as the wing leading edge or cockpit. Each time a match is made between model part or subpart and an image feature, the state-space is updated and the evidence accrued increases for that particular model instantiation. The actions are the queries generated from match or mismatch at each stage of the interrogation procedure.

Typical models[27,28] for uncertainty representations include: (1) probabilistic Bayesian; (2) Dempster-Schafer; (3) fuzzy logic, symbolic constraint; and (4) logic or rule based. Theoretical bases for choosing one model over another are an ongoing area of research. In practice, the uncertainty representation scheme must support a self-consistent method of evidence accrual or the decision process will never converge to a stable solution.

The inferencing mechanism is the generation of more global relationships or conclusions from the local relationships, network structure, and observations. This is accomplished by reducing network uncertainty as more information regarding relationships or parameters becomes available, and calculation of global consistency of the network in supporting one or more alternate hypotheses.

Probability representation schemes usually employ Bayesian probability networks to infer posterior distributions. Inference algorithms available for probability representation include Bayes formula, Pearl's algorithm, Shachter's algorithm, hierarchical Kalman filter, symbolic probabilistic inference, and Speigelhalter's algorithm. Symbolic constraints schemes use simultaneous equations and inequalities to infer constraints, unique solutions, or empty solutions. Symbolic constraint inference mechanisms include term rewriting,

symbolic algebra, or mathematics. Logic representation schemes use rule bases and logic definitions to infer logic data bases, unique solutions, or empty solutions. Inference algorithms for logic representation schemes include prolog, chaining resolution, Rete, and Boyer-Moore algorithms. Dempster-Schafer representations employ belief functions to infer revised belief functions. Inference algorithms for Dempster-Schafer approach are the rules of combination, and true upper and lower probabilities. Fuzzy logic schemes represent networks as membership functions and use T-norms algorithms to infer revisions to these membership functions. A detailed discussion of these representation schemes and inference algorithms is beyond the scope of this chapter. A more detailed treatment of these topics is contained in Fung et al.[28]

4.6.5 Artificial Neural Network Algorithms

There is a debate in the computing and ATR communities regarding whether neural networks are algorithms or computing architectures. Because the majority of neural network developers emphasize the computing architecture character of neural networks, the subject is treated from that perspective here, with the awareness that many of the algorithms implemented on neural networks may be unique to neural networks. Functions performed particularly well by neural networks are those that require both many parallel operations and adaptive learning. In terms of ATR applications, there are currently two primary portions of ATR processes being implemented on artificial neural networks. These processes are segmentation and target classification. These implementation methods and other aspects of neural network methods are discussed briefly in Sec. 4.7.

4.7 ATR ARCHITECTURES

Implementation techniques or architectures for ATR include digital (and sometimes analog) electronic processing or computing, optical processing techniques, and artificial neural network implementation methods. There have been many texts written on each of these architectural approaches for various applications including object recognition. This section, of necessity, only introduces the various implementation schemes and provides a few examples.

Figure 4.21 shows in simplified form how each level of filtering (invariant filter through matched filter) can be implemented on any of the available architectures. If we equate the statistical pattern recognition approach primarily with an invariant filter scheme, the template matching with the matched filter scheme, and the combination of all levels of filtering with the model-based reasoning (MBR) ATR algorithms, we note that each of the ATR algorithm approaches can be accomplished by one or more of the architectural implementations. For any of the algorithm approaches, it is possible to have different functions accomplished by digital, optical, or neural network processors. In each section that follows, a brief description is provided for the processing architecture, and then an indication is provided for which type or portion of ATR algorithms runs best on the architectures.
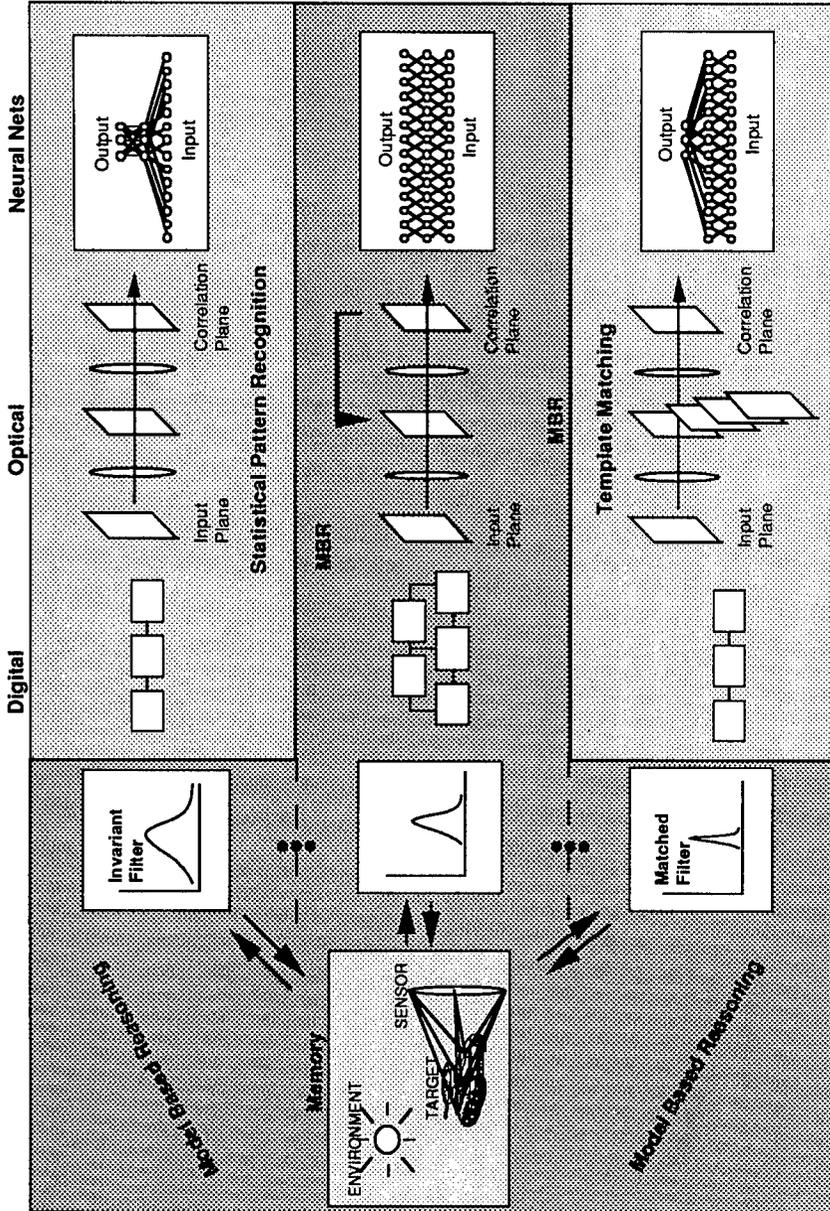
Fig. 4.21 Architectures for various ATR algorithm approaches, using the paradigm of invariant through matched filter operations and the three primary hardware architectures generally employed for ATR processing.

4.7.1 Digital Implementations

Computing Architectures. There are a variety of ways to connect simple computers to form more powerful computing architectures. The connection structures possible include (1) lines or buses, (2) stars, (3) trees, and (4) polygons or rings. The choice of one of these topologies over another depends on the tradeoffs of independence and communication, and speed and efficiency of processing. Completely connected architectures, where every computer is linked to every other computer, are the most flexible but also the most complex to build. Most of the prototype networks emulating complete connectivity have linked their computers through some type of common ring or bus using high-bandwidth wires, through a common memory, or through a crossbar grid of switches consisting of an array of $N \times N$ wires with each of N computers linked with $N \times N$ switches. These completely connected architectures are suitable for only small numbers (2, 4, 8, 16, or 32) of computers. Shuffle-exchange reconfiguring networks that link M nodes to N nodes, using $\log(N)$ banks, each with M switches, can be used to approximate a complete crossbar switching grid to $M \log(N)$ switches.

Pipeline and Array Processors. One of the more simple structures, and among the most powerful, is the pipeline of independent computers. A single array of computers works on each input in serial fashion, and after the pipe is full, outputs emerge in the same order as the inputs, containing the results of the entire sequence of computing. Each processor repeatedly executes the same instruction on a sequence of data flowing through the pipe. If the pipeline has N processors, and d is the time required to fetch and decode the next instruction, then once the pipeline is full the program will execute up to dN times as fast as a single processor computer. The processor that requires the most time in the pipeline will determine the overall speed, because all other processors must wait for it to finish. A short pipeline of powerful processors gives the Cray, the Cyber-205, and other similarly designed supercomputers their power.

Pipelines are very useful whenever images are to be transformed in simple steps as they stream by. Two-dimensional pipeline arrays are available that have a small computer dedicated to each individual pixel of the image and store intermediate results in a dedicated register or memory cell as appropriate. This type of pipeline produces a powerful image transform processor for basic add, subtract, multiply, and divide operations.

A special category of array processors, sometimes called *cellular array processors,* uses a pipeline of processors specially designed to effect 3×3 or larger sized window functions. These array processors compute any logical or arithmetic operation whose operands are the center pixel of the window plus the eight, or other appropriate number for higher than 3×3 windows, nearest neighbors. These array processors may be either floating point processors, which generally compute 32-bit floating point operations, or fixed-point processors, which generally compute 8-bit arithmetic or 1-bit logical operations on each input. Special purpose computer chips are often built to speed up frequently computed processes such as windowing or convolution operations. So-called systolic arrays, emphasizing careful structuring of gates, can be fabricated on special purpose chips to accomplish convolutions and thresholding operations.

Pyramid Processors.   An especially attractive multicomputer architecture for image processing and pattern recognition is a pyramid structure. The base is a large array, with successively smaller arrays on top of it. The base layer can be the size of an image, in which case it is often called a *retinal layer*. A pyramid can be thought of as a tree structure wherein each parent contains several children in the next largest or lower array, and each parent may be linked to nearest neighbors at its own level to form interior arrays at every level.

A number of variations of pyramid processors are possible, emphasizing different converging hierarchical local-neighborhood structures. Each processor can be linked either to four neighbors in a square array, to six neighbors in a hexagonal array, or to more neighbors in a more complex array. Each layer, each subarray within each layer, each pyramid, and so forth, may be given its own controller. Each successive increasing or decreasing size layer of processors can be made more powerful. Processors can have more than one parent, providing overlap. A layer of processors can be sandwiched between two layers of memories, with either several child memories or several parent memories or both. A network of traditional computers can be linked to the pyramid nodes, or a common memory can be used to link pyramid and network.

A pyramid structure allows information to be converged as the image is transformed, simulating a bandwidth reduction of information processing at successively higher levels. The pyramid structure generally provides an efficient means of image noise reduction at the lower levels. The structure requires that the algorithm be written to preserve or coalesce the more salient characteristics in the image at higher levels. It effects an important reduction in the distance required to transfer messages over the image plane. This distance reduction is obtained by passing messages up the pyramid chain and back down another branch, thereby taking advantage of the geometric progression of the tree structure over the arithmetic progression of horizontal or linear transfer across the image plane.

Asynchronous Computing Network Structures.   The largely synchronous designs of pipelines, arrays, systolic arrays, and pyramids are not suitable for more cognitive level computing tasks that may be desired in target recognition algorithms. For example, trying to recognize objects in different portions of the image in a priority based on intermediate measures of target likelihood is a task requiring more independent, directed computing. Asynchronous network structures provide one with the flexibility of many independent processors, each with their autonomous control and complete programmability, in a multicomputer environment. Many asynchronous networks have been designed including[29] variations on the ring, star, lattice, tree, cluster, clustered clusters, snowflake, lens, N-cube, hypercube, hypertree, X-tree, and a host of other possible graph structures. Although the graph structures and the number of processors that can be linked are infinite, the practical aspects of implementing these, and mapping algorithms onto these various topologies, are very new areas of research.

The chief limitation of the asynchronous network is the long time required for communication between the component computers. To illustrate the limitation, a multicomputer's operating system typically executes thousands of

instructions to transmit a packet of information, where each packet is any-where from 1 to 2000 bytes. Thus, it takes thousands of times longer to execute a transfer of information than to operate on that information.

In contrast, in a synchronous array each computer can receive information from one of its four, six, or eight nearest neighbors in one instruction time. Although the asynchronous networks are adaptable and flexible, they are slow, and hence inappropriate and very wasteful for numerically intensive repetitive tasks.

To overcome the long communication times between computers, the cluster of clusters architecture is employed where communication required between higher level clusters is rare. The architecture is patterned after the function-ality expected in terms of frequency of communication required as well as length of information to be transmitted.

Reconfigurable Networks.   Reconfiguring networks have primarily been used in $N \log(N)$ banks of computer switches to allow N pieces of information to be shuffled among N computers, processors, or memory components. Much smaller sets of switches can be employed at a variety of levels, to effect partial recon-figuration. Programmable switches have been used to change the characteristic 3×3 to hexagonal kernel neighborhood configurations used in array pro-cessing. A tree can be reconfigured into an array. It is possible to intersperse switches with computers in a two-dimensional grid network. It is also possible to reconfigure the linkages between controllers and processors. The benefit of these systems is that many different topologies could be represented by the same basic network, as needed at different stages of the required algorithm process.

Matching ATR Algorithms with Processor Architectures.   Before attempting to describe the mapping of ATR algorithms to digital processor types, it is important to note several things. First, the configuration of digital processing employed usually depends on the stage of development for a given algorithm. Second, it is highly dependent on the overall processing speed required in image frames per second, which ultimately may be translated into operations per second or instructions per second. Third, it is highly dependent on the particular choice of algorithm components and order in which these are to be computed. Fourth, as a general rule, the level of adaptive control and feedback inherent in the process will drive a need for more adaptive and cognitive processing at the expense of highly parallel, computationally intensive, re-petitive processing. Finally, it is possible to configure any number of combi-nations of the architectures discussed, pipeline, array, pyramid, asynchronous, and reconfigurable, in a hybrid or multimodal computing configuration.

Template-matching schemes tend to be implemented on highly parallel ar-ray processors to implement the typically employed convolution or Fourier transform operations. Control of the process, with decisions for which templates to process and quality of match for each, are handled with more conventional computers.

Statistical pattern recognition algorithms are readily implemented on dig-ital processors using a combination of highly parallel and serial processors. Generally the front-end prescreening and segmentation routines are run on cellular arrays, array processors, or pyramid processors that process every pixel

in the image. The feature extraction routines may be run at higher levels of a pyramid architecture. Pattern classification and matching routines are generally accomplished on serial computers.

Model-based reasoning algorithms are usually run on special-purpose asynchronous network structures that combine serial and parallel processing in a distributed processing strategy. As with statistical pattern recognition algorithms, the front-end preprocessing and initial region of interest detection algorithms employ massively parallel two-dimensional array processors that operate on the entire image plane. Cellular array processing is often used for region of interest detection operations such as thresholding or erode and dilate to be implemented on these array processors. Architectures for the model-based algorithms include various hybrid combinations of topologies for computing including connection machines, butterfly networks, hypercube, graph structures, and pyramid and general-purpose computers. Generally communication between individual processors occurs on some common blackboard.

In reality, most ATR algorithms are run on some distributed hybrid processing network, where coarse-to-fine processing strategies are implemented as a series of operations on computers operating in parallel. The exact configuration of final hardware processors is almost as unique as the algorithm suites implemented and relies heavily on very high speed integrated circuit and very large scale integrated circuit technologies.

4.7.2 Optical Processing Implementations

Implementing statistical pattern recognition algorithms on an optical processor can be accomplished by creating composite Fourier transform filters for all possible viewing angles and target types and placing these in the transform plane of the optical system. Reference transform filters for these implementations are spatial light modulators that are either analog film modulators or digital devices, such as light-emitting diodes.

Some of the earliest implementations of optical matched filtering were applied to type character recognition. In particular, VanderLugt[30] demonstrated spatial frequency filters for nine different characters, generated on photographic film. It was necessary to mechanically swap the various filters into and out of the coherent optical transform path in these early systems. This was a very time consuming procedure. More recent implementations of the optical matched filtering procedure have employed spatial light modulators, which have the advantage of nonmechanical switching of filters at high speeds. A large number of target signature templates can now be stored in optical holographic crystals. In practice, these reference templates are compared to newly acquired sensor imagery using the techniques described in Sec. 4.6.3.

4.7.3 Neural Network Implementations

Artificial neural networks have a long history associated with the biological and medical community. Computer simulated neural networks for image/signal recognition processing applications are well documented in a number of recent surveys.[31–33] Several organizations are working on the development of very large scale integrated circuit components that are specialized for the

implementation of neural networks. This section discusses specific applications to ATR, with brief discussion of particular approaches that appear very promising.

Neural network architectures are based on models for exploring many competing hypotheses in parallel. The network topology, node characteristics, and learning or training rules collectively specify the neural network model. The learning rules specify an initial set of weights and indicate how the weights should be adapted during use of the network.

Neural networks can be applied to the implementation of a variety[34] of ATR algorithm strategies. The implementation of a statistical pattern recognition algorithm on a neural network is accomplished by generalizing the target representations in the intermediate network layers. To force generalizations associated with the invariant decision filters, a neural network would implement few intermediate nodes between input and output layers. A template-matching algorithm can be implemented by inserting many intermediate hidden layer nodes in a neural network, to retain specificity inherent in the reference target templates. The configuration of a neural network for model-based reasoning algorithms is most likely to be a combination of the described types, depending on the requirement of the algorithm at a particular processing stage.

Neural architectures for ATR applications generally fall into one of two categories, either (1) neural vision or (2) neural classifiers. Neural vision schemes for ATR should provide at least one of the following: (1) boundary completion for occluded targets, (2) normalization of spatial discontinuities from varying luminance values, or (3) segmentation based on texture information. Neural classifiers for ATR applications should at least (1) maintain parallel search with very complex image or signal patterns or (2) be able to detect novel patterns and create new categories for such.

Neural vision models are primarily generated to emulate biological vision systems. These models most often perform contrast enhancement of image data through the competitive interactions of on-center and off-surround structures. A few models also incorporate cooperative-competitive behavior to perform edge and contour detection. A specific class of models have cooperative-competitive designs with feedback between layers that enable them to fill in occluded boundaries and eliminate noise. An example of the latter class is the boundary contour system[35] (BCS). The BCS network consists of three competitive layers, a cooperative layer, and a feedback loop to the first competitive layer. The adaptive feedback mechanism for filling in missing boundary information makes this neural net scheme attractive because this is a serious limitation in other edge-based region segmentation algorithms.

Lippman's[32] comparison of neural network classifiers with traditional statistical classifiers is helpful for those working in the area of ATR. The functional forms of traditional statistical and neural network classifiers are compared in Fig. 4.22.

In this diagram, inputs and outputs are seen to be passed serially, and computations are performed serially in the more traditional statistical classifier. The statistical classifier has N inputs, with associated symbols or extracted target features. An algorithm computes which of M classes most closely resembles the input in terms of pretrained exemplar weightings of the symbols or features by class. Weighting parameters for the target features are estimated

Traditional Classifier

Neural Net Classifier
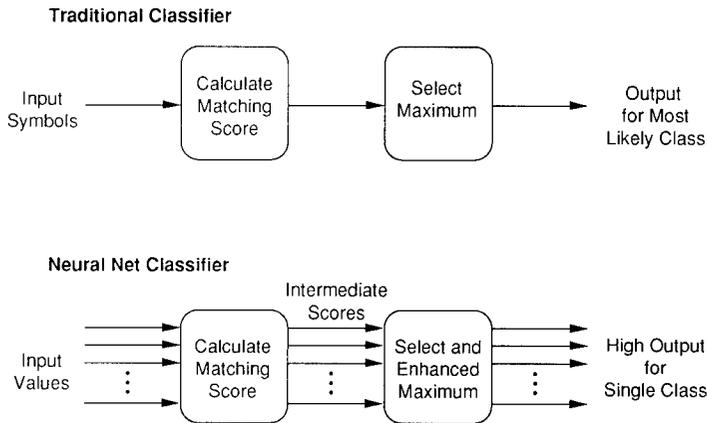
Fig. 4.22   Schematic comparison of traditional and neural network classifiers.

from training data in advance, and then held constant during the application of the algorithm. Statistical classifiers are discussed in detail in Sec. 4.6.2.

In an adaptive neural net classifier, the inputs and outputs are in parallel, and all internal computations are performed in parallel via the node structure. In these classifiers there are N input connections; usually each is associated with a portion of the input (e.g., pixels in an image chip of a target signature). The first stage computes matching scores and outputs these scores in parallel to the next stage via M outputs. The second stage selects the maximum and enhances it, the highest weight of all outputs defining the best match of the possible M classes. Internal parameters are generally adapted during use of the neural net classifier, and the weighting on output values defines the most likely class match. Adaptation will more likely produce the correct response for succeeding input patterns that are similar to the current pattern.

A taxonomy of six neural network types that can be used as classifiers for fixed patterns is shown in Fig. 4.23 as a guide to the many available classifiers described in the literature. The taxonomy breaks six neural network types[32] first into those that operate on binary valued inputs versus those that operate on continuous valued inputs. Next, the networks are divided into those that are trained with supervision and those that are not supervised during training. Networks trained with supervision are generally used as classifiers or associative memories. Examples of supervised networks are the Hopfield,[36] Hamming,[37] and single- and multilayer perceptrons.[2] Networks trained without supervision are generally used to quantize vectors or to form clusters. No information is provided as to which class is correct during the training stage. Examples of the unsupervised learning networks are the Carpenter/Grossberg classifier,[38] and the Kohonen self-organizing feature maps.[39] These networks are provided with no information regarding "correctness" of class outputs during training, thus forming self-organized clusters based on similar (arbitrary) characteristics among input data. All networks can be trained adaptively, although some, such as the Hopfield and Hamming networks, are generally
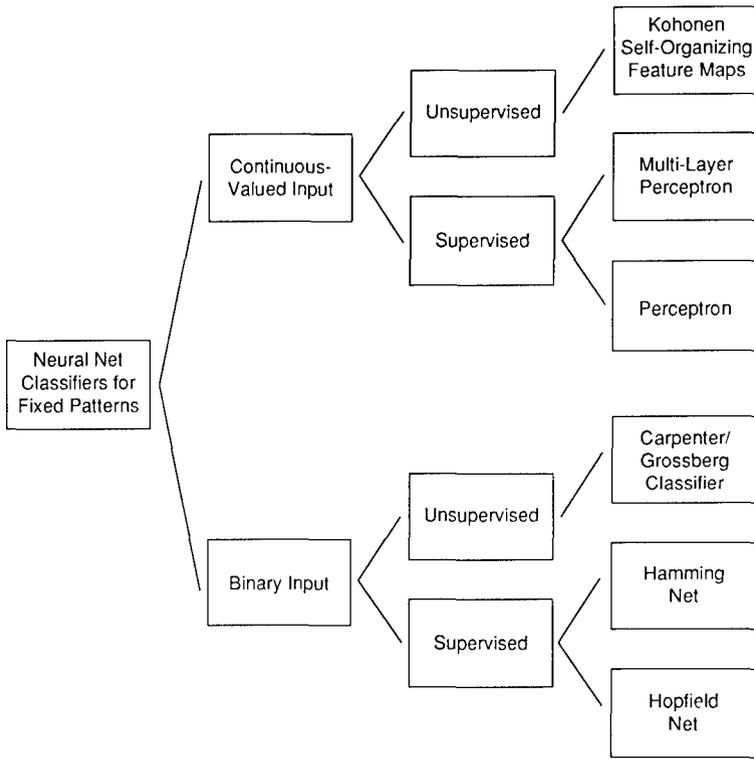
Fig. 4.23 Taxonomy of six neural networks that can be used as classifiers for fixed patterns.

employed with fixed weights. Lippman[32] also discusses the functional correspondence of certain classical and neural network algorithms.

Supervised learning algorithms are similar to traditional statistical classifiers in that they are provided with a set of weights for each class exemplar during the training stage. The Hamming network is in fact an implementation of the optimum classifier for binary patterns in random noise.[37] The perceptron structure performs fundamentally the same calculations as required for a Gaussian classifier when weights and thresholds are selected appropriately. Perceptrons can also be set up to perform a convergence procedure[32] that is fundamentally different than a Gaussian classifier.

The unsupervised learning networks can be more challenging to use in ATR algorithms simply because the network forms its own internal feature representations, associations, and generalizations. Kohonen's network forms a prespecified number of clusters similar to the K-means algorithm, where K is the number of clusters formed. Unlike the K-means algorithm, the Kohonen network has only one presentation of each new pattern, and the weights are modified after each presentation. These types of algorithms can be useful for certain portions of ATR algorithms where the best features to employ to cluster the target objects are not obvious to the algorithm designer. Caution must be employed to ensure that the internal network features (intermediate parti-

tioning) and weightings are not trained specifically to undesired or peculiar features in a limited set of input data.

There is much ongoing work in the field of applying artificial neural networks to pattern recognition and image processing. A more in-depth discussion of this work is not possible in the limited space of this volume. Before closing, some important observations regarding neural network applications to ATR should be reiterated. Portions of ATR algorithms requiring large numbers of complex relationships to be computed in parallel, and which are not easily programmed explicitly are perhaps most appropriate to implement in neural network architectures. Neural networks are not a panacea for the problems associated with developing robust ATR algorithms. Artificial neural networks have a long history that has only recently been applied to the ATR application. Neural network architectures are highly data driven; that is, they generally require large amounts of training data. Neural network architectures, in their current designs, tend to be very specific to the algorithm processes they support. With time, it is hoped that the neural network architectures will become more generally applicable to ATR algorithms as well as less reliant on requiring all possible sample conditions to be represented in ATR training sets.

4.8 ATR COMPONENT PERFORMANCE EXAMPLES

A performance curve relates the trade-off between two contraindicating performance measures. For an ATR, these pairs of performance measures usually relate to the trade-off between making errors of commission or omission. Depending on the ATR functional components, typical pairs are the probability of detection versus probability of false alarm, the detection rate versus false alarm rate, and the probability of correctly classifying vehicles versus probability of incorrectly classifying nonvehicles as vehicles. The trade-off curve can be used to choose an operating point to meet cost and/or effectiveness criteria.

This systems analysis requires a performance model that relates ATR performance to the trade-off under consideration for the variables of interest. These typically include the various scenario variables and the system design parameters. Performance measures considered in an ATR performance model for these analyses are probability of detection and false alarm rate. In other electro-optical systems analyses, performance models are often based on distributional models for the threat and environmental variables and their effects on target signature. For ATR applications, the development of distributional models for natural and cultural clutter is still a distant goal. This forces a reliance on phenomenological descriptions of the scenarios in terms of the scenario variables or in terms of the image characteristics, which are a first step toward describing the required probability distributions.

The same analyses that yield performance models are used to evaluate the measured performance of an ATR. Performance evaluations are necessary in formulating the ATR concepts and in optimizing the performance of ATR. They are particularly important tools for the definition of required sensor characteristics and algorithm parameters. Performance curves and the corresponding confidence intervals can be used to compare ATR algorithm design options and to determine sensitivity to sensor characteristics and environmental condi-

tions. They both are tools that if misapplied will confuse rather than clarify issues.

4.8.1   Performance Data

Figures 4.24 and 4.25 show the detection performance curves for a variety of point-of-interest detectors on two data sets. The first data set had a wide range of infrared imaging conditions, which yielded contrasts and target sizes that ranged from undetectable to fully identifiable. The second data set contained a narrower range of contrasts at two resolutions and emphasized higher clutter conditions, which made detection more difficult. Each of these two sets was obtained with a different infrared imaging sensor and at a different location. This, along with variation in other collection parameters such as time of day, provided a reasonable range of scenario parameters and some variation in ATR system parameters.

   The performance curves were plotted by sorting the point of interest detections in decreasing confidence order. Each detector reported a decision confidence statistic along with the location. Then the fraction of true targets found versus a normalized count of false alarms was plotted parametrically as a function of a cutoff on the confidence values.

   A good algorithm has a high correlation between its confidence value and actual detected targets. This means there should be relatively more detections at the head of the list than at the bottom. The resulting performance curve should quickly arch upward and have a constantly decreasing slope as in Figs. 4.24 and 4.25. This slope indicates that as the confidence threshold is
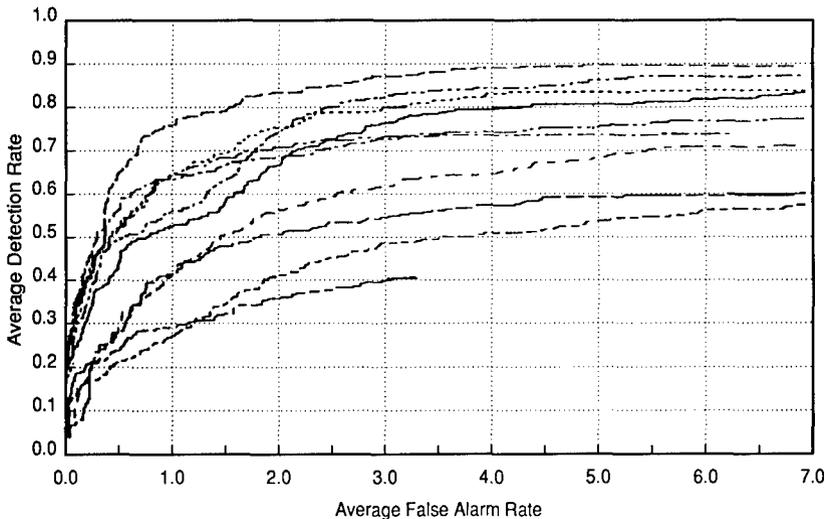
Fig. 4.24   Comparison of detection algorithms over Aberdeen Proving Grounds data set (107 images, 310 targets) as a function of threshold variation.

lowered, the incremental detection rate diminishes while the false alarm rate increases. The performance curves were generated[40] from data reported on two image data sets collected at the U.S. Army Aberdeen Proving Ground (APG), Maryland, and the U.S. Marine Corps reservation at Twenty-Nine Palms (29P), California.

For detector subcomponent characterization, only certain portions of these image data sets were included in the experiment. From the APG data set, 107 images were chosen containing 310 targets, whereas for 29P data set, 64 images were selected containing 184 targets. Figure 4.24 summarizes the results available from 10 detection algorithms on the APG data set. Figure 4.25 shows the results for 12 algorithms on the 29P data. Note that these plots do not reflect actual expected ATR performance for any given scenario. The data sets were specifically chosen to represent a wide range of targets and environmental conditions. Expected performance varies from easy classification to impossible detection. Therefore, the plots enable comparison between algorithms based on average performance.

These performance plots can be used to select average operating points for algorithms. Two simple choices for selecting operating points are to fix the fraction of detected targets or fix the false alarm rate. Where these choices intersect, the performance curve for an algorithm determines the required threshold on confidence. Algorithm behavior can then be studied in greater depth at the particular operating points selected. This also allows a common basis for comparisons between algorithms. However, these curves do not allow performance to be estimated for other conditions or subsets of the conditions used in the original experiments.
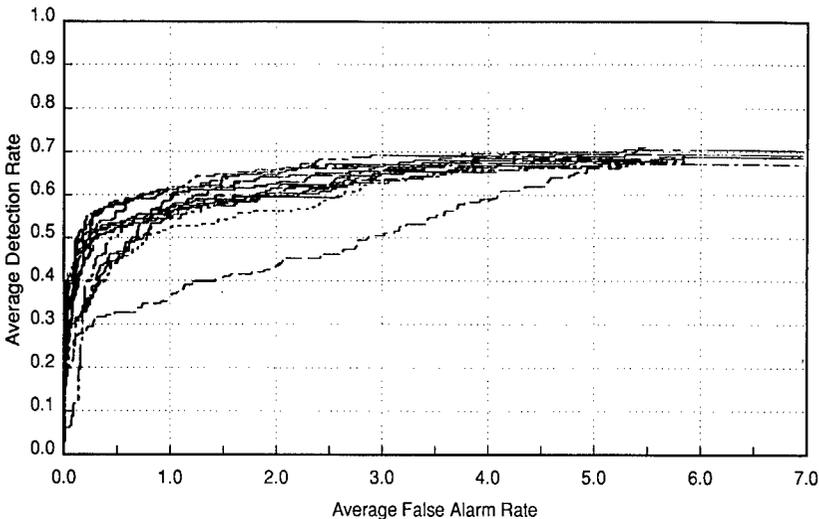
Fig. 4.25 Comparison of detection algorithms over 29 Palms data set (64 images, 184 targets) as a function of threshold variation.

4.8.2   Detection Performance Analysis

The data from these same experiments were used to develop a performance model for three of the best detectors in Figs. 4.24 and 4.25. The confidence reported by the detectors was modeled as a random variable whose distribution depends on the target metrics. Analysis of variances (ANOVA) techniques were used to evaluate this dependence for various metrics and combinations. The product of two image metrics from Table 4.2, RN_0 and TIR^2 was found to produce the simplest form and best reduction of uncertainty. The number of resolution cells on target RN_0 is a target metric. The target/background interference ratio (TIR) is a target to background metric. This dimensionless product is roughly an integral measure of the target signature strength relative to background. The model developed for the distribution of confidence of targets given the target metrics is non-parametric. The mean of the distribution was obtained from the analysis of variance (ANOVA) fit to the confidence values for targets. The shape of the distribution was modeled by the histogram of the difference between confidence values and corresponding expected value based on target metrics. This distribution was normalized by a smoothed version of the measure probability of detection versus the same target metrics. By intersecting these distributions of detection responses estimated at different values of $RN_0 * TIR^2$ with the experimental distribution of false alarm responses, families of performance trade-off curves are generated. Figure 4.26 depicts the detector performance curves for an algorithm at three values of the metric $RN_0 * TIR^2$. As the metric decreases, the plateau value decreases, indicating a lower proportion of correctly detected targets. In addition, as the metric decreases, more false alarms are produced for each correct detection.
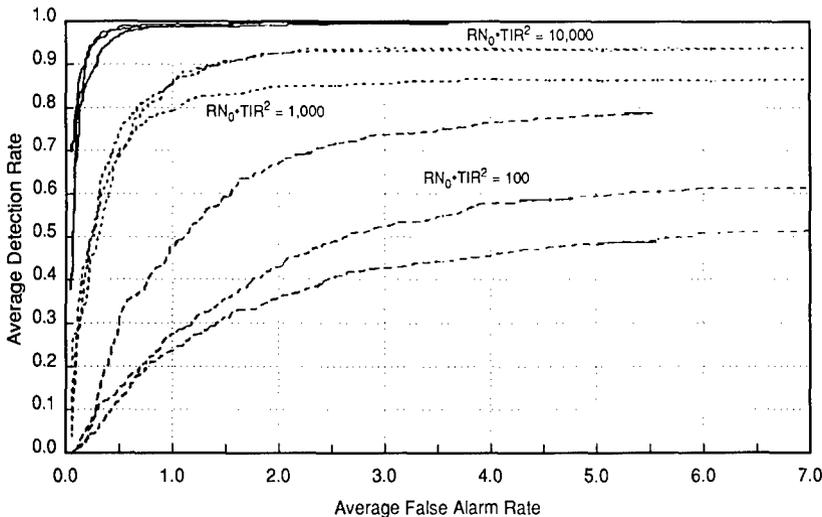
Fig. 4.26   Dependence of detection algorithms on target strength as a function of threshold variation averaged over background of characterization data set (530 targets).

It is curves such as these that are necessary for enlightened mission analysis studies.

Intuition suggests that independent of the background, the resolution cells and contrast of a target will partially determine its detectability. In some cases, the contrast can be traded for increased size or vice versa. Visual support for these claims is provided by the IR images of targets shown in Fig. 4.27. The grid of targets in this figure are arranged left to right by increasing size (number of resolution cells RN_0) and top to bottom by increasing contrast. Along any diagonal from the bottom left to the top right the product of RN_0 and TIR^2 is approximately constant. Indeed, the analysis described suggests that for several algorithms, targets along a diagonal have the same detectability on average.
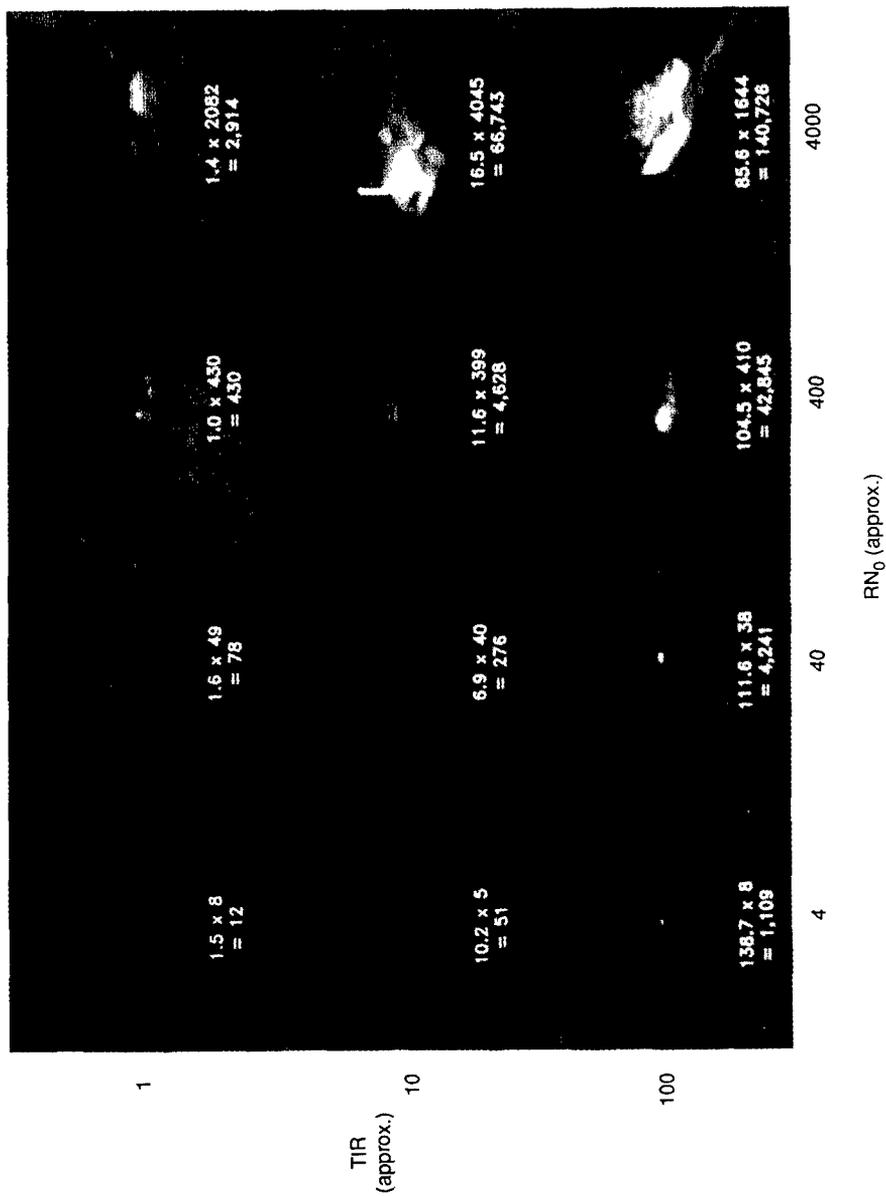
Fig. 4.27 Visual display of target detectability over the range of values of integral contrast.

REFERENCES

1. Automatic Target Recognizer Working Group (ATRWG), Department of Defense, *Target Recognition Definition and Performance Measures,* Document 91–002, Appendix II, IIT Research Institute/GACIAC, Dayton, OH (June 1991).

2. M. Minsky and S. Papert, *Perceptrons, An Introduction to Computational Geometry,* MIT Press, Cambridge, MA (1969).

3. J. Johnson, "Analysis of image forming systems," in *Proceedings of the Image Intensifier Symposium,* pp. 249–273 (1988).

4. ATRWG, *Image Metrics for ATR Evaluations,* Document 90–002, IIT Research Institute/GACIAC, Dayton, OH (June 1990).

5. L. G. Clark and V. J. Velten, "Image characterization for automatic target recognition algorithm evaluations," *Optical Engineering* **30**(2), 147–153 (Feb. 1991).

6. R. Emaus, *Detector Performance Model,* Vol. 6, Autonomous Infrared Techniques (AIRT) Program, Final Technical Report, U.S. Army Contract MDA 903-84-C-0814 (Apr. 1988).

7. K. S. Fu, *Syntactic Methods in Pattern Recognition,* Academic Press, New York (1974).

8. B. Bhanu, "Automatic target recognition: state of the art survey," *IEEE Transactions on Aerospace and Electronic Systems* **AES-22,** 364–379 (1986).

9. B. J. Schachter, *Analysis and Evaluation of IR Target Detection, Segmentation and Classification Algorithms,* Final Report, Contract DAAL02-85-C-0085, Westinghouse Electric, Baltimore, MD (1988).

10. M. J. Lahart, T. Jones, and F. Shields, Night Vision and Electronic Sensors Directorate, private communication (1988).

11. R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis,* p. 271, John Wiley & Sons, New York (1968).

12. J. M. Tennenbaum, A. C. Kay, T. Buford, G. Falk, J. Feldman, G. Grape, R. Paul, K. Pringle, and I. Sobel, "The Stanford hand-eye project," in *Proceedings of the International Joint Conference on Artificial Intelligence,* D. A. Walker, L. M. Norton, Eds., pp. 521–526 (1969).

13. E. C. Hildreth, "The detection of intensity changes by computer and biological vision systems," *Computer Vision, Graphics and Image Processing* **22,** 1–27 (1983).

14. W. Brown, T. Crimmins, and S. Sternberg, *An Introduction to Image Algebra,* The Environmental Research Institute of Michigan, Ann Arbor (1985).

15. S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis Machine Intelligence* **PAMI-6,** 721–741 (1984).

16. S. Hsu and J. A. Famighetti, "A stable structure approach to automated scene understanding and awareness with multipolarized SAR image data," *Journal of Imaging Technology* **13,** 195–201 (1987).

17. M. J. Lahart, "Estimation of error rates in classification of distorted imagery," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6,** 535–542 (1984).

18. C. J. Stone, "An asymptotically optimal window selection rule for kernel density estimates," *Annual of Statistics* **12**(4), 1285–1297 (1983).

19. S. P. Smith, *CAPIR Final Report,* Contract DAAL02-85-C-0151, Northrop Research and Technology Center, Palos Verdes, CA (1987).

20. B. Efron, *The Jacknife, and the Bootstrap, and Other Resampling Plans,* SIAM, Philadelphia (1982).

21. T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory* **IT-18,** 21–27 (1967).

22. K. Fukunaga and D. M. Hummels, "Bayes error estimation using Parzen and k-NN procedures," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-9,** 634–643 (1987).

23. D. H. Foley, "Considerations of sample and feature size," *IEEE Transaction on Information Theory* **IT-18,** 618–626 (1972).

24. A. Rosenfeld and A. C. Kak, *Digital Picture Processing,* Chap. 8, Academic Press, New York (1976).

25. J. W. Goodman, *Introduction to Fourier Optics,* Chap. 7, McGraw-Hill, New York (1968).

26. E. G. Zelnio, "ATR paradigm comparison with emphasis on model-based vision," in *Conference on Model-Based Vision Development and Tools,* R. L. Larson and H. N. Nasr, *Proceedings of the SPIE* **1609**, 2–15 (1992).

27. D. Morgan, T. Miltonberger, and G. Orr, "A sensor algorithm expert system," in *Digital Image Processing Applications, Proceedings of the SPIE* **1075**, 367–372 (Jan. 1989).

28. R. Fung, K. Chang, D. Morgan, and T. Miltonberger, *Evidence Accrual,* Final Report TR3280-001, Contract F33615-90-C-1482, Wright-Patterson Air Force Base, OH (Apr. 1991).

29. L. Uhr, "Parallel architectures for image processing, computer vision, and pattern perception," Chap. 18 in *Handbook of Pattern Recognition and Image Processing,* T. Young and T. Fu, Eds., Academic Press, New York (1986).

30. A. VanderLugt et al., "Character reading by optical spatial filtering," in *Optical and Electro-Optical Information Processing,* J. T. Tippet, Ed., p. 125, MIT Press (1965).

31. J. A. Anderson and E. Rosenfeld, *Neurocomputing, Foundations of Research,* The MIT Press, Cambridge, MA (1988).

32. R. P. Lippman, "An introduction to computing with neural nets," *Association for Computing Machinery SIGARCH Computer Architecture News* **16**(1), 7–25 (1988).

33. ATRWG, *Artificial Neural Systems,* ATRWG White Paper, IIT Research Institute/GACIAC, Dayton, OH (Apr. 1988).

34. ATRWG, *Comparison of Statistical Pattern Recognition, Model-Based and Neural Network Approached for Automatic Target Recognition,* ATRWG White Paper, IIT Research Institute/GACIAC, Dayton, OH (June 1991).

35. M. Cohen and S. Grossberg, "Neural dynamics of brightness perception: features, boundaries, diffusion, and resonance," *Perception and Psychophysics* **36**(5), 428–456 (1984).

36. J. J. Hopfield and D. W. Tank, "Computing with neural circuits: a model," *Science* **233**, 625–633 (Aug. 1986).

37. R. P. Lippman, B. Gold, and M. L. Malpass, *A Comparison of Hamming and Hopfield Neural Nets for Pattern Classification,* Technical Report TR769, MIT Lincoln Laboratory, Cambridge, MA (June 1987).

38. G. A. Carpenter and S. Grossberg, "Neural dynamics of category learning and recognition: attention, memory consolidation, and amnesia," in *Brain Structure, Learning, and Memory,* J. Davis, R. Newburgh, E. Wegman, Eds., AAAS Symposium Series (1986).

39. T. Kohonen, *Self Organization and Associative Memory,* Springer-Verlag, Berlin (1984).

40. J. Sherman, "Development and use of confidence intervals for automatic target recognition evaluation," in *Automatic Object Recognition,* H. N. Nasr, Ed., *Proceedings of the SPIE* **IS 7,** 144–169 (1990).

Directed Energy Systems

Gary Golnik
W. J. Schafer Associates
Chelmsford, Massachusetts

CONTENTS

5.1 INTRODUCTION

Directed energy (DE) systems include high-energy lasers (HELs) for weapon and fusion applications, laser radar systems, neutral particle beams, laser communication systems, laser systems for material processing and many others. This chapter deals with high-power systems; the goal is to aid in the initial conceptual design of directed energy systems. Various types of DE systems are described and the emphasis is on the transport and control of the energy, rather than on its creation.

Directed energy applications came into being not long after the development of the first lasers, beginning with weapon and fusion systems. These efforts developed into vigorous programs at the Army, Navy, Air Force, Defense Advanced Research Projects Agency (DARPA), Strategic Defense Initiative Organization (SDIO), Lawrence Livermore National Laboratory (LLNL), Los Alamos National Laboratory (LANL), the University of Rochester, and many other laboratories around the world.[1-5] (Note that Ref. 3 has more than 700 references to almost all unclassified laser work through 1974.) As lasers grew more sophisticated, and the ancillary detector, coating, control system, and other technologies developed, applications such as laser communication and laser radar became possible.

All of these systems required similar functions. Creation of laser energy was first. The device and resonator design determined the output properties of the resulting beam. The beam transfer system (BTS) shaped the beam, corrected its phase, and transferred the beam to the directing aperture. To achieve significant range, the directing aperture was usually a telescope of diameter considerably larger than the original beam. The transfer process included some losses in beam energy, control of the lost energy, and stabilization of the beam's pointing angle. The beam director pointed the beam at a target and maintained the target aim point for the duration of irradiation. Simultaneous achievement of these functions, often at high speed and in the presence of structural vibrations, required sampling, sensing, and control systems. To design such systems, an optical system engineering methodology emerged simultaneously at many company and government laboratories.

5.2 ELEMENTS OF A DIRECTED ENERGY SYSTEM

Figure 5.1 shows the key elements of a DE system. These include beam formation, beam transfer, beam expansion, tracking, and control systems. De-
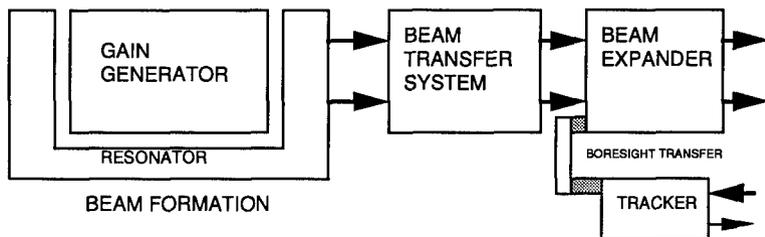
Fig. 5.1 Elements of a directed energy system.

pending on the requirements of the specific application, some of the elements may be minimal. For example, a laser welding system that delivers energy to a fixed point does not require a tracking system to track a target. A laser weapon system does.

Beam formation is the creation of the coherent beam by the initial resonator, oscillator, or beam source. Many high-energy lasers consist of a single gain generator and resonator. Laser fusion systems and master oscillator power amplifier (MOPA) lasers are examples of distributed gain systems. Intervening amplification stages are part of the beam transfer system. Amplifier power extraction is discussed in Sec. 5.5.

The beam transfer system connects the beam source with the beam expander. In addition to a simple transfer of energy over the required beam path, most control functions take place in the beam transfer system. Tilt control, often called *autoalignment,* and adaptive optics beam cleanup to remove aberrations are the two primary functions. The BTS often includes a beam sampler. Fine pointing of the beam, or larger angle optical retargeting, is also done in the BTS.

The beam expander system may be a large telescope or a small focusing mirror, depending on the application. For long-range applications, the beam expander is as large as practical. A larger beam expander provides a smaller spot size on the target. Larger beam expanders usually have larger magnification, which reduces the effect of jitter on far-field spot spreading. The telescope primary mirror may provide a sample of the beam by grating patches, commonly called *holographic optical elements* (HOEs). HOEs provide a sample of the wavefront as it leaves the system, thus including all system aberrations. The beam expander is able to tilt, most often on a gimbal, to provide gross pointing of the beam. The boresight transfer system, connecting the beam expander and a separate aperture tracker, is part of the beam expander system.

The tracker system provides the information required to point a beam at a target. The target may be cooperative or uncooperative. Trackers can be active or passive. An active tracker illuminates the target and receives the returned illuminator energy. A passive tracker observes the ambient signature emitted or reflected by the target. The sensing scheme employed may be imaging, laser radar, centroid track, or many others. Whether separate aperture or shared aperture, trackers are most often mounted on the beam expander. A separate aperture tracker has its own light gathering telescope, boresighted to the main DE beam axis by auxiliary optical surfaces. A shared aperture tracker uses the main beam expander as its telescope, sharing the aperture through a grating or beamsplitter.

In addition to these basic functional elements, a DE system may include a low-energy laser for calibration purposes, cooling systems for mirrors and thermal management, auxiliary sources and sensors for alignment and calibration, and a computer control system. Beam clippers and dumps for energy management, isolation systems for vibrational isolation, and electrical power systems (for either auxiliary or primary power, depending on the type of DE system) provide supporting functions. A beam path conditioning system; a contamination control system, including shrouds, doors, and windows; and a structural assembly complete the system. These auxiliary systems may have a direct influence on DE system performance.

Ground-based laser (GBL) systems have other elements to assist in atmospheric compensation and relay the beam to cooperative relay mirrors. A beacon laser is used to maintain pointing between the GBL and its relay mirror, and between the relay mirror and the fighting mirror. The simplest relay is a large flat mirror. The fighting mirror usually has a receiving telescope, a beam control system, and a transmitting telescope.

5.3 DIRECTED ENERGY SYSTEM EXAMPLES

5.3.1 High-Energy Lasers

High-energy laser (HEL) is a generic term that has come to mean high average power lasers. High-instantaneous-power (pulsed) lasers are discussed later. Most high-energy lasers are gas lasers, although arrays of diode lasers can achieve high average power. Carbon dioxide (CO_2) lasers have been used for many years in industrial processes (welding, etc.). High-power CO_2 lasers typically use an electrical discharge through a flowing gas to create gain. Gas dynamic CO_2 lasers generate the gain through combustion followed by expansion through a supersonic nozzle. Continuous wave (cw) chemical lasers (HF, DF) also generate gain by combustion followed by expansion. Other high-power chemical lasers include oxygen iodine and carbon monoxide. Free-electron lasers generate gain by coupling energy from accelerated particles to an optical field. High average power potential comes from a long train of very short pulses.

5.3.2 Particle Beams

Some DE applications use neutral and/or charged particle beams. The principles of operation are considerably different from lasers, although there are many analogous subsystems (magnetic lenses instead of optical mirrors). Coverage of such a broad and separate subset of DE systems is beyond the scope of this chapter.

5.3.3 Laser Fusion Systems

Inertial confinement fusion (ICF) research over the last three decades included a variety of laser systems.[6] ICF systems typically operate in a single-pulse regime or at relatively low repetition rates. ND:glass, Nd:YAG, excimer, and CO_2 systems have been used. Laser fusion systems use an oscillator to produce a laser pulse, and amplify it by a series of amplifiers of various shapes and increasing sizes. The initial beam splits into many individual beams (each beam separately amplified). The beams focus into a target chamber, their arrival timed so they simultaneously irradiate a fusion target uniformly from all sides. The incident energy causes a reaction, condensing the target rapidly, and beginning fusion. Practical fusion power plants would require a succession of such pulse-target interactions to maintain continuous power output.

5.3.4 Laser Radars and Illuminators

As strategic applications have increased, both laser radar and illuminator systems have become high-power systems. An illuminator is a high-power laser

whose output beam spreads to cover an angle optimized between the conflicting constraints of coverage area and power on target. The illuminator energy reflects from the target and returns to the tracker. Depending on the system, the illuminator may be cw or pulsed.

Laser radar systems are the subject of an extensive literature.[7,8] A laser radar sends out a train of pulses that reflect from the target and return to the receiver. The time of flight, Doppler frequency shift, and other beam characteristics are used to determine information about the target. Distributed targets can be interrogated throughout their volume (monitoring of pollution gases in the atmosphere is an example). Laser range finders are a subset of this technology.

5.3.5   Laser Communication Systems

Laser communication systems, for both low-earth-orbit and deep space communication, are very similar to other HEL systems, except for modulation of the outgoing energy and a cooperative receiver.[9] A cooperative receiver allows the use of beacon technology to simplify the pointing and tracking problems. Less power is needed, because sensitive detectors and receiving telescopes can be used. Laser communication systems transmit data at a much higher rate than radio frequency systems. The many advantages of laser communication systems provide relief for many system requirements. For a deep space mission, a relatively small receiver antenna is feasible. Probes need no longer be large antennas with small instrument packages attached.

5.3.6   Future Applications

Directed energy systems have other possible applications, including laser propulsion, laser debris clearing in space, and laser power beaming. Laser propulsion provides a means of limiting the expended mass needed to achieve orbit, or to make orbital transitions.[10,11] A high-energy laser heats an absorbing disk on the base of the craft. The disk's material is ablated and ejected, providing thrust. The system is *monopropellant* and does not require oxidizer or fuel.

Much debris exists in low earth orbit, the result of the natural environment as well as a generation of space launches. This debris could limit activity in space, or require shielding and avoidance. A relatively low power laser, combined with a tracking system, could destroy debris. Ablation-induced thrust, causing reentry into the atmosphere, is another option.

Laser power beaming can be a practical means to transport power through space, depending on the system constraints.[12] Such a system contains a laser located at a convenient source of energy. The laser beam is transmitted through a telescope to a receiver at the remote site. The cost of transporting and erecting a power system at a remote site may be considerably greater than beaming an equivalent amount of power. Conversion of electrical energy into optical energy (and the converse) increases in efficiency each year. A candidate for a receiver location might be a hypothetical settlement on the moon, particularly in its early days when capital equipment must be raised from within the earth's gravity well.

5.4   ENGINEERING FORMULAS—SUBSYSTEM PERFORMANCE

Conceptual design of a DE system begins by determining the overall system requirements. A conceptual design follows. The conceptual design forms the basis for determining an initial performance assessment. The performance assessment forms the basis for a detailed set of requirements. The requirements guide the design to completion. One of the difficulties in designing a directed energy system is that the requirements for such a system are complex and interactive.

There are two general approaches to deriving the detailed system require-ments. One is to specify the requirements for the functional elements of the system. The second is to specify the requirements by performance categories [power, wavefront error (WFE), etc.]. A matrix formulation combining both methods yields valuable insight (Table 5.1). The rows of the matrix represent the functional elements of the system. The columns represent the performance characteristics of the system. The entries at each matrix location represent the requirement. Entries in this section follow the subsystem performance budget (Fig. 5.2). To use this matrix, first determine which subsystem or pa-rameter is of interest. All entries with either a primary or secondary require-ment are discussed in this chapter.

This section provides the various formulas, tables, and figures necessary to perform initial performance calculations for the directed energy system. The results of these calculations provide the values needed in the budgeting meth-odology in Secs. 5.10 to 5.16. After comparing the resulting system performance to the desired goals, repeat the calculations until the system performance is adequate.

This formulary represents a series of approximations, computer code results, and practical experience. The constraints of each system are unique, but all have a common need: *margin*. Margin cannot be calculated and is difficult to defend to a program manager. Although costly, margin is essential. Using these approximate formulas without adequate margin can lead to a system

Table 5.1   Directed Energy System Design Matrix

Irradiance on Aimpoint Characteristics Subsystem	Power	Diameter	BQ(WFE)	BQ (B/D, J)	BQ (Diff., Prop., Other)	Wavelength	Range	Margin
Beam Formation Gain Generator	P	N	S	S	P	P	N	P
Beam Formation Resonator	P	S	P	P	P	N	N	P
Beam Transfer System	S	S	P	P	P	N	N	P
Beam Expander System (and propagation to target)	S	P	P	P	P	N	P	P
Tracker	N	N	N	P	N	N	P	P
Target Coupling	P	S	S	S	S	P	N	N

P - PRIMARY   S - SECONDARY   N - NO REQUIREMENT
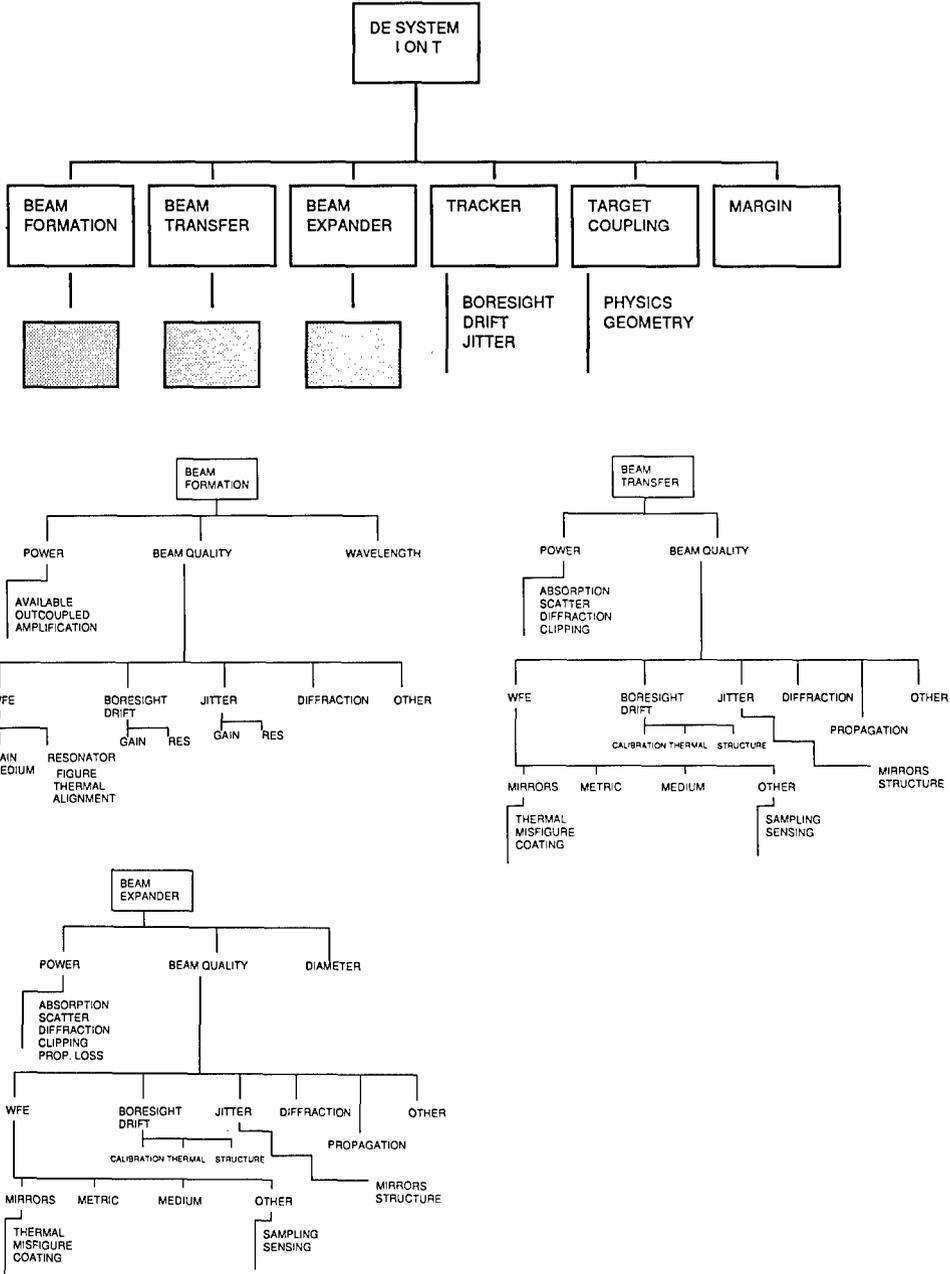
Fig. 5.2   Subsystem performance budget.

design that appears credible, but will be very difficult to build. All of the system requirements for a DE system can be expressed in terms of irradiance on the target and time to retarget. Irradiance on the target includes irradiance on the target aim point, time on target, and target coupling. The use of a common parameter for all requirements allows the requirements to be compared and traded directly.

A useful but potentially confusing term is beam quality (BQ). BQ is defined here as the square root of the ratio of ideal power on target to actual power on target, a dimensionless quantity:

$$\text{BQ} = \left(\frac{P_{\text{ideal}}}{P_{\text{actual}}}\right)^{1/2} . \tag{5.1}$$

A perfect BQ is 1.0, with a degraded BQ greater than unity. Lack of precision in the definition of *ideal* can lead to a BQ of slightly less than 1.0, especially in laboratory measurements. An increase in the numerical value of BQ thus represents a degradation in performance; references to "improving the beam quality" mean decreasing the numerical value of BQ.

Definitions of beam quality include a peak or on-axis Strehl[13] definition, a phase-only definition, and a power-in-the-bucket definition. Sometimes only very specific terms are used to define contributors to beam quality degradation. Beam quality, in this chapter, includes all effects that reduce the amount of energy delivered to the target spot. Figure 5.3 shows the relationships among beam quality definitions. Because beam quality is a ratio of actual power to ideal power, the definition of *ideal* is critical.

Fig. 5.3   Beam quality definitions.

The Strehl ratio is a measure of the peak on-axis intensity of a far-field beam compared to the ideal on-axis intensity. A Strehl ratio of 1.0 represents a perfect beam, whereas a ratio less than 1 indicates a degraded beam. The phase-only beam quality definition is the square root of the inverse of the Strehl ratio. The power-in-the-bucket (PIB) definition is the square root of the ideal encircled energy within a defined radius "bucket" to the actual encircled energy within the same radius. Power-in-the-bucket means the power contained within a defined radius around the centroid of the far-field spot.

5.5   BEAM FORMATION

The characteristics of the DE device (or gain generator) and resonator determine the characteristics of the output beam. The device generates the power output and introduces phase variations that affect the beam quality. The device determines the output wavelength. Other contributors to beam quality, such as partial aperture lasing, also come from the device. In addition, the device may be a source of vibration that affects the remainder of the optical system. The resonator shapes the optical properties of the system, dominating the beam quality effects. Section 5.5.2 discusses the various contributors to the beam quality of the beam formation system.

5.5.1   Power

Power is extracted from a gain medium by an optical system. This section discusses the amount of power that is available from an ideal gain medium, extraction or outcoupling of that power, and amplification effects in systems that have external amplifiers.

5.5.1.1   Available Power.   The volume of available gain and factors such as chemical efficiency determine the amount of available power. The details of available power are unique to each lasing medium. This section gives examples of such calculations but is not exhaustive. Available power is determined by the gain volume and the available energy per unit volume. The cross-sectional area of the gain medium is limited by the width of the optical mode that can usefully be extracted from that medium. Free-electron lasers couple energy to the optical mode from an electron bundle and the gain volume is defined by the beam dimensions. Various parameters are used to characterize the available energy density as a function of volume. For chemical lasers, these include δ and σ, the power density and the specific power. Figure 5.4 shows available power as a function of geometric parameters for various available power densities. This figure can be used to calculate the size of the gain region needed to extract a particular amount of power.

Several factors limit the gain region size. Amplified spontaneous emission (ASE) limits the length of the gain medium. The gain medium must be short enough so that ASE does not dominate. The aspect (height-width) ratio of the gain medium is limited by a resonator configured for reasonable power extraction. High-energy lasers are characterized as either linear or cylindrical. A cylindrical laser has a central combustor and an annular gain region. Other laser types are linear. For rectangular cross-section linear devices, aspect ratios much in excess of 10 are difficult to achieve. For cylindrical resonators, the
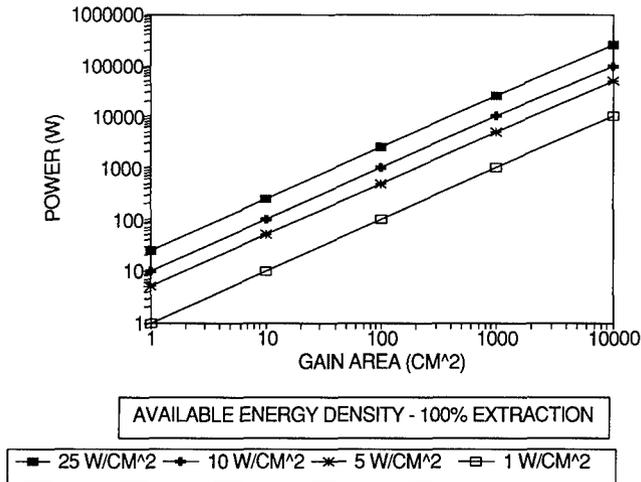
Fig. 5.4 Available power from gain media, laser output power.

availability of optics limits the diameter of the cylindrical gain region. Pumping of the gain region limits its extent in many pulsed systems.

A resonator never couples perfectly to a gain medium. Careful calculation of the mode geometry is necessary to determine the true gain volume.

5.5.1.2 Outcoupled Power.
Outcoupled power is often calculated by what is known as *Rigrod analysis*. A simpler methodology due to Schwamb is clearer.[14] (Note that similar formulations have been used by many researchers for many years, but few have appeared in print, and none in the simple form of Ref. 14.) This analysis considers the available gain and the losses as well as the saturation intensity to calculate an extraction efficiency. Losses within the resonator are not simple power losses because of the interaction between the gain and the losses. Extracted intensity can be calculated by

$$I_{\text{ext}} = g_0 I_{\text{sat}} L \left(1 - \frac{g}{g_0} \right) , \tag{5.2}$$

where

g_0 = small-signal gain
L = length of the gain medium
I_{sat} = medium-specific saturation intensity,

and

$$g = \frac{1}{L} \ln \left[\frac{1}{(R_1 R_2)^{1/2}} \right] , \tag{5.3}$$

where R_1 and R_2 are the reflectivities of the resonator end mirrors. Equation (5.3) applies for standing wave oscillators. A traveling wave (ring) oscillator
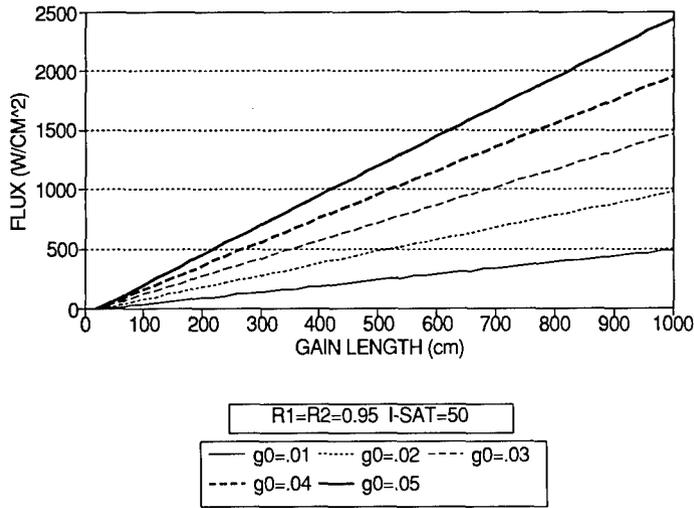
Fig. 5.5 Extracted flux from oscillator.

can be described by a similar equation, but without the square root in the denominator of the logarithm. Figure 5.5 shows extracted flux as a function of various gain and resonator parameters.

5.5.1.3 Amplification. Two general types of directed energy systems use amplification outside of the oscillator or resonator. The first type, a master oscillator power amplifier (MOPA) configuration, uses an oscillator that is as large as possible, given the ASE limit, followed by an amplifier subject to the same limit.[a] The initial resonator power calculation determines the amount of energy extracted from the oscillator. The second type of system (for example, laser fusion systems) uses a series of amplifiers. Limitations on pumping the gain restrict these amplifiers in size and shape. Typical amplifiers come in the form of either a rod or a disk. In pulsed systems, the gain and the input pulses must match temporally and spatially.

The available energy from the amplifiers is determined by the available gain and the magnitude of the irradiance input to the amplifier according to

$$I_{\text{ext}} = I_{\text{in}}(e^{gL} - 1) \; , \tag{5.4}$$

where

$$g + \frac{1}{L}\frac{I_{\text{in}}}{I_{\text{sat}}}(e^{gL} - 1) = g_0 \; . \tag{5.5}$$

For example, assume that the entering intensity is just at the saturation intensity, L is 100 cm, and the small-signal gain is 0.02 per centimeter. The

---

[a]The oscillator power is sometimes reduced to allow a better quality oscillator. Examples include injection locking.

value of g is 0.00792. The output intensity is 1.21 times as large as the input intensity. This factor is the *stage gain,* or net increase factor due to the amplifier stage.

5.5.2   Beam Quality

The six main categories of BQ effects (Fig. 5.2) are wavefront error, boresight/drift, jitter, diffraction, propagation, and other (miscellaneous) effects. All but propagation are important in the beam formation system. Wavefront error dominates, but diffraction effects are also important.

5.5.2.1   Wavefront Error

Wavefront error in the beam formation system is dominated by two primary sets of aberrations: those of the gain medium and those of the resonator that extracts energy from the gain medium. A variety of individual effects contribute in each of these categories.

Gain Medium Aberrations.   Several effects cause aberrations in the gain medium: flow-induced density variations in the medium (phase aberrations), variations in the index resulting from gain (resonant dispersion), mixing nonuniformities, and thermal density variations. Depending on the type of gain medium, optical path difference (OPD) effects can be significant. For a low-pressure gas laser system (low gas density), optical path variations as a result of density gradients cause small effects. For higher pressure gas lasers, solid-state lasers, or glass lasers, large variations occur. In some amplifiers, such as flashlamp-pumped disks in a laser fusion system, variations in energy deposition cause variable heating. The resulting wavefront error comes from thermal deformation of the medium or density gradients induced by the energy deposition.

The optical path difference as a function of medium density variation is

$$\mathrm{WFE} = \frac{1}{2\sqrt{3}} \frac{n-1}{\lambda} \frac{\Delta\rho}{\rho} L \quad [\text{rms waves}] \;, \tag{5.6}$$

where

$\Delta\rho/\rho$ = peak-valley variation in density
$n-1$ = average index of refraction variation
λ = wavelength
L = path length,

and the leading numerical factor $[1/(2\sqrt{3})]$ converts peak-to-valley (P-V) variation to rms. Figure 5.6 shows WFE as a function of the length-density product $(\Delta\rho/\rho)L$ and the index wavelength $(n-1)/\lambda$.

Resonator Aberrations.   Many sources of resonator aberrations exist: mirror thermal distortion, mirror fabrication error, mirror misalignments, coating phase aberrations, and coating dispersion. The magnitude of phase aberrations is strongly affected by the mode-medium interaction, or the effect of multiple pass gain. A *loaded cavity* (with gain) is affected less by an aberration of a given strength than is the corresponding resonator without gain.
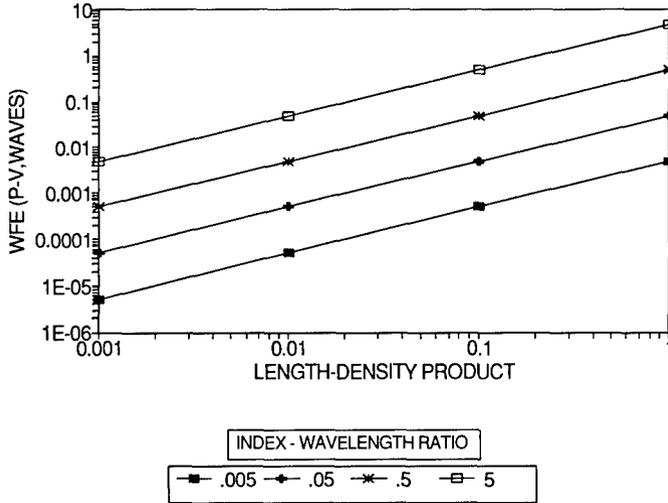
Fig. 5.6 Gain medium OPD effects in gas lasers.

Resonator aberrations can have a dominating effect on overall beam quality. The optical mode passes through the resonator many times. The resonator mirrors superimpose their own distortion on the resulting phase front. In addition, misalignment of the resonator mirrors can cause beam quality to be less than perfect. Coatings on the mirrors can also be nonuniform, leading to further OPD effects.

Resonator misalignment effects are simple in two-mirror resonators but become more complicated for multiple-mirror resonators. Detailed geometric or physical optics models are appropriate for analysis of such complex resonators. Severe misalignment of a two-mirror resonator can cause a wavefront error, but generally displaces the optical axis until a "critical angle" is reached at which output power goes to zero. The critical angle is defined as the misalignment angle at which the optical axis reaches the edge of the smallest resonator mirror. For a positive branch confocal unstable resonator, the critical angles for a primary mirror are[15]

$$\theta_{\rm PM} = \frac{r_{\rm sm}(M - 1)^2}{4ML} \;,$$ (5.7)

where

$r_{\rm sm}$ = radius of the small, convex, or secondary mirror
M = resonator magnification
L = resonator length.

For the secondary mirror,

$$\theta_{\rm sm} = \frac{r_{\rm sm}(M - 1)^2}{2L(M + 1)} \;.$$ (5.8)

If both mirrors are misaligned by an equal amount,

$$\theta_{\text{both}} = \frac{r_{\text{sm}}(M - 1)^2}{2L(3M + 1)} \quad . \tag{5.9}$$

For a negative branch resonator (concave-concave),

$$\theta_{\text{both}} = \frac{r_{\text{sm}}(M + 1)^2}{2L(3M - 1)} \quad . \tag{5.10}$$

Figure 5.7 shows the critical angle for two-mirror resonators as a function of magnification. The effective secondary mirror size is often determined by an output coupler or "scraper." The geometric parameters of the resonator must be chosen with care. In general, an unstable resonator with a magnification approaching 1 (i.e., a Fabry-Pérot resonator) has increasingly difficult alignment tolerances. A higher magnification resonator has equivalently lower misalignment errors, but does not outcouple power as efficiently.

Although negative branch resonators have lower alignment sensitivities than positive branch resonators, their internal mode geometry makes them less efficient. Negative branch resonators lead to higher mirror flux loadings and poorer geometric coupling to the gain medium.

Misfigure of the resonator optics as a result of fabrication directly affects the output quality. Most resonator mirrors are specified by their rms surface quality. The resulting effect on output phase error depends on the resonator configuration. Figure 5.8 shows typical misfigure errors. There is no simple formula to determine misfigure (or polishing) error. Before computer-controlled polishing, a number of practical equations representing particular mirror fabrication processes were in vogue. Now most surfaces can be polished to good quality, to a diameter as large as 8 m and an optical speed of nearly $f/1$, to
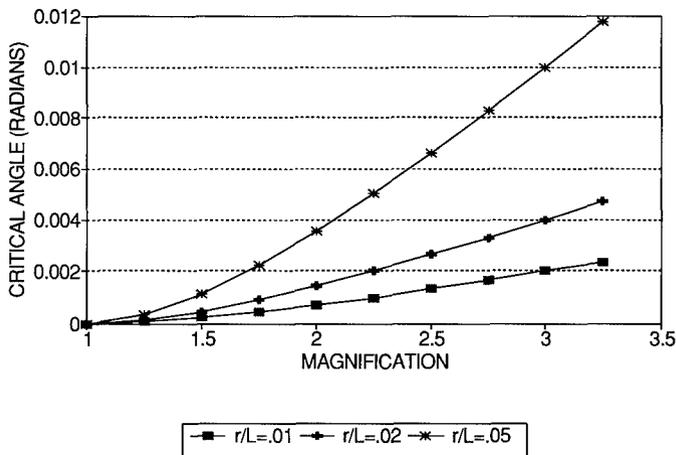
Fig. 5.7   Resonator critical angle for two-mirror confocal positive branch.
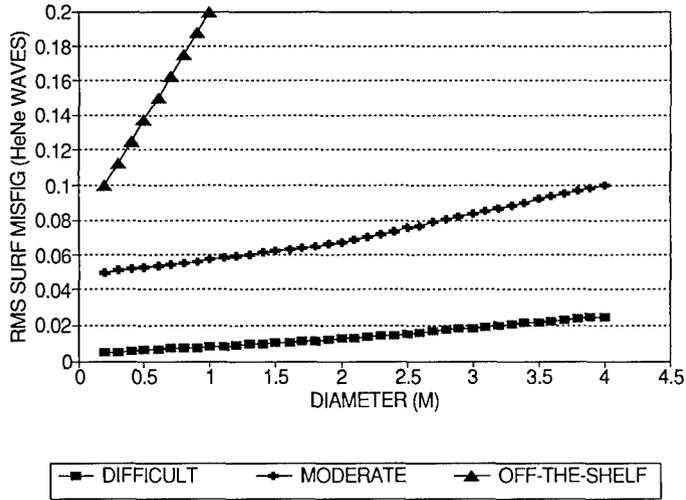
Fig. 5.8   Misfigure surface error effects (2 × for WFE).

first order, although at substantial cost. The three curves represent best performance to date (difficult), moderate difficulty, and off-the-shelf hardware. Larger misfigures are certainly possible, but most HEL systems start with "interferometric" quality optics and better.

The Hubble Space Telescope is an example of the lower curve. The Hubble primary mirror remains one of the most precise mirrors ever fabricated at large size (despite an undesired spherical error).

Coatings can be nonuniform in their thickness. This thickness nonuniformity is equivalent to a mirror fabrication error. The availability of coatings with 1% or better thickness tolerances can reduce the magnitude of such effects (particularly for thin coatings). However, for multiple-layer stacks with relatively poor thickness control processes, this effect can be a significant one.

All random (uncorrelated) variations in surface figure can be combined as a single root-sum-square (RSS) value. If the variations are spatially correlated, they are directly summed. For uncorrelated errors,

$$\delta_{\text{rms}} = \left[\sum\nolimits_n (\delta_1^2 + \delta_2^2 + \dots + \delta_n^2) \right]^{1/2} , \tag{5.11}$$

where the δ_n are the individual rms errors. If the errors are correlated,

$$\delta_{\text{rms}} = \sum\nolimits_n (\delta_1 + \delta_2 + \dots + \delta_n) . \tag{5.12}$$

For example, consider five errors each of magnitude 0.1 waves. If the errors are uncorrelated, Eq. (5.11) gives a net error of 0.22 waves. For correlated errors, Eq. (5.12) yields 0.5 waves. For equal size uncorrelated errors,

$$\delta_{\text{rms}} = \sqrt{n}\, \delta_1 , \tag{5.13}$$

where δ_1 is one of the errors. This relationship can be used early in the budgeting process when most errors are assumed to be of comparable size.

The transformation to equivalent resonator effects depends on the resonator. A useful way to determine the relative effects starts by calculating the equivalent number of resonator passes (the number of times the Fresnel core must transit the resonator before it expands to fill the output aperture). For each resonator pass, each resonator mirror affects the beam once. Although only that portion of the mirror filled by the mode, on that pass, contributes to the wavefront error, the full mirror effect is included to be conservative. The number of equivalent resonator passes is

$$N = \frac{\ln(a/2a_{\text{core}})}{\ln(M)} \; ,$$

(5.14)

where M is the magnification, a is the resonator radius, and

$$a_{\text{core}} = (\lambda L)^{1/2} \; ,$$

(5.15)

where L is the resonator length (mirror separation) and λ is the wavelength.[16] The equivalent wavefront error is

$$\text{WFE}_{\text{res}} = 2\sqrt{N}(\delta_1^2 + \delta_2^2)^{1/2} \; ,$$

(5.16)

where δ_1 and δ_2 are the resonator mirror surface errors. The factor of 2 in Eq. (5.16) converts mirror surface error to wavefront error.

5.5.2.2   Boresight/Drift

Boresight and drift are static and low-frequency (less than 1 Hz) variations in the output pointing direction. Higher frequency variations are called *jitter*. Boresight and drift effects, although small, appear in the gain medium. Larger effects occur in the resonator.

Gain Medium.   Gain media are subject to effects such as density and gain gradient beam steering. Such an effect occurs because the density in one portion of the gain medium is either higher or lower than that in another region. These effects are typical, for instance, of chemical lasers, in which the density is higher near the nozzle and lower away from the nozzle. The resulting effective wedge of optical thickness causes beam steering similar to that of a glass prism. The magnitude of the effect depends on the strength of the density gradient. Because the effect does not exist until the medium is established, the initial appearance of the effect causes a boresight error. The error is

$$\theta_{pv} = \frac{n-1}{r} \frac{\Delta \rho}{\rho} L \; ,$$

(5.17)

where

r = beam radius
n = index of refraction

$\Delta\rho/\rho$ = density variation
L = path length.

For $n - 1 = 1 \times 10^{-6}$, a density variation of 1%, a path length of 1 m, and a beam radius of 10 cm, the steering error is 1×10^{-7} or 100 nrad.

Resonator. Resonator boresight effects occur primarily because resonator alignment does not perfectly determine the optical axis. This difference between the predicted or initial alignment and the initial pointing of the beam is the dominant effect in boresight errors. Such errors cannot readily be assessed analytically. These errors can range in magnitude from a few microradians up to several tens of microradians, depending on the accuracy of alignment techniques and the complexity of the resonator. An estimate of the magnitude of boresight error can be obtained by substituting the expected misalignment of the individual resonator mirrors into the equations for resonator misalignment. For a two-mirror resonator,

$$\theta_{beam} = \frac{R_{PM}\theta_{PM}}{L} , \tag{5.18}$$

where R_{PM} is the radius of curvature of the larger concave mirror (PM), θ_{PM} is the misalignment angle of that mirror, and L is the mirror separation (or resonator length). The effect of the other mirror is

$$\theta_{beam} = -R_{SM}\theta_{SM} , \tag{5.19}$$

and the resonator length is

$$L = \frac{R_{PM} + R_{SM}}{2} , \tag{5.20}$$

where R_{SM} is the radius of curvature of the small convex mirror (SM) and θ_{SM} is the misalignment angle of the SM. Use of theodolites and good quality optical references can align larger systems to individual accuracies of 5 to 10 μrad. For a typical system with $L = 1$ m, $R_{PM} = 4$ m^{-1}, $R_{SM} = -2$ m^{-1}, and a 5-μrad initial misalignment of each mirror, the two-mirror-resonator boresight error is

$$\theta_{beam} = \frac{4(5 \times 10^{-6})}{1} = 20 \ \mu\text{rad} \quad \text{(PM)} , \tag{5.21}$$

$$\theta_{beam} = -(-2)(5 \times 10^{-6}) = 10 \ \mu\text{rad} \quad \text{(SM)} . \tag{5.22}$$

The output errors are larger than the individual mirror errors. Such an amplification is a cause for concern and requires very tight resonator alignment requirements in most systems.

Drift effects in the resonator are caused primarily by relatively low temporal frequency thermal variations, which can cause significant tilts. A thermal analysis code can determine the drift magnitude. Equations (5.21) and (5.22) can then be used to convert the individual drifts into a net output drift.

5.5.2.3 Jitter

Gain Medium. Jitter as a result of gain medium effects is small for most lasers because of the lack of mechanisms for maintaining a fluctuating gradient. The drift equation [Eq. (5.17)] can be used if a dynamic gradient exists, where the density variation is then a function of time:

$$\frac{\Delta\rho}{\rho}(t) \ .$$

(5.23)

Resonator. Jitter resulting from the resonator can be significant. As shown in earlier equations, the resonator output direction depends on the relative tilt of the internal resonator mirrors. The same equations yield jitter when rms jitter for individual mirrors is used. If both mirrors move simultaneously and the angles are small,

$$\theta_{\text{out}} = 2\theta_{\text{PM}}\left(\frac{M}{M-1}\right) \ ,$$

(5.24)

where the magnification M is

$$M = -\frac{R_{\text{PM}}}{R_{\text{SM}}} \ .$$

(5.25)

Figure 5.9 shows the magnitude of these effects for typical resonator mirrors. Note that a Fabry-Pérot (plano-plano) resonator has a magnification of 1 and thus an infinite output misalignment. In practice, this means that such a resonator must be aligned to better than the diffraction angle (λ/D).
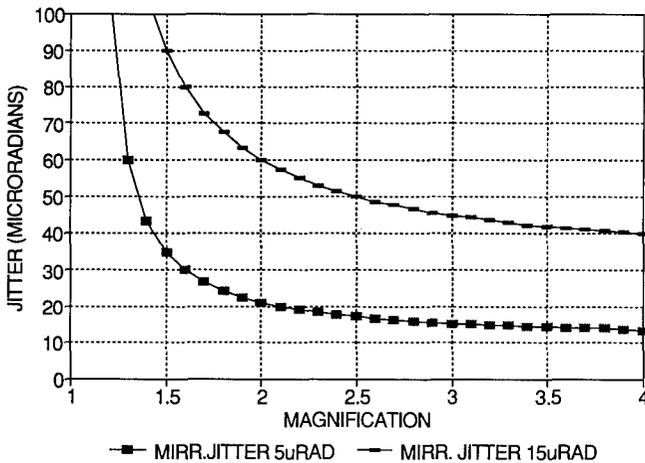
Fig. 5.9 Resonator output jitter.

5.5.2.4   Diffraction

Diffraction determines the fundamental mode properties of the resonator. Lasing occurs because the resonator provides a low-loss path for amplifying spontaneous emission from the gain medium. The lasing process starts with photons emitted near the optical axis. The region of the gain contributing to the initial lasing is the Fresnel core, defined previously. As the Fresnel core expands to fill the mode, diffraction determines the basic structure of the mode. Because most gain media are nonuniform, the diffracted mode is modified by the gain interaction. The resulting mode combines both gain and diffraction effects. Lossier higher order spatial modes can coexist with the fundamental resonator mode if gain is high enough. Resonator designers must consider the mode separation and try to suppress higher order modes. Some waveguide lasers use multiple modes as a means of extracting more energy, but the beam quality is poorer. The resulting mode properties determine a net intensity as well as a net phase. In a simple resonator, the lowest loss mode may be a Gaussian intensity with a flat phase. In more complicated resonators, the mode may have a complex intensity profile and the phase may be slightly aberrated. Propagating this mode to the far field results in an "intrinsic" beam quality, exclusive of all other effects. Typical intrinsic beam qualities range from 1.001 to 1.03.

5.5.2.5   Other Beam Quality Effects

Although the diffractive and gain properties determine the fundamental mode, that mode may be somewhat different for each lasing line in a multiline laser. The differences result primarily from the spatial gain variations for each wavelength. These variations differ substantially from laser to laser. Single-wavelength lasers such as Nd:YAG are not affected. Chemical lasers have strong variations that lead to *partial aperture lasing*. Each wavelength partially fills the device output aperture. Because the individual wavelengths do not use the full aperture, they have an equivalently smaller beam diameter, a larger far-field spot, and lower energy density in the far field. Figure 5.10
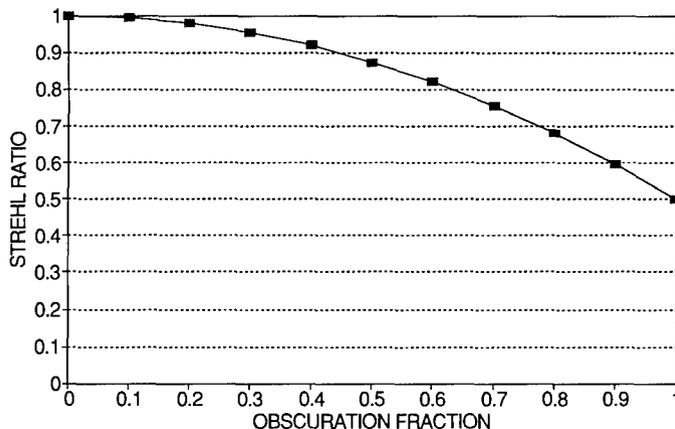
Fig. 5.10 Partial aperture lasing—obscuration of one wavelength in a two-wavelength laser.

shows the Strehl ratio for a laser with two wavelengths. Each wavelength has 50% of the total power, but different obscurations. In this example, one wavelength is obscured, whereas the other is not. Note that each wavelength focuses to a slightly different spot size even if all fill the aperture.

5.5.3   Wavelength

The output wavelength of the DE system directly affects the spot size of the beam and can also indirectly affect other parts of the DE system that are wavelength sensitive. The line center wavelength is determined by the lasing transition, but the wavelength may be pulled slightly by gain effects. The linewidth may be pressure or Doppler broadened. The output wavelength is also affected by resonator longitudinal mode structure. A typical HF laser gain profile is 300 to 400 MHz wide (full width at half maximum). If only a single longitudinal mode is operating under the gain curve, that mode may be anywhere in the frequency width of the gain, with the corresponding wavelength variation. Most long laser resonators have many modes operating simultaneously, because the mode spacing is $c/2L$ (c/L for a ring resonator). With many modes running, the average output frequency variation cannot exceed one-half of the mode spacing. For a ring resonator of length 25 m, the mode separation is 12 MHz. Depending on saturation, 25 to 33 modes operate simultaneously. The net frequency variation is one half the mode spacing, or 6 MHz.

5.6   BEAM TRANSFER SYSTEM

The beam transfer system (BTS) is often of unity magnification (assumed here). Intermediate levels of magnification sometimes occur, particularly where there are fabrication constraints on expensive elements such as fast steering mirrors (FSMs) or deformable mirrors (DMs). The transfer system may correct the output jitter, correct the wavefront, prevent decentration (or beam walk), provide for optical retargeting, and move the beam through the DE system's physical constraints. If the beam expander includes a gimbal, the BTS must include gimbal-follower mirrors.

5.6.1   Power

Power effects in the beam transfer system are simple to calculate. These effects include absorption, scatter, clipping, and diffraction. The effect of amplifiers in the beam transfer system must include the incident energy and the saturation characteristics of the amplifier (see Sec. 5.5.1.3).

5.6.1.1   Absorption.   Absorption losses can be calculated quite simply by multiplying together 1 minus the individual absorption of the various mirrors in the train, and subtracting the result from unity. Use care in assuming that absorption is 1 minus the reflectivity because this is only true for mirrors with relatively large absorptions. High-performance optics have relatively small absorption, both in the coatings and in the bulk material. Low scatter resulting from excellent surface quality and low absorption very often are achieved only simultaneously. Figure 5.11 shows ranges of typical reflectivity values for a
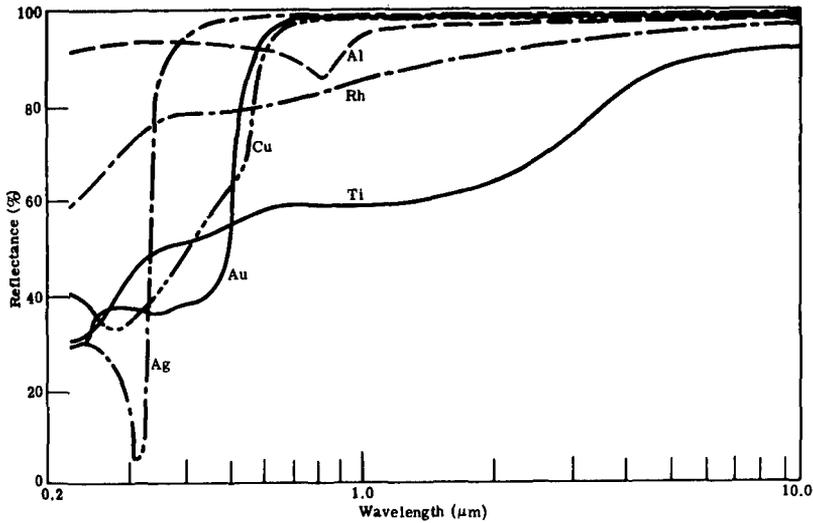
(a)
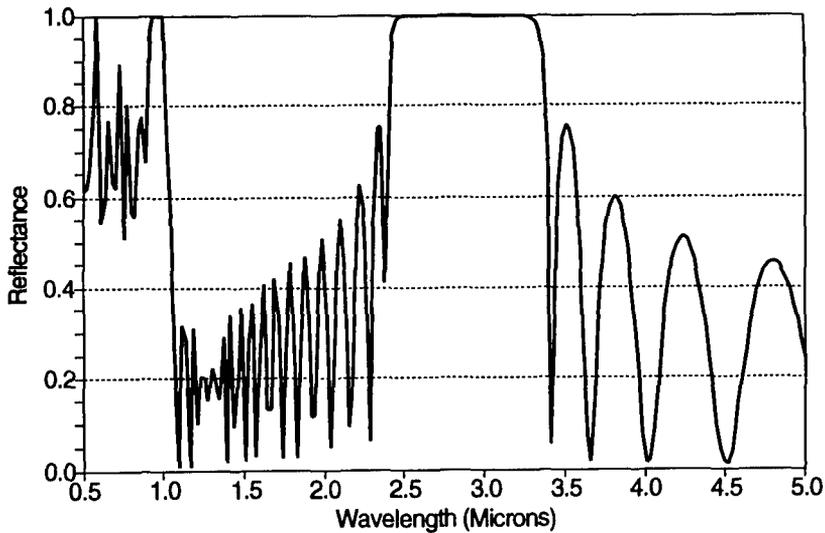
(b)

Fig. 5.11   (a) Reflectivity for various metals (Ref. 18, pp. 7–81) and (b) example of a prediction of a broadband 99% reflectivity multilayer dielectric mirror coating.

variety of bare metals and multilayer dielectric (MLD) coatings used in directed energy systems. MLD technology can achieve very high reflectivities (> 0.999).

Beam path absorption causes power loss, depending on the concentration of absorbers. Even in a vacuum chamber, a laser system can suffer appreciable absorption losses if strong absorbers such as ground-state lasing species are present. The intensity loss ratio for absorption is[17]

$$\frac{I_{\text{out}}}{I_{\text{in}}} = \exp(-0.01 L \alpha_T) \ , \tag{5.26}$$

where

I_{out} = output power after the absorbing path
I_{in} = input power entering the absorbing path
L = path length in kilometers
α_T = total absorption of all species in percentage per kilometer.

The term α_T is a summation of other individual absorption terms. For example, for HF propagation in air, CO_2 and H_2O are strong absorbers. Typical values for a laser line under varying conditions are given[18] in Sec. 5.17.3.

5.6.1.2 Scatter. Scatter effects on power are typically simpler to calculate than the distribution of scattered energy. The distribution is important because of its effect on adjacent structures within the directed energy system. The simplest means of calculating scatter uses the total integrated scatter (TIS) as shown in the following equation[19]:

$$\text{TIS} = \left(\frac{4\pi\delta \cos\theta}{\lambda}\right)^2 , \qquad (5.27)$$

where

δ = rms surface roughness
θ = angle of incidence
λ = wavelength of the incident light.

Total integrated scatter is all energy lost to the beam as a result of scattering from the surface into the hemisphere in front of the mirror. This parameter is valid in the region of rough to moderately rough materials and is not necessarily valid for superpolished surfaces with rms roughnesses of the order of 1 to 10 Å. Figure 5.12 shows typical total integrated scatter as a function of roughness.
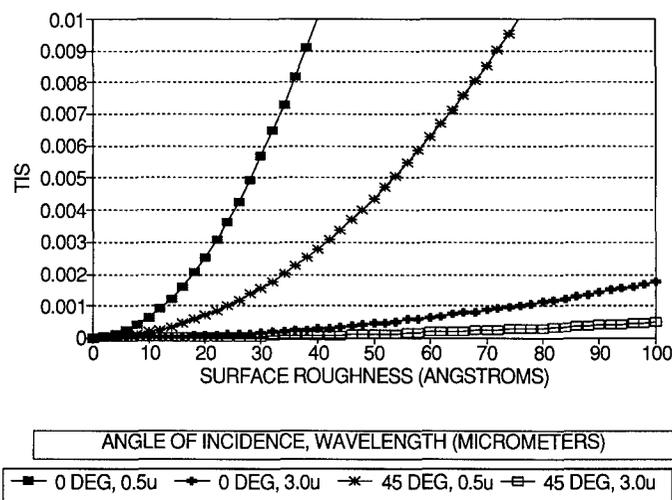
Fig. 5.12 Total integrated scatter.

Scatter is a very directional phenomenon, usually characterized by a bidirectional reflectance distribution function (BRDF). The TIS is only useful as an approximate loss calculation. Do not use it for design of power management equipment. Diamond-turned optics evidence correlated scatter effects, called *near-angle scatter*. The scatter results from a grating effect caused by the turning grooves. As the name suggests, the scatter is confined to a relatively narrow range of angles near the beam propagation direction.

5.6.1.3 Diffraction/Clipping.

Diffraction losses can be an important power loss effect in the beam train, depending on the length of the beam train and the ratio of diameter to length. The Fresnel number of the propagation is an approximate guide to determine whether a significant amount of diffraction can be expected. The collimated Fresnel number is

$$N_c = \frac{r^2}{\lambda L} ,$$

(5.28)

where

r = beam radius
λ = wavelength
L = propagation distance.

For a noncollimated beam, r^2 in Eq. (5.28) can be replaced by $r_1 r_2$, where r_1 and r_2 are the radii at each end of the propagation. A Fresnel number of 5 or less implies significant diffraction. The far-field condition for a circular aperture is

$$L \gg \frac{\pi r^2}{\lambda} \quad \text{or} \quad N \ll \frac{1}{\pi} .$$

(5.29)

Figure 5.13 shows the diffraction patterns for circular and rectangular apertures for various Fresnel numbers.[20]

Clipping losses in the beam transfer system (exclusive of clipping of diffraction on the outer diameter) can be characterized by an area ratio loss for beams that are reasonably uniform in irradiance profile. Nonuniform beams can be characterized by an irradiance weighted loss. The power loss resulting from obscuration is

$$P_{\text{loss}} = P_0(1 - \varepsilon) ,$$

(5.30)

where P_0 is the total power in the unobscured beam and ε is the area ratio,

$$\varepsilon = \frac{A_o}{A} ,$$

(5.31)

where A_o is the obscured area and A is the total beam area. The effective beam quality loss is discussed in the following sections. A power loss and a beam quality loss should not be assigned to the same obscuration (see following discussion).
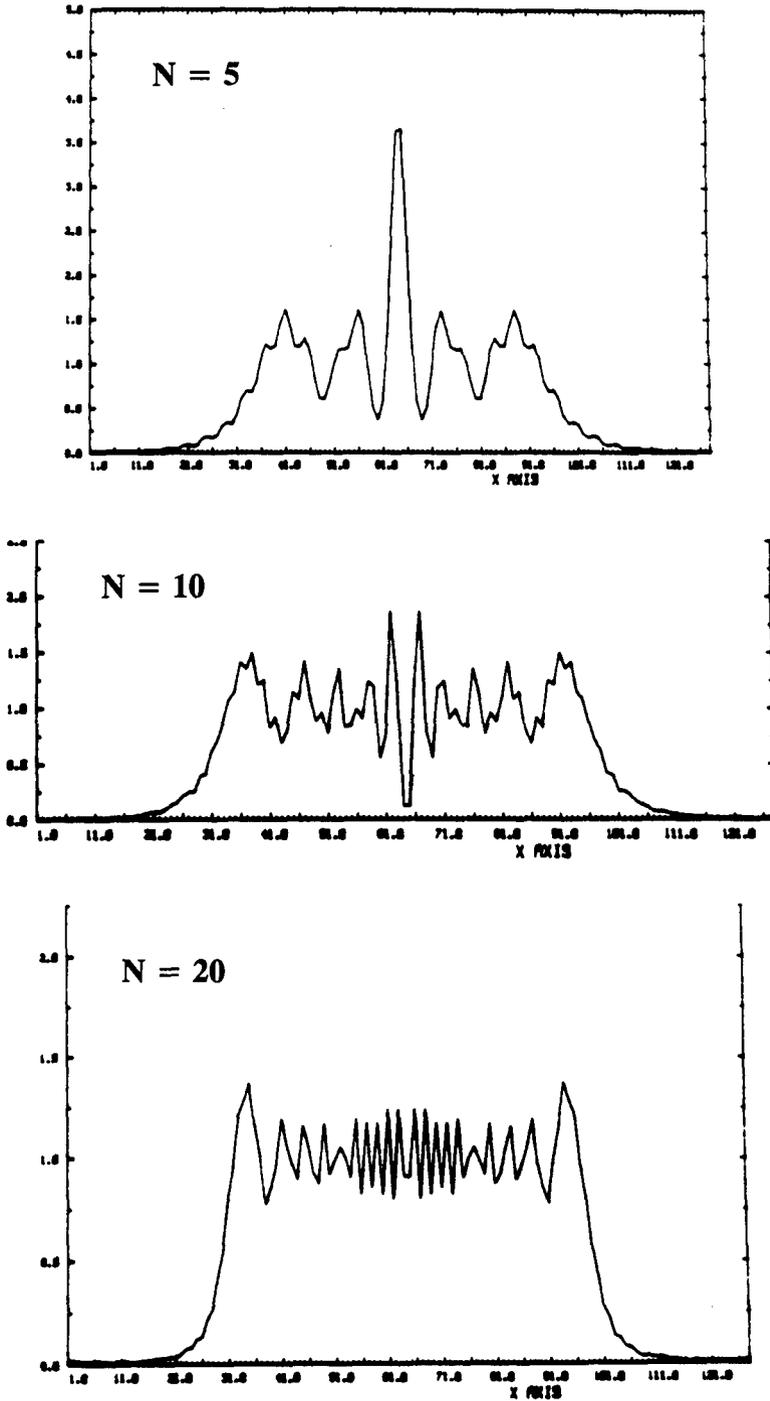
Fig. 5.13 Near-field intensity at three different Fresnel numbers (propagation distances) for a uniformly illuminated circular aperture (scales arbitrary).

5.6.2   Beam Quality

Beam quality dominates the performance of the directed energy system. Because irradiance on target is inversely proportional to the square of beam quality, even relatively small beam quality effects can cause dramatic reductions in system performance.

5.6.2.1   Wavefront Errors.

Wavefront errors in the beam train are somewhat easier to calculate than those within the resonator because each optical surface reflects or transmits the beam only once. This section discusses the wavefront effects resulting from mirrors (or lenses and windows), metric errors (misalignments, etc.), the beam path medium, and other wavefront-related effects, such as sampling and sensing.

Mirrors.   The primary cause of distortion in mirrors is either fabrication error or thermal distortion as a result of absorbed energy. Zernike polynomials or other aberration polynomial sets are useful in characterizing distortions. Coating OPD effects also affect the wavefront error. For all of these effects, double the magnitude of a surface distortion when calculating wavefront error (normal incidence).

Mirror thermal distortion consists of two separable effects: thermal bowing and beam mapping. Thermal bowing is the gross curvature of the optical surface resulting from absorbed energy and the mirror's constrained edges. Beam mapping is the out-of-plane surface response to local variations in absorbed energy. Complicated models of thermal distortion provide detailed calculations of the thermostructural response of various materials. For conceptual designs, two simpler formulas have been developed. Thermal bowing is described by

$$\text{WFE}_{\text{bow}} = \frac{4\pi}{\lambda} \xi_{\text{bow}} \frac{r^2}{25} \alpha P_0 \ , \tag{5.32}$$

where

ξ_{bow} = thermal bowing coefficient (2×10^{-9} to 9×10^{-9} cm/W)
r = mirror radius in centimeters
λ = wavelength in centimeters
α = mirror absorptivity
P_0 = incident power in watts.[21]

This equation assumes that the centroid of the beam is at the center of the mirror. If r is replaced by $r - r'$, where r' is the centroid location radius, the equation still holds. This thermal distortion is derived from the performance of cooled metal mirrors. Thermal bowing is primarily a focus error at normal incidence. At an angle, the error introduces some astigmatism.

Thermal mapping is a localized distortion of the mirror because of a local nonuniformity in irradiance:

$$Z_{\text{map}}(x,y) = R_t R_m \xi_m \alpha I(x,y) \ . \tag{5.33}$$

The associated wavefront error is

$$\text{WFE}_{\text{map}} = \frac{2\pi}{\lambda} \frac{\displaystyle\iint 4(Z_{\text{map}} - Z_{\text{ave}})\, dx\, dy}{\text{area}} \,, \tag{5.34}$$

where

ξ_m	=	thermal mapping coefficient, typically 5 to 100 Å W^{-1} cm^{-2}
area	=	area of the beam
$I(x,y)$	=	incident flux in watts per square centimeter
Z_{ave}	=	average piston component of the distortion
R_t	=	thermal diffusion function
R_m	=	structural spreading function.[22]

The structure of the mirror and its thermal properties strongly influence the mapping. If the two R above are 1, beam irradiance maps directly (thermal conductivity very low and the structure relatively flexible). For a conductive surface (metal mirrors), the R terms are a four-parameter double Gaussian.

Metric Errors. Because the beam transfer system is normally a unity magnification system, misalignment effects do not typically cause beam quality aberrations. If an interim telescope is used, alignment effects within that telescope affect the beam quality. These effects are equivalent to those discussed in the beam expander section.

Beam Path Medium. Density gradients in the medium in the BTS cause wavefront errors. Index mismatching at interface planes can cause large effects. Thermally induced wavefront effects resulting from beam absorption can include thermal lensing. The beam formation section described beam path effects.

Other Effects. Most modern directed energy systems use control systems to align, reduce jitter, or clean up beam phase aberrations. Control system analysis for DE systems is a topic deserving of a handbook chapter by itself. Three types of control systems are typically used in DE systems: focus control, tilt control, and adaptive optics control of higher order phase aberrations. This section gives some approximate rules for determining correctability of aberrations, but does not deal with the specifics of the control systems themselves. Adaptive optics systems are described elsewhere in this Handbook. The budgeting methodology described in Secs. 5.10 to 5.16 should be followed whether correction systems are used or not. Two budgets, one for uncorrected performance and the other for corrected, should be constructed. The uncorrected budget is useful in determining control system dynamic range.

    If the directed energy system has sampling and sensor systems, the errors inherent in these systems must be included in the error budgets. Many beam transfer systems include linear gratings, often in a grating rhomb configuration, to sample the beam. In addition to the normal aberrations associated with the mirror surfaces, these elements are subject to relative misalignment

aberrations. Absorbed energy can cause further errors if the grating pattern is distorted. Grating distortion causes a pointing error:

$$\theta_{GR} = \frac{n\lambda}{d\cos\theta}\left(\frac{\Delta d}{d}\right) , \tag{5.35}$$

where

   d = grating spacing
   Δd = change in grating spacing caused by distortion
   n = diffracted order
   θ = diffracted angle.

Sensing systems include a variety of errors dependent on the type of sensor used. The sensor optical train affects the sensed beam. Because the control system will try to correct all sensed errors, errors in the sensor can actually induce errors in the main beam path. Other errors include sensor algorithm errors, aliasing of spatial and temporal errors, and undersampling. DE system errors are often divided into common-path and non-common-path terms. Common-path errors affect both the transmitted and sensed beams and can be corrected by a control system. Non-common-path errors affect only one of the paths. If the error occurs only in the transmitted path, it is not corrected. If the error occurs only in the sensed path, a control system may impose the error on the transmitted beam, trying to "correct" the sensed beam.

   Control systems can induce errors. These control system errors typically arise because the control system amplifies errors outside of either its temporal or spatial bandwidth. Net performance improves because the control system corrects large beam errors while inducing smaller control errors. The control and sensor system may also not provide an accurate representation of the beam. Such errors must be taken into account when establishing the initial uncorrected performance of the beam as determined by a sensing system. In general, all aberrations can then be divided into correctable or uncorrectable and a correctability factor calculated.

5.6.2.2 Boresight/Drift.
Boresight and drift are not usually major concerns in the beam transfer system because of the relative ease of alignment and of the availability of convenient locations for beam limiting apertures (clippers). If an autoalignment system is available, drift is minimized. If not, structural and fabrication concerns arise. Even in a controlled system, thermal effects may be of concern.

Thermal Effects. Any thermal error that results in a beam tilt will cause beam drift. A tilt is caused by a net irradiance ramp. Most optical elements respond slowly enough that their tilt component of phase distortion is considered drift. The tilt term is

$$\theta_t = \frac{qI(y)\alpha\xi}{r} , \tag{5.36}$$

where

$I(y)$ = one-dimensional tilt component, in the y direction, of the
        intensity in watts per square centimeters
α = absorptivity of the mirror
ξ = thermal distortion coefficient of the mirror in $\text{Å W}^{-1}\text{ cm}^2$
r = radius of the mirror
q = unit conversion factor in angstroms per radius.

The tilt component of the intensity is simply the best straight-line fit ramp from one edge to the other. The total intensity is irrelevant because it simply causes a piston error.[b] If the optical element distortion varies as a function of position because of the details of construction of the mirror, then a more complicated functional form of ξ must be used.

Even if an alignment system removes drift, the alignment system may not accurately represent the average tilt of the beam. If an auxiliary beam is used that is considerably smaller than the main beam, the alignment beam may sample a local tilt that is completely different from the average tilt. The simplest example is an alignment beam propagating outside of the HEL clear aperture. The region of the mirror used by the alignment beam is colder (less distorted) and does not represent the average (distorted) pointing direction of the HEL beam.

Boresight errors in the beam transfer system occur as a result of structural changes in the system between the time it is aligned and the time the HEL operation begins. Careful design and operational practice include stiffening the system to minimize motion and performing final alignment immediately before the beam is activated.

5.6.2.3 Jitter.

The beam transfer system is usually a significant contributor to system jitter. It has the majority of optical elements in many systems, and it spans the most distance in the DE system structure. System dynamics determine the relative proportion of the various effects. If cooled mirrors are used with a stiff structure and a relatively quiet environment (for example, space applications), then mirror jitter dominates. If uncooled mirrors are used in a relatively long structure subject to large disturbances (for example, a laser material processing system), then structural effects dominate. Jitter is usually spatially uncorrelated, but may be temporally correlated depending on the modal frequencies of the system. The temporal structure of the jitter is generally expressed as a power spectral density (PSD). The rms value of the PSD is the square root of the integrated PSD, and is used to budget the performance of an individual mirror.

Mirror Motions. If there are n mirrors in the beam transfer system, then the net jitter caused by the mirrors, in the beam coordinate system, is the RSS of the uncorrelated rms jitters:

$$\sigma_T = \left[\sum (\theta_{\text{rms},n}^2)\right]^{1/2} ,\tag{5.37}$$

---

[b]A piston error is a uniform out-of-plane growth of the entire mirror surface, causing no net optical path variation across the beam.
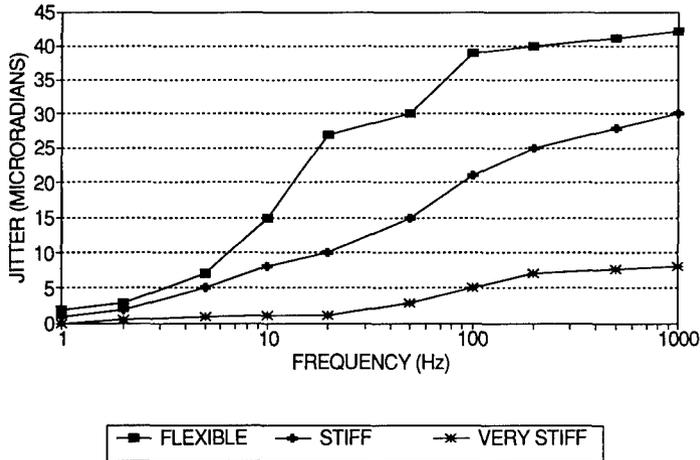
Fig. 5.14 Integrated jitter over the frequency range for three laser systems of different structural quality.

where σ_T is the total jitter due to mirrors in radians and $\theta_{rms,n}$ is the individual rms tilt of the n'th optical element. Typical rms jitter for individual elements varies widely but is generally in the range 1 to 20 μrad. Figure 5.14 shows the integral of jitter PSDs for three typical systems.

Structural Motion. The stiffer the structure, the less the jitter amplitude, but the higher the frequency of the jitter induced. All structures have low-order fundamental modes that can induce jitter in the optical system. For an arbitrary structure, the response in the x direction as a function of time is

$$y^2(x,t) = \sum_j \left[\frac{\phi_j(x_f)\phi_j(x)}{M_j\,\omega_j^2} \right]^2 \int_0^\infty S(\omega)H_j H_j^* \, d\omega \; , \qquad (5.38)$$

where $\quad M_j = \int \rho\phi_j^2 \, dv \; ,$

and $\quad H_j = \dfrac{1}{[1 - (\omega/\omega_j)^2 + 2i\tau_j(\omega/\omega_j)]} \; .$

Here

$\quad x_f \quad$ = point at which the force is applied

$\quad \phi_j(x_f)$ = modal displacement in the j'th mode in the direction of the applied force

$\quad \phi_j(x) \quad$ = mode shape

$\quad \omega_j \quad$ = frequency of the j'th mode

$\quad S(\omega) \quad$ = power spectrum of the applied force

$\quad \tau_j \quad$ = critical viscous damping ratio in the j'th mode

$\quad H_j^* \quad$ = transpose of H_j.

5.6.2.4 Diffraction Effects. Diffraction contributions to beam quality in the BTS are small. Nonuniform irradiance distributions cause an effective reduction of irradiance on aim point. Diffractive spreading of the beam can smooth an initially spiked irradiance, but can also introduce diffractive ripples in the intensity.

5.6.2.5 Propagation. Propagation in the beam train is generally not of concern except for very long paths (where reimaging can often be used to alleviate effects). Propagation from the beam expander to the target is discussed in Sec. 5.7.

5.6.2.6 Other Effects. Clipping of the beam causes a reduction in irradiance on the target. This reduction results from a combination of both power and aperture effects. Section 5.6.1.3 discusses power loss effects. The loss of a portion of the beam or the obscuration of a portion of the beam results in an effectively smaller area. Excluding the energy loss, this effectively smaller area leads to a less intense far-field spot.

Consider a specific example. The effect of a central obscuration on Strehl ratio for a focus error has been calculated by Mahajan[23] to be

$$
S_f = 1 - \left[\frac{(1 - \varepsilon^2)^2 \kappa}{2\sqrt{3}} \right]^2 \, ,
\tag{5.39}
$$

where ε is the obscuration radius (unit circle) and κ is the aberration strength coefficient. The effect of obscuration on the unaberrated beam is

$$
S_{\text{obs}} = (1 - \varepsilon^2) \, .
\tag{5.40}
$$

Section 5.13.4 discusses obscurations in more detail. The combined effect can be written as

$$
S = S_{\text{obs}} S_f \, ,
\tag{5.41}
$$

or

$$
S = \left\{ 1 - \left[\frac{(1 - \varepsilon^2)^2 \kappa}{2\sqrt{3}} \right]^2 \right\} (1 - \varepsilon^2) \, .
\tag{5.42}
$$

The resulting functional dependence is illustrated in Fig. 5.15. For weak aberrations, the Strehl falls off as the obscuration increases. For strong aberrations, the Strehl is also strongly reduced when the beam is unobscured. Some net improvement can be achieved by reducing the aperture of a badly aberrated system.

The effect of obscuration on Strehl ratio for a multiple-wavelength beam is:

$$
S_m = \sum_n F_n (1 - \varepsilon_n^2) \, ,
\tag{5.43}
$$

where ε_n is the area obscuration for the n'th wavelength, and F_n is the fractional power in the n'th wavelength. For example, for $n = 3$, $F_1 = 0.5$, $F_2 = 0.3$, $F_3 = 0.2$ and $\varepsilon_1 = 0.1$, $\varepsilon_2 = 0.2$, $\varepsilon_3 = 0.3$:
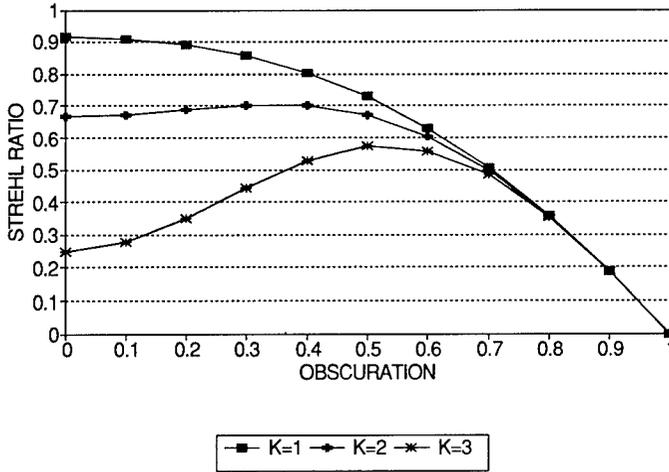
Fig. 5.15   Effect of apertures on Strehl ratio for three focus magnitudes (κ).

$$S = 0.5(0.99) + 0.3(0.96) + 0.2(0.91) = 0.965 \ . \tag{5.44}$$

Considering only the aberration, increasing the obscuration gradually reduces the aberration effect. The Strehl actually decreases rapidly because of the power loss effect. The Strehl resulting from a pure obscuration (no power loss) actually increases, because the power is effectively redistributed to the outer portion of the annulus. Power and BQ/Strehl effects are not usually separated in the literature, causing difficulty in power budgeting. Remembering that both BQ and Strehl are defined relative to "ideal" power can help keep budgets clear. Section 5.10 discusses BQ/Strehl relationships.

5.7   BEAM EXPANSION AND POINTING

This section discusses transport of the beam to the target, accomplished by a large telescope called a *beam expander*. The final focusing optic in a laser fusion system performs the same function. Propagation effects after the beam expander are also included (discussed in both the power and beam quality sections).

Beam expanders for directed energy systems have typically been two-mirror telescopes (Fig. 5.16), although three-mirror telescopes are used in wider field-of-view systems. Telescope design must balance the often conflicting requirements of magnification, primary to secondary mirror spacing, alignment sensitivity, aberrations, and primary mirror diameter. Each of these parameters is constrained by the system application. Figure 5.17 shows the variation in primary-to-secondary separation as a function of magnification, $f/\#$, and input beam diameter. Performance of the telescope is more difficult to achieve at low $f/\#$ (extreme curvatures) and high magnification. A higher $f/\#$ requires a long system.

For large directed energy systems the output aperture is a key parameter in determining brightness and is typically as large as possible. A large aperture

PM focal length f1 $Z=d\ f/\#\ (M-1)$
SM focal length -f2 $f/\#=f1/D$
M=D/d
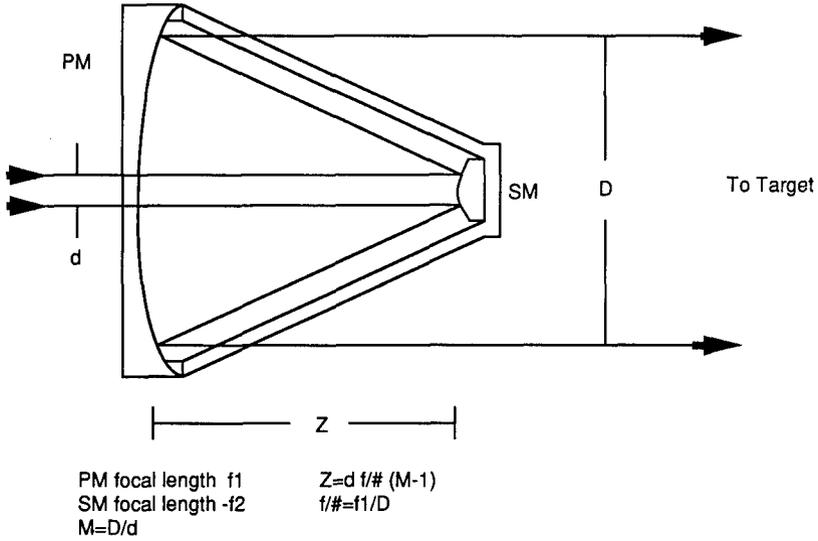
Fig. 5.16 HEL beam expander (two mirror).
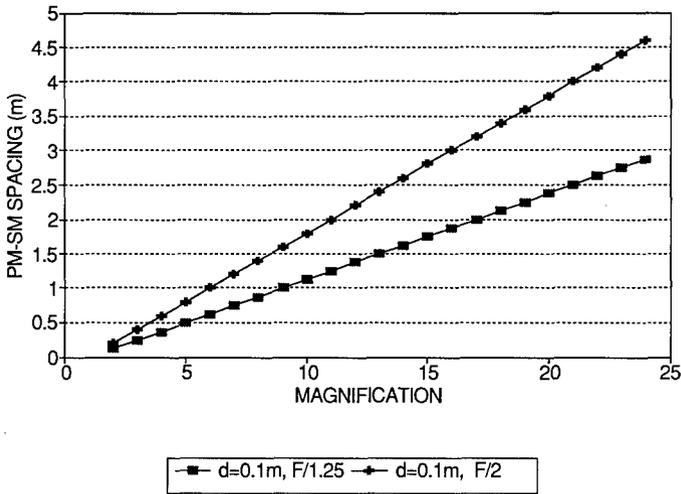
Fig. 5.17 Beam expander size.

system must be optically very fast (low $f/\#$) for the primary-to-secondary mirror spacing to be reasonable. For a slow system, the primary-to-secondary spacing is large, and the structural task of maintaining the alignment between the primary and the secondary is difficult. The relative speed of the system and the equivalent magnification determine how tight the telescope alignment requirements will be. The minimum input beam size is often determined by the distortion characteristics of the mirrors.

5.7.1   Power

Power losses in the beam expander are primarily caused by absorption, scattering, and clipping. Losses resulting from diffraction and the medium are small because of the relatively short distances that the beam travels. Losses during propagation after the beam expander result primarily from absorption in the medium between the DE system and the target, although scattering and nonlinear effects can be important, depending on aerosol density and beam intensity, respectively. Diffraction losses between the beam expander and the target are also significant (much of the transmitted energy misses typical targets because they are relatively small). If an edge-fluence requirement is placed on the target interaction, only 20 to 50% of the power leaving the beam expander may fall in the central portion of the first Airy disk.

5.7.1.1   Absorption.   Although the beam expander system (BES) has only two mirrors, the primary mirror is usually large and has higher absorption than the smaller mirrors in the system. In addition, the primary mirror may have holographic optical elements (HOEs), gratings that diffract a portion of the incident beam. The power loss due to absorption is simply:

$$P_{\text{loss}} = P_0[1 - (1 - \alpha_p)(1 - \alpha_s)] \ , \tag{5.45}$$

where

P_0 = incident power
α_p = primary mirror absorption
α_s = secondary mirror absorption.

The secondary mirror absorption should be as low as that of a beam transfer system mirror, except that very fast aspheric secondaries are difficult to polish to the excellent surface finish required for very low absorption. Low-absorption multilayer dielectric coating stacks have been successfully applied to large primary mirrors, and low absorption is now more a matter of cost than engineering in the range of absorption from 0.02 to 0.001.

5.7.1.2   Scatter.   Scatter for both mirrors in the beam expander can be treated in the same fashion as the beam transfer section mirrors. Secondary mirror scatter is usually comparable to BTS mirrors. Scatter for the primary mirror is generally higher because the surface finish is poorer. If the primary mirror is segmented, scatter may be increased by the presence of segment edges (or power management clipper edges).

5.7.1.3   Diffraction/Clipping.   The beam expander is often a complicated structure and has many obscurations that clip the beam. A symmetric two-mirror beam expander usually has a central obscuration that is determined by the size of the secondary mirror and the secondary mirror mounting structure. The secondary mirror support struts may also obscure the beam. If the primary mirror is segmented, the mirror segment gaps further obscure the beam. All of these effects can be accounted for by the total geometric obscuration area, as discussed earlier (Sec. 5.6.2.6). Diffraction is small over the limited paths found in the beam expander, but can be a dominant effect in reaching the target.

Diffraction effects during propagation can be treated exactly for a limited number of cases, and have been modeled extensively for many others. The far-field spot distribution for a uniformly illuminated unobscured circular aperture is

$$I = I_0 \left[\frac{2J_1(x)}{x} \right]^2 , \qquad (5.46)$$

where

$$x = \frac{2\pi r z}{\lambda L} , \qquad (5.47)$$

and

λ = wavelength
r = near-field radius
I_0 = irradiance at the center of the far-field spot
L = the propagation distance
z = the far-field coordinate.[24]

Equation (5.47) sometimes appears with the term z/L replaced by v, the far-field angular coordinate. The integral of the far-field diffraction pattern yields the encircled energy:

$$I(z) = [1 - J_0^2(x) - J_1^2(x)]I_0 , \qquad (5.48)$$

where the J_n are the Bessel functions of zeroth and first order.

For an unaberrated, uniform intensity beam, Fig. 5.18 shows the encircled energy as a function of far-field radius and the intensity as a function of far-
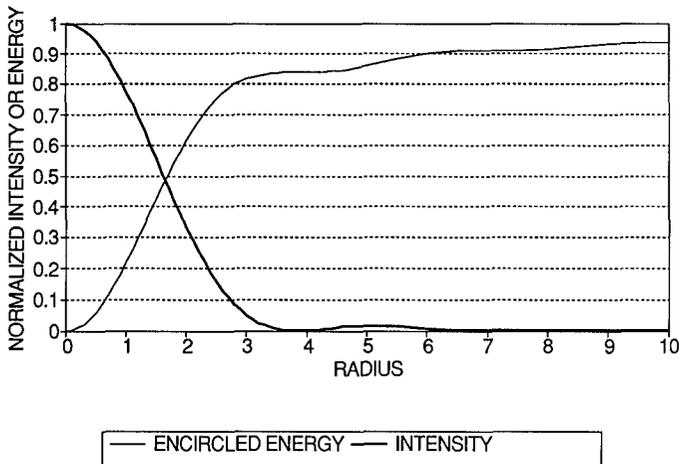
Fig. 5.18   Far-field distribution; intensity and encircled energy for a uniformly illuminated circular aperture.

field radius. At the 50% intensity point, the encircled energy is only 50% of the total. The literature contains many near- and far-field aperture distributions.[25]

5.7.1.4 Propagation Losses.

Propagation losses between the primary mirror and the target depend on the beam power, the intensity distribution, the intervening medium, and conditions within that medium. Absorption over various atmospheric paths has been the subject of previous chapters of this Handbook. Evaluating laser absorption is difficult because the fine structure of the absorption spectra are absent in low-fidelity atmospheric calculations or measurements. Section 5.17 gives tabulations of absorption for various key laser lines and conditions. Scattering from aerosols is important: most atmospheric propagation is limited to "cloud-free lines of sight." Nonlinear effects in the medium are intensity and power dependent. Extreme nonlinearities can cause beam deflections that are so severe that they are more power loss mechanisms than phase errors. Nonlinearities are also dependent on the wave form and pulse train of the beam. The cw lasers are not as susceptible to nonlinearities because of low intensity. The upcoming beam quality section discusses blooming and turbulence.

5.7.2 Diameter

The output diameter of the DE system is a key, but not a free, parameter. Minimization of the far-field spot requires a large diameter if long ranges are to be achieved. There are many constraints on the diameter, chief among them structural issues, fabrication issues, and aberration sensitivities. Because of the complexity of the constraints, it is best to choose a diameter based on a simple top-level brightness calculation, followed by an optimization exercise using nomographs or other optimization tools. Sample nomographs are given in this section, although the design constraints shown on the nomographs are particular to the system being designed and should not be taken to be generally applicable.

A *nomograph* is a series of interrelated graphic representations of key parameters, chosen so the interrelationship among dependent parameters is clear and linked in such a way that the physical relationships among the parameters can be determined graphically. Figure 5.19 shows a nomograph for optimizing diameter. The graphs show the brightness and power as a function of power/diameter product.

The design constraints are chosen based on the achievability of a given parameter and are shown as dashed lines. All points within the constraints are not equal. Each parameter has an extreme representing the most difficult and expensive design point. Note that other combinations of parameters may be more important in other systems. Section 5.18.1 contains an example that determines the diameter of an illuminator output aperture.

An effective diameter must be chosen to balance the available volume among the various parts of the system. Guard bands, assigned to protect components and allow for jitter and decentration, are set at 2.5% of the large aperture and 5% of the smaller apertures. These guard bands affect the allowable tilt and decentration of the beam through the system.
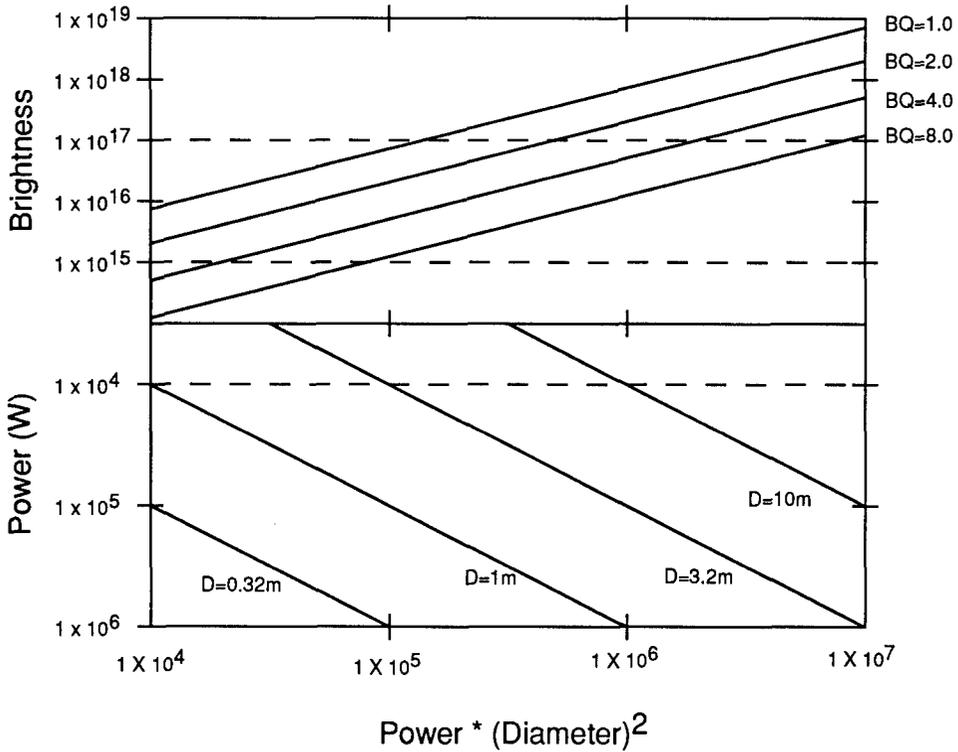
Fig. 5.19 Nomograph for choosing diameter.

5.7.3 Beam Expander Beam Quality

The calculation of beam quality for a beam expander system follows very closely the calculation for the beam transfer system, but with more emphasis on large optics and propagation issues. Metric errors within the telescope system can be significant. This same discussion applies to intervening telescopes or re-imaging optics in the beam transfer system.

5.7.3.1 Wavefront Effects

As with the BTS, wavefront effects can be a dominant contributor to system performance. Fabrication and maintenance of a good quality optical system for the BES is difficult and requires careful design.

Mirrors.   Mirrors and other optical elements in the beam expander are subject to thermal distortion, misfigure, and coating OPD errors. Medium effects in the beam expander are negligible because of the short path lengths (but see the upcoming propagation section). Thermal distortion of the secondary mirror is similar to the thermal distortion of a BTS mirror and is not repeated here. Thermal distortion of the primary mirror is usually somewhat different, because the PM is usually a low coefficient of thermal expansion (CTE) material,

the coating absorptivity is usually higher because of the larger size, and the mirror structure is more flexible because of the larger aspect ratios. Segmented mirror distortion is further complicated because thermal bowing of a segment (primarily spherical aberration) is a higher spatial frequency error over the full aperture. The thermal distortion of a PM is

$$\delta_{PM} = \delta_{bow} + \delta_{map} \ , \qquad\qquad (5.49)$$

where the bowing and mapping terms are the same as in the BTS section.

Misfigure of the BES mirrors is larger than for the BTS mirrors because of their curvature and because they are often paraboloids or hyperboloids. The effects of variation in focal length are discussed under metric errors. Misfigure is a complicated function of the diameter, the $f/\#$, the material, and the polishing process. The achievable rms misfigure is primarily a matter of cost, up to some limit (see Fig. 5.8). Telescopes such as the Hubble PM have been polished to accuracies of $\lambda/80$ (rms, surface, 0.633 μm). Mirrors of up to 50 cm diameter have been polished to $\lambda/200$ (rms, surface, 0.633 μm). Mirrors are routinely available to $\lambda/20$ (rms, surface). More expense can be expected for very thin substrates, very lightweight ("eggcrate") designs, extreme aspheres, etc. The structure of the resulting aberrations is also complex and depends on the polishing method.

Large lap polishing tends to smooth the surface and leave fewer high-spatial-frequency errors, but tends to roll edges and is not appropriate for steep surfaces. Small lap polishing tends to leave more mid- and high-spatial-frequency errors. Modern computer-controlled polishing uses small laps because they provide greater control of material removal.

Coating wavefront error effects on large optics tend to be either center-to-edge (spherical) or edge-to-edge (tilt) terms, depending on the geometry of the coating chamber. Coating thickness variations of 1% are feasible.

Metric Errors.   Two types of metric errors are important: errors in the focal lengths of the components and misalignment of the components. Maintenance of the focal length of a telescope mirror is nontrivial during the fabrication process because it usually involves a separate metrology test such as profilometry. The effect of an error in the focal length of one or both elements is

$$\Delta_f = \Delta_{f_a} f_b \left(\frac{z - f_a}{z^2} \right) \ , \qquad\qquad (5.50)$$

where

$$z = f_a + f_b - d \ ,$$

and

$$\begin{aligned}
f_a &= \text{focal length of the secondary mirror} \\
f_b &= \text{focal length of the primary mirror} \\
\Delta_{f_a} &= \text{error in the focal length of the secondary mirror} \\
d &= \text{primary-secondary separation.}
\end{aligned}$$

Similarly, the effect of a change in the separation of the telescope elements is

$$\Delta_f = \Delta_d \left(\frac{f_a f_b}{z^2} \right) , \tag{5.51}$$

where Δ_d is the error in telescope mirror separation. The change in the effective focal length of the system can be treated either as a range error or as an equivalent wavefront error by calculating the effective sag of the surface.

Metric error as a result of the misalignment of the BTS elements can be analyzed with geometric optics (ray trace) codes for complicated telescopes. For a simple two-element paraboloid-paraboloid (Mersenne) telescope, the WFE resulting from misalignment is

$$WFE_{PM} = \frac{D\theta_{PM}}{96\sqrt{8}\lambda(f/\#)^2} , \tag{5.52}$$

where

D = primary mirror diameter in meters
θ_{PM} = primary mirror misalignment angle in radians, one-axis
$f/\#$ = $f/\#$ of the telescope.[26]

The error resulting from motion of the secondary mirror has both tilt and translation contributions and is primarily coma:

$$WFE_{SM} = \frac{1}{48\sqrt{8}\lambda} \left[\frac{\theta_{SM} r_{SM}}{(f/\#)^2} + \frac{\Delta x_{SM}}{2(f/\#)^3} \right] , \tag{5.53}$$

where

θ_{SM} = misalignment angle of the secondary mirror
Δx_{SM} = decentration of the secondary mirror
r_{SM} = radius of the secondary mirror.

Figure 5.20 shows the effects of misalignment for a fast and a slower telescope. The example uses Eqs. (5.52) and (5.53) and assumes equal amounts of secondary mirror misalignment in radians, secondary mirror decentration in meters, and primary mirror misalignment in radians.

Sampling/Sensing. It is often desirable for control system purposes to sample the wavefront of a directed energy system as far downbeam toward the target as is possible. The best system is one in which the beam is sampled at the primary mirror of the telescope, just before leaving the directed energy system and heading toward the target.[c] Most recent systems use holographic optical

---

[c]For a DE system that includes correction for propagation in the atmosphere, a returning beacon or guide star may be used to sample the atmosphere. This return wavefront usually passes farther back into the DE system, toward the source. Including the optical train errors along with the atmospheric errors allows a single sensor to provide the data necessary to correct many errors.
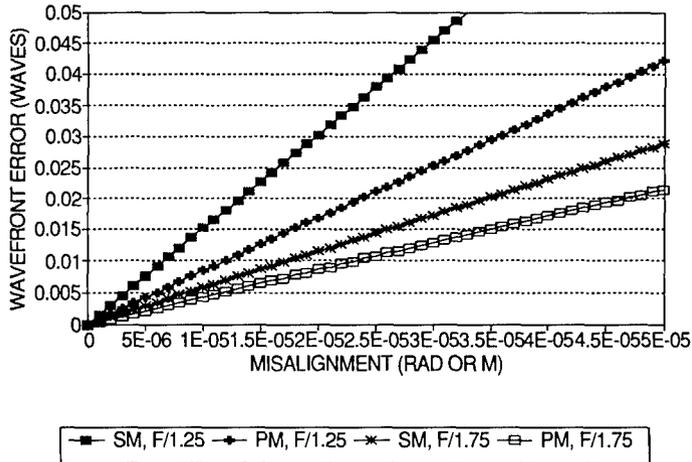
Fig. 5.20   Beam expander aberrations.

elements to sample the beam. These holographic gratings are effectively segments of an equivalent zone plate placed in patches on the primary mirror, with the patch size and spacing determined by the fabrication processes utilized. The beam-sampling fidelity is thus a function of the individual grating parameters as well as the geometry of the grating set.

Directed energy systems use HOEs to produce a set of Hartmann[d] spots. The HOE beams focus to the Hartmann spot plane. Each of the spots in the Hartmann plane can be monitored for positional changes. Because each spot is in the far field, the amount of displacement of the spot is proportional to the tilt across the HOE patch. This so-called Hartmann sensing is thus equivalent to a Hartmann test of an optical system. The errors introduced by the system that senses the Hartmann spots also affect the sensed OPD of the system and are important for determining the basic uncorrected system performance. The Hartmann spots also can cause optical path difference variations, depending on fabrication processes.

5.7.3.2 Boresight/Drift

Boresight and drift errors in the BES result from two primary effects: the misalignment of the boresight transfer system between the BES and the tracker, and pointing of the primary mirror (PM) and the secondary mirror (SM). Pointing errors may result either from structural deformation of the SM support structure or movement of the mirrors themselves. Thermal and fabrication errors are small. Structural stability is often dominated by gravity effects (or absence of gravity on a system calibrated in gravity). The effect of a gravity-induced sag on output pointing direction is

---

[d]Hartmann spots provide a sampling of the wavefront similar to a traditional Hartmann test,[27] but the slopes of the individual samples are measured by monitoring the deviation of each beam in the far-field plane.

$$\theta_{gs} = \frac{2\theta_{sag}}{M} ,$$ (5.54)

where θ_{sag} is the angle of the tangent to the deflected SM support at the SM and M is the magnification. The WFE resulting from a gravity sag can be calculated by the earlier equation for SM misalignment.

Drift of the mirrors themselves is simply

$$\theta_D = 2\theta_{PM} + \frac{2\theta_{SM}}{M} ,$$ (5.55)

where the θ_i are the drifts of the two mirrors.

Fabrication error in the tracker boresight transfer system causes a direct boresight error. The magnitude depends on the type of system used. A simple periscope system may have errors as large as 10 μrad. An extended corner cube may have errors of the order of 1 μrad. Sophisticated reference transfer systems may have submicroradian errors.

Boresight errors are usually expressed in target space, after the effect of the telescope magnification. Because PM and boresight transfer system errors are directly in target space, they dominate most budgets.

5.7.3.3   Jitter

Mirror Motions.   The output beam jitter induced by the beam expander primary mirror is simply twice the primary mirror jitter. The ratio of secondary mirror motion to output beam jitter is dependent on the parameters of the telescope. For a two-mirror telescope, the relationship between the output angle and the secondary mirror angle is

$$\theta_{out} = \frac{2\theta_{SM}}{M} ,$$ (5.56)

where M is the magnification. A telescope reduces jitter by the inverse of its magnification: if an optical system has 10 μrad of jitter entering a telescope of magnification 5, the jitter leaving the telescope is 2 μrad. Designers often speak of *input space* and *output space,* or *compact beam space* and *telescope space* to describe the effects of the telescope. The secondary mirror is in input space. The primary mirror is in output space.

Structural Motions.   Secondary mirror jitter is determined by the local mirror jitter and by the structural modes of the BES metric structure. In general, the two orthogonal low-order tilt modes (azimuth and elevation) of the BES structure dominate the structural contribution to jitter, although higher order modes can have measurable effects. There are no simple ways to determine the structural motions of arbitrary systems. A finite element computer model such as NASTRAN is usually used to analyze the motions.

5.7.3.4   Diffraction Effects

Diffraction effects and the resulting power losses were discussed in Sec. 5.7.1. The structure of the far-field spot is critical for many target interactions. For

convenience, a perfect near-field beam is considered to have uniform near-field intensity.[e] The amount of useful far-field energy is dependent on two primary parameters, the peak intensity and the edge fluence. A beam that is too small may penetrate (burn through) a target but not accomplish its purpose because the area of effect is too localized. The energy in the first bright ring of the diffraction pattern is sometimes used as a rule of thumb but is often optimistic because the edge fluence at the edge of the first bright ring is low (ideally zero).

5.7.3.5 Propagation

Beam quality degradation during propagation is due either directly to medium OPD (turbulence is an example), thermal absorption (thermal blooming is an example treated elsewhere in this Handbook), or nonlinear effects (such as self-focusing). Atmospheric turbulence is the subject of extensive investigation and experimental and analytical description as discussed in detail elsewhere in this Handbook. Turbulence can cause the beam to spread, change beam angle, or vary in intensity distribution. All of these change the amount of energy on target. Changes in intensity distribution are not considered here. Beam spreading can be characterized by

$$\theta \approx \left(\frac{1}{k^2 a^2} + \frac{1}{k^2 + \rho_0^2} \right)^{1/2} , \tag{5.57}$$

where

$$k = \frac{2\pi}{\lambda}$$

$$a = \text{beam radius} \tag{5.58}$$

$$\rho_0 = \left(\frac{\pi^2}{k^2 C_n^2 L} \right)^{3/5} ,$$

where C_n^2 is the atmospheric structure constant, L is the propagation distance, and C_n^2 ranges from 1×10^{-14} near the ground to 1×10^{-18} at 20 km (Table 5.2). These approximations are true for uniform turbulence (not varying with altitude). Changes in beam angle are

$$\langle \alpha^2 \rangle = \frac{D_\phi}{k^2 \rho^2} , \tag{5.59}$$

---

[e]The optimum energy distribution for propagating energy to the far field is a nonuniform distribution, generally Gaussian or at least apodized in some fashion. *Optimum* is a relative term depending on the desired structure in the far field, since maximum peak intensity on axis is not necessarily the best case for interacting with specific targets.

Table 5.2   Value of C_n^2 for Various Heights (data from Ref. 28)

HEIGHT (M)	DAY	NIGHT
18.5 - 240	1.7E-14(H/18.5)$^{-1}$	3.4E-14(H/18.5)$^{-1}$
240 -880	1.3E-15	6.5E-15
880 - 7000	8.87E-16(H/1000)$^{-3}$	8.87E-16(H/1000)$^{-3}$
7000 - 20500	6.34E-18(H/1000)$^{-1/2}$	6.34E-18(H/1000)$^{-1/2}$
> 20500	0	0

Measured above a 3.1 km mountain peak

where

$$D_\phi = 1.46 k^2 C_n^2 L \rho^{5/3} \; , \tag{5.60}$$

where ρ is the beam radial coordinate and L the propagation distance (Ref. 28, pp. 31–35, and Ref. 29).

Thermal blooming is an example of a series of distortions induced in the medium by the presence of a high-energy beam. Blooming may also take place within the beam transfer system. If no wind is present and for a Gaussian intensity profile, thermal blooming causes primarily a defocus term, because the lowest index of refraction occurs near the center of the beam, creating an effective negative or defocusing lens. If wind is present, the beam turns toward the direction from which the wind is blowing, commonly called *blooming into the wind*. The beam also changes shape, taking on a characteristic crescent shape with the arms of the crescent pointing toward the wind flow. An approximate formulation derived by Tyson and based on experiments and wave optics calculations relates the Strehl ratio to the blooming strength N_b by

$$S = \frac{1}{1 + K' N_B^m} \; , \tag{5.61}$$

where K' ranges from 0.0625 and $m = 2$ (Gaussian[30] beam) to $K' = 0.01$ and $m = 1.2$ (uniform beam).[28]

Nonlinear effects include the opposite of thermal defocus: self-focusing. The nonlinear index of refraction varies in such a way that an effective concave lens is created. Although the power of the lens is at first small, it is related to the intensity. As the beam converges, the effect grows stronger, converging the beam still more. This phenomenon can occur on a small scale within the beam, due to hot spots, and causes the local beam areas to break down, leaving intensity holes in the beam. Other nonlinear effects include stimulated Raman scattering and stimulated Brillouin scattering.

5.7.3.6   Other Effects

The frequency stability of the laser can affect the ability of a HOE grating sampler to determine the true focus of the beam. The wavelength stability is related to the frequency stability by

$$\Delta\lambda = \frac{-\lambda^2 \Delta\nu}{c} \ .$$
(5.62)

The apparent defocus is related to the wavelength by

$$\text{WFE}_{\text{rms}}(\text{waves}) = \frac{D_{\text{HOE}}}{\sqrt{3}\lambda}\left(\frac{m\Delta\lambda}{d\cos\theta}\right) \ ,$$
(5.63)

where

D_{HOE} = HOE spot diameter in meters
m = grating order
λ = wavelength
$\Delta\lambda$ = change in wavelength
d = grating spacing in meters
θ = grating diffracted angle.

For a typical HOE, the cosine of the diffracted angle is nearly 1.

5.8   TRACKING

The purpose of this section is to describe the effects of tracker performance and tracker relative alignment on the performance of the rest of the directed energy system. Figure 5.21 defines the terms used in this section. The field of view of the tracker, in angle space, is shown. The left image depicts the sit-
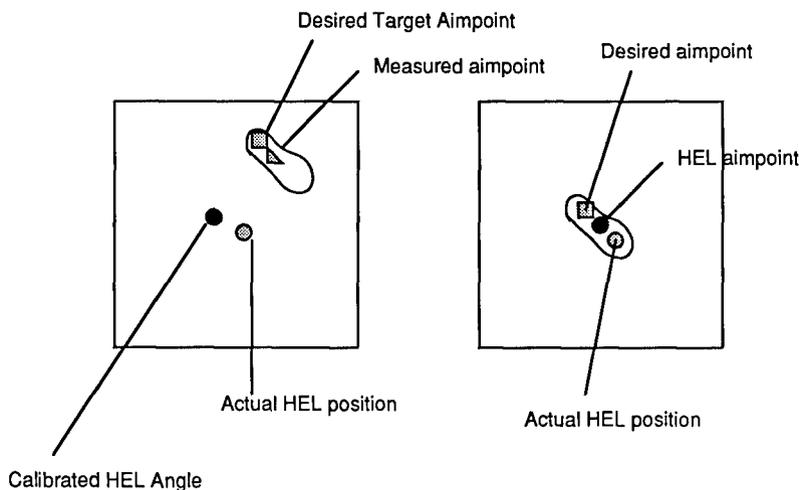
Fig. 5.21   Definition of tracking terms.

uation at target acquisition. The tracker calibration places the assumed HEL position at the center of the field of view. The actual HEL position is different by a boresight error. The extended object under track is characterized by an observed aim point. The desired aim point is different by a track error. The track error may increase during the irradiation time (drift). The instantaneous uncertainty in aim point location is dominated by tracker jitter.

After acquisition, the tracker commands the HEL, centering it on the target aim point. The right image shows the situation at the beginning of irradiation. Boresight and track error have been converted to an aim point error.

5.8.1   Boresight/Drift

The transfer of tracker boresight was described in the BES section. Drift of the tracker is determined primarily by structural effects. Thermal effects generally do not cause a drift, except as they affect the local area of the PM used to transfer boresight. This effect is identical to that described in Sec. 5.6.2.2. Structural effects can be analyzed in detail by a code such as NASTRAN.

Point Ahead.   Point ahead refers to a deliberate boresight error introduced to compensate for the finite time of propagation of a DE beam to a distant moving target. The point ahead angle is dependent only on the relative velocity of the DE system and the target and is independent of range. If an active tracker is used, the tracker illuminator must "lead" the target, whereas the tracker return signal will show a target lag. The point ahead angle θ is simply

$$\theta = \frac{\Delta v}{c} , \tag{5.64}$$

where Δv is the relative velocity in the plane orthogonal to the direction of propagation and c is the speed of light.

Relative velocity in the direction of propagation does not affect the point ahead, although it could conceivably require a focus correction at short ranges. An error in the point ahead angle has the same effect as a system boresight error. Measurement of the target relative transverse velocity is difficult. For a relative velocity of 1 km/s, the point ahead angle is 3.3 μrad. For a typical system of 1-m aperture and 1-μm wavelength, with an edge fluence requirement of 50% of peak, the relative velocity must be measured to about 30 m/s (see example in Sec. 5.18).

5.8.2   Jitter

Jitter within the tracker unit is usually reduced by referring the tracker reference to an inertial reference unit (IRU). The primary remaining component of jitter is rigid-body motion of the entire tracking unit, which cannot be distinguished from DE system output beam jitter in terms of effect on target. Because tracker jitter occurs in output-beam space, its magnitude must be a small portion of overall system jitter.

5.9 TARGET COUPLING

Many directed energy systems have been designed based on the brightness of the system or its irradiance on aim point. The description of the effectiveness of the system, whether it is a laser fusion system, a weapons system, or a metal working system, is incomplete without consideration of the coupling of the energy to the target. Target coupling is dependent on the physics of the interaction and its geometry.

5.9.1 Target Physics

Target physics determines how efficiently the incident energy couples into the target. A portion of the energy is reflected and scattered, the rest is absorbed. The absorption depends on target material, roughness, the wavelength of the incident irradiation, beam size, etc. Absorption physics are time and intensity dependent. The properties of the surface may change thermally or structurally during the period of irradiation. If material ablates, or if a plasma forms at the surface, the absorption characteristics can change dramatically. Figure 5.22 shows a simple overview of target-beam interaction illustrated by approximate burn-through times for bare aluminum. Below a minimum deposited energy, burn-through is not achieved, because the material dissipates the heat. At intermediate energy, penetration is achieved by melting. At higher energies, the material vaporizes. Finally, a separated plasma forms just above the target surface, which prevents energy from being coupled into the surface. Note that all four regimes may exist simultaneously over a beam spot on target because of variations in beam intensity. The exact mechanisms are dependent on material, energy density, pulse shape, and target structure. Target interaction physics for laser fusion targets is a separate discipline in itself and is beyond the scope of this treatment.

The threshold phenomenon is interesting. Most targets have a minimum energy before damage occurs. The minimum is dependent on the thermal diffusivity of the target. If the target can diffuse the heat rapidly enough, the beam will never damage the target, regardless of how long the irradiation lasts.
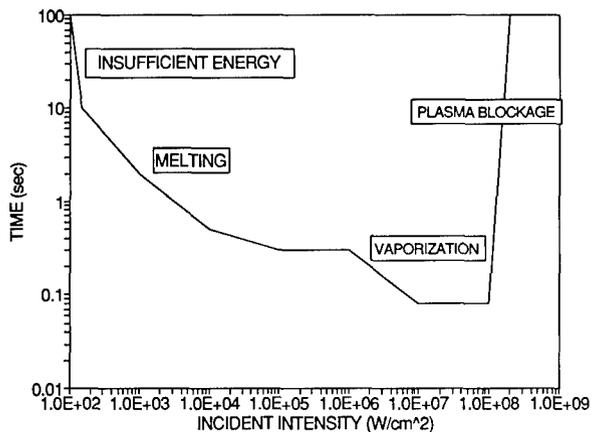
Fig. 5.22 Penetration time for bare aluminum, 2.7 μm (approximate).

5.9.2   Coupling Geometry

Geometric coupling includes two primary effects, the effective angle of incidence of the beam on the target plane and the geometry of the target plane. Consider a beam incident on a plane surface. The surface tangent to the plane is at a complex angle to the beam. The area irradiated by an incident beam is increased as the beam spreads to intercept a target plane at an arbitrary angle. The effective energy density is reduced (the effective spot size is increased). The effective geometric coupling factor is

$$I_{\text{eff}} = I_0 \cos\alpha \cos\beta \ , \tag{5.65}$$

where α and β are the direction cosines of the target trajectory, shown in Fig. 5.23. Target shape further complicates the geometry. First project the beam into the plane tangent to the target surface by Eq. (5.65). Then project the resulting ellipse onto the target structure. For some cases of interest the geometry is simple. For example, a cone-shaped object aimed directly at the DE system has an effective reduction in energy proportional to the cosine of the cone angle as long as the beam does not overfill the target.
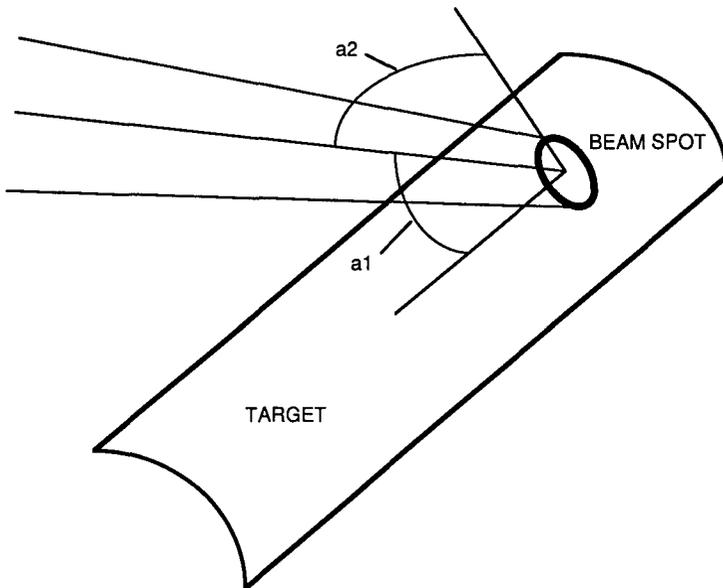
Fig. 5.23 Coupling geometry.

5.10   ENGINEERING FORMULAS—SYSTEM PERFORMANCE

Section 5.4 introduced the matrix of system requirements. A structured set of requirements, called *budgets,* guides the design of a DE system. One set of budgets summarizes overall system requirements. If all of the various parameters of the system (power, beam quality, etc.) meet these budgets, the system will perform adequately. The second set of budgets defines subsystem performance. The second set is useful for designing individual system elements. The two budget types flow from the same matrix formulation, and contain the same information, but are sorted to be convenient for their different purposes. Figure 5.24 shows the form of the performance budget. Each of the terms and its contributing factors is described here. The subsystem performance budget appeared in Sec. 5.4 (Fig. 5.2).

The budgeting process requires a common parameter that allows comparison of the effects of design changes in disparate systems. Many candidates have been proposed and used by various portions of the DE community. The most rigorous parameter that is sufficiently general is irradiance on target (IoT). Irradiance on target consists of the irradiance that can be placed on the target aim point (IoA), the effective time that the beam spends on the target, and coupling of the beam energy to the target. Time on the aim point is selectable and is not discussed further. Target coupling was described in Sec. 5.9.
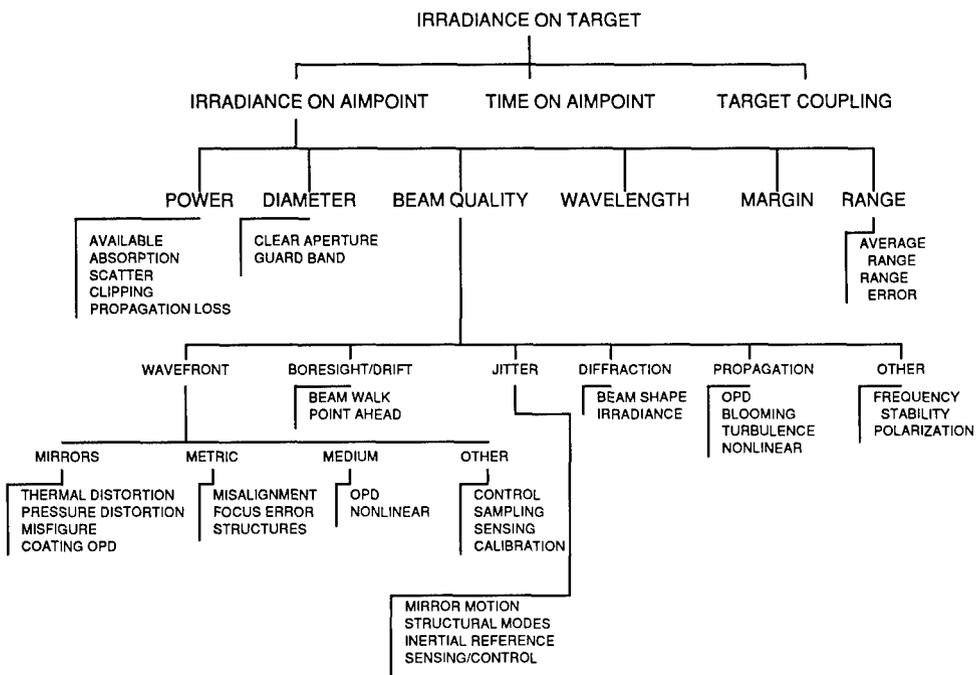
Fig. 5.24   System performance budget.

Sections 5.4 to 5.9 showed how to develop a subsystem performance budget. The designer also needs to understand how the overall system performs. This section shows how to combine the various subsystem effects into an overall performance budget.

IoA is the encircled energy in a defined radius around the aim point and is a function of six primary items, as shown in Fig. 5.24. Power, diameter, beam quality, and wavelength can conveniently be summarized by brightness, a useful approximate indicator of system performance. Brightness is

$$B = \frac{\pi P D^2}{4\lambda^2 BQ^2} \, ,$$
(5.66)

where

P = power at the system exit aperture
D = system effective output diameter
λ = system wavelength
BQ = system beam quality.

The power term must include all power losses, including obscurations. The diameter must include the effective area. Another useful system parameter is the Strehl ratio:

$$I_s = \frac{I_A}{I_I} \, ,$$
(5.67)

defined in terms of brightness as

$$I_s \approx \frac{B}{L^2} \propto \frac{1}{BQ^2} \, ,$$
(5.68)

where

I_A = actual peak intensity at the center of the far-field distribution
I_I = ideal peak intensity
B = brightness
L = range to the far-field spot
BQ = beam quality.

Using the approximation, brightness divided by the square of the range, in Eq. (5.68) also yields the energy density, in watts per square meter, in the small region near the peak. Calculation of IoA depends on the definition of spot size on target. If an edge fluence criterion of 50% of the peak intensity is used, the encircled energy is about 50%, and the approximate IoA reduces to

$$\text{IoA} \approx \frac{0.5P}{BQ^2} \, .$$
(5.69)

Equation (5.69) includes the diffractive effect described in Sec. 5.13.4, so the beam shape term should be omitted from the BQ if Eq. (5.69) is used. The radius of the spot for this criterion is approximately

$$z = \frac{\lambda L}{2D} .$$
(5.70)

The approximations break down if the spot is too small. A rigorous derivation of the IoA includes higher order terms. Consequently, the preceding budgeting formulas are valid for only small errors (BQ less than about 1.7). System performance is generally adequate only when errors are small, however, so the budget process functions well in most cases. The quadratic relationship between IoA and most system errors is key.

A commonly used, although potentially somewhat confusing, parameter is the phase contribution to beam quality, which can be related to the wavefront error by the Marechal approximation:

$$BQ_p = \exp[0.5(2\pi\delta)^2] ,$$
(5.71)

where δ is the wavefront error in waves. A more general definition of beam quality includes nonphase terms:

$$BQ = BQ_p * BQ_{np} .$$
(5.72)

Note that the multiplying of beam quality terms implies that the effects are uncorrelated, both temporally and spatially. The multiplication of beam qualities is equivalent to root sum squaring the individual phase terms, which leads to beam quality degradation. For uncorrelated errors,

$$BQ = BQ_1 BQ_2 = \exp[0.5(2\pi\delta_1)^2] \exp[0.5(2\pi\delta_2)^2] ,$$
(5.73)

$$BQ = \exp[0.5(2\pi)^2(\delta_1^2 + \delta_2^2)] .$$
(5.74)

If the errors are correlated,

$$\delta = \delta_1 + \delta_2 ,$$
(5.75)

and

$$BQ = \exp\{0.5[2\pi(\delta_1 + \delta_2)]^2\} ,$$
(5.76)

or

$$BQ = \exp[0.5(2\pi)^2(\delta_1^2 + \delta_2^2 + 2\delta_1\delta_2)] ,$$
(5.77)

where the δ_n are the individual wavefront errors. For example, consider two errors of equal magnitude, 0.1 waves. In one case, let the two errors be the result of thermal distortion at successive identical mirrors in the beam train. The two errors are spatially correlated. In the second case, let the errors be the random misfigure of the two mirrors. The second two errors are not spatially correlated. For the two cases:
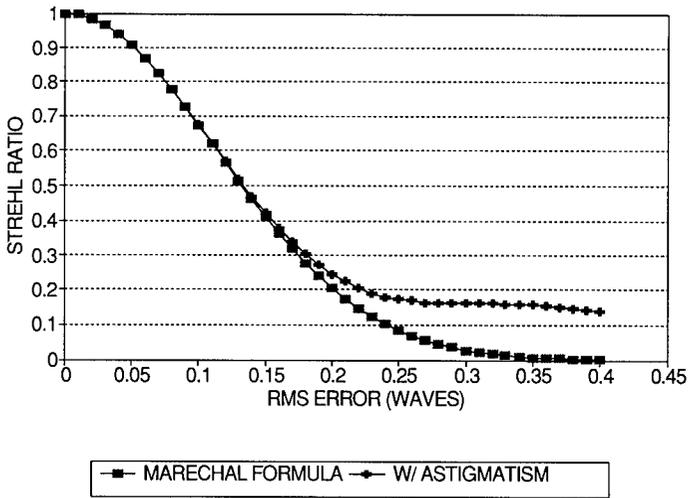
Fig. 5.25 Wavefront error comparisons.

Case 1: Correlated $BQ = \exp[19.74(0.01 + 0.01)] = 1.484$,

Case 2: Uncorrelated $BQ = \exp[19.74(0.01 + 0.01 + 0.02)] = 2.203$.

$$(5.78)$$

Figure 5.25 and Table 5.3 show the relationship among rms wavefront error, Strehl ratio, and beam quality (wavefront definition) for various magnitude errors. Also shown is peak intensity for a uniformly illuminated, unobscured circular aperture with an astigmatic error.[31] For wavefront errors greater than $\lambda/10$, the approximate definitions become increasingly less accurate and the structure of the particular aberration must be considered.

Because of the inexact nature of some of these commonly used forms, and the overlap among their terms and applications, IoA is a better overall system parameter. IoA can be calculated directly by a computer code or can be inferred by building a series of approximate models.

Table 5.3   Wavefront Error, Strehl, and BQ for Various Magnitude Aberrations

DELTA WAVES	DELTA 1/WAVES	STREHL	BQ	STREHL W/ ASTIGMATISM
0	-	1.0000	1.0000	1.0000
0.01	100.00	0.9961	1.0020	0.9961
0.02	50.00	0.9843	1.0079	0.9843
0.03	33.33	0.9651	1.0179	0.9651
0.04	25.00	0.9388	1.0321	0.9388
0.05	20.00	0.9060	1.0506	0.9060
0.06	16.67	0.8675	1.0736	0.8676
0.07	14.29	0.8241	1.1016	0.8243
0.08	12.50	0.7767	1.1347	0.7771
0.09	11.11	0.7263	1.1734	0.7271
0.1	10.00	0.6738	1.2182	0.6752
0.11	9.09	0.6202	1.2698	0.6225
0.12	8.33	0.5664	1.3288	0.5700
0.13	7.69	0.5132	1.3960	0.5186
0.14	7.14	0.4613	1.4724	0.4692
0.15	6.67	0.4114	1.5591	0.4225
0.16	6.25	0.3640	1.6575	0.3792
0.17	5.88	0.3195	1.7691	0.3397
0.18	5.56	0.2783	1.8956	0.3044
0.19	5.26	0.2405	2.0393	0.2734
0.2	5.00	0.2062	2.2024	0.2468
0.21	4.76	0.1753	2.3881	0.2246
0.22	4.55	0.1480	2.5997	0.2065
0.23	4.35	0.1239	2.8411	0.1922
0.24	4.17	0.1029	3.1173	0.1814
0.25	4.00	0.0848	3.4339	0.1736
0.26	3.85	0.0693	3.7976	0.1683
0.27	3.70	0.0562	4.2164	0.1650
0.28	3.57	0.0453	4.7000	0.1632
0.29	3.45	0.0361	5.2597	0.1624
0.3	3.33	0.0286	5.9093	0.1622

5.11   POWER BUDGETS

The power budget contains five terms: available power, absorption losses, scatter losses, clipping losses, and propagation losses. Calculation of the net power budget starts with a calculation of power from the device (as in Sec. 5.5.1). For a system without further amplification, power continually decreases. If amplifiers are used, intervening losses affect the input power to the amplifier, but not the total system output (except indirectly). For the remainder of this discussion, the system will not contain amplifiers.

Absorption losses can be summed simply by

$$\alpha_{ab} = 1 - (1 - \alpha_1)(1 - \alpha_2) \dots (1 - \alpha_n) \; , \tag{5.79}$$

where the α_n are the individual absorptivities. Note that absorptions are not additive, although the difference is small for very low absorptions. For example, if there are 10 absorbers, each with absorption 0.1, α_{ab} is 0.65.

In a similar fashion, total scatter losses can be calculated by

$$\alpha_s = 1 - (1 - TIS_1)(1 - TIS_2) \dots (1 - TIS_n) \; , \tag{5.80}$$

where the TIS_n are the individual element scatter losses.

Total clipping losses include obscurations and clipping. Obscuration losses can be calculated by geometrically summing all obscurations into a net obscuration (often best done graphically or through a ray-trace equivalent). For a uniform irradiance distribution, the net loss is

$$\alpha_o = \left(1 - \frac{A_o}{A_b}\right) , \tag{5.81}$$

where A_o is the obscured area and A_b is the total beam area. Clipping losses represent power that expands to a diameter larger than the beam clear aperture. The power is clipped to protect the system. The amount of clipped power depends on the clipper aperture. Summing the losses for each section of the beam path with significant path length (long enough for significant diffraction), the total clipped losses are

$$\alpha_c = 1 - (1 - \alpha_{c1})(1 - \alpha_{c2}) \dots (1 - \alpha_{cn}) , \tag{5.82}$$

where the α_{cn} are the individual clipping losses.

Propagation losses include absorption, scatter from aerosols, and nonlinear effects:

$$\alpha_p = 1 - (1 - \alpha_{pa})(1 - \alpha_{ps})(1 - \alpha_{pn}) , \tag{5.83}$$

where

α_{pa} = absorption over the path to the target
α_{ps} = loss resulting from scatter
α_{pn} = effective absorption due to nonlinear effects.

Combining all of these effects yields the net power budget expression:

$$P_{\text{net}} = P_a(1 - \alpha_{\text{ab}})(1 - \alpha_s)(1 - \alpha_o)(1 - \alpha_c)(1 - \alpha_p) , \tag{5.84}$$

where P_a is the available power, P_{net} is the net power at the target, and the α were defined earlier. The net output power is not present in the central lobe of the far-field spot. The distribution of energy must be calculated as shown in the beam quality section. Recall that each of the α_n was defined in Secs. 5.5.1, 5.6.1, and 5.7.1 and can be calculated using the equations found there.

5.12   DIAMETER BUDGETS

Although choosing a diameter is difficult, budgeting the chosen value is simple. The total effective diameter is simply the sum of the clear aperture diameter and the guard band. Note that the diffraction effects of nonuniform apertures are discussed in Sec. 5.13.4. A guard band is an annulus around the outer diameter of the primary mirror clear aperture (as well as other mirrors). The guard band serves two purposes, allowing for a certain amount of beam walk and protecting the outer mounting structures of the primary mirror. The size of the guard band is usually 1 to 2% of the output diameter for a large primary

(more for a smaller) focusing mirror. The limiting diameter of the system is usually the telescope primary mirror, because the largest mirror is usually the most difficult to fabricate. Wide field-of-view systems may require oversize beam transfer mirrors because of retargeting-induced beam walk.

5.13   BEAM QUALITY BUDGETS

Beam quality includes all physical effects that determine the variation in size and structure of the beam on its aim point, relative to an ideal beam. The peak intensity in the far field can be calculated for arbitrary effects (small). The structure of the far field is much more complicated and depends on the type and symmetry of the aberration. By itself, a peak intensity criterion is insufficient, because of edge fluence requirements in target interactions. For budgeting purposes, peak intensity can be used as long as the edge fluence requirement is monitored.

The beam quality terms to be used in the IoA budget can be summarized as

$$BQ = BQ_{WFE}BQ_{drift}BQ_{jit}BQ_{diff}BQ_{prop}BQ_{other} \ , \tag{5.85}$$

where the BQ terms result from wavefront error, boresight/drift, jitter, diffraction, propagation, and other effects. This approach is a simplification valid for a point design or over a relatively narrow range of parameters. Some beam quality effects are dependent on power and diameter.

For many years, the HEL device community has used a narrower definition of beam quality, with several different and sometimes conflicting definitions. The HEL beam control community has commonly used the phase-only definition of beam quality [Eq. (5.71)].

Beam quality is affected by all wavefront perturbations of the beam, but also includes such nonphase effects as polarization, spectral content, partial aperture lasing, and irradiance distribution. Beam quality is a dimensionless quantity. A perfect beam has a beam quality of 1. An aberrated beam has a beam quality greater than 1. Most common definitions of beam quality contain approximations that are increasingly invalid as beam quality increases. At beam qualities much in excess of 1.7, actual irradiance distribution at the aim point must be calculated, and an effective power-in-the-bucket definition is usually used. Power-in-the-bucket is the encircled power in a defined area around the aim point.

5.13.1   Wavefront Effects

All system errors that cause one portion of the beam to traverse a different pathlength are called *wavefront effects*. These include geometric aberrations such as coma or astigmatism, misalignment of powered surfaces, thermal distortion, etc. By convention, effects that occur after the beam has left the system exit aperture are referred to as propagation effects. Correlated and uncorrelated effects are budgeted separately because they sum differently.

Wavefront errors can be summarized as:

$$BQ_{WFE} = BQ_{med}BQ_{res}BQ_{bow}BQ_{map}BQ_{PM}BQ_{SM}BQ_{fig} \ . \tag{5.86}$$

The individual BQ terms were described in Secs. 5.4 to 5.9 and are

BQ_{med} = gain medium density variation
BQ_{res} = resonator WFE (misfigure and misalignment)
BQ_{bow} = thermal bowing
BQ_{map} = thermal mapping
BQ_{PM} = primary mirror misalignment
BQ_{SM} = secondary mirror misalignment
BQ_{fig} = BQ resulting from all mirror misfigures outside of the resonator.

5.13.2 Boresight/Drift

Uncorrelated and correlated boresight and drift can be budgeted separately, but correlated boresight errors are generally small. Correlated tilts in general add in a coherent fashion, using vector addition in the beam coordinate system. Uncorrelated tilts add in an RSS fashion:

$$\sigma_{bd} = (\sigma_{bd1}^2 + \sigma_{bd2}^2 + \cdots + \sigma_{bdn}^2)^{1/2} , \tag{5.87}$$

where the σ_{bdn} are the individual uncorrelated boresight or drift errors. The effects of boresight and drift are negligible as long as the beam remains within the aim point tolerance, because the beam shape is not affected (Fig. 5.26). If the target size is larger than the aim point size, which is larger than the beam size, the BQ is not affected as long as the drift remains less than d_1. For drift a, the irradiance on aim point drops rapidly. If the allowable aim point size equals the beam size, drift effects begin immediately. Thus,

$$BQ = 1 \quad \sigma_{bd} \leq (\theta_a - 2\theta_b) , \tag{5.88}$$

ANGLE SPACE
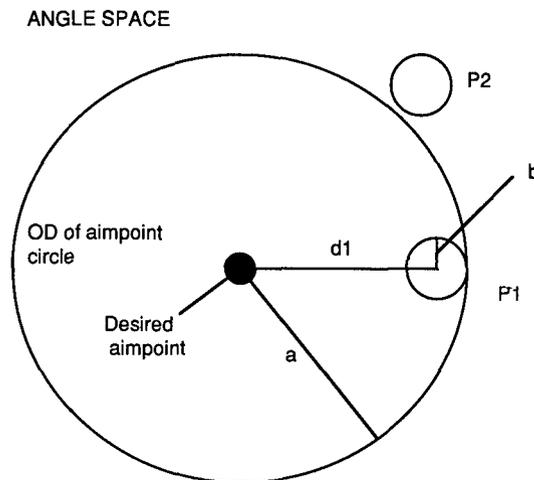
Fig. 5.26 Definition of boresight and drift variables.

$$BQ = \infty \quad \sigma_{bd} \geq (\theta_a + \theta_b) \ , \tag{5.89}$$

where an infinite BQ implies no irradiance on aim point. In between these two cases,

$$BQ_{\text{drift}} \approx \left[\frac{1}{0.92 - (0.84\sigma_{bd}/2\theta_b)} \right]^{1/2} - 0.0425 \ , \tag{5.90}$$

where θ_a is the aim point angular radius and θ_b is the beam angular radius. This formula is an approximation based on the encircled energy for a circular aperture. In this case, the beam angular radius is approximately $1.22\lambda/D$, where D is the near-field beam diameter.

5.13.3 Jitter

Jitter affects the shape of the beam. For individual uncorrelated jitter effects, the total jitter σ_T is

$$\sigma_T = (\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2)^{1/2} \ , \tag{5.91}$$

where the σ_n are the individual jitters. The effect of jitter on the far-field spot depends on the probability distribution of the jitter. For Gaussian jitter, an effective BQ can be defined as (Ref. 28, p. 16)

$$BQ_{\text{jit}} = \{1 + [1.57\sigma_T (D/\lambda)]^2\}^{1/2} \ . \tag{5.92}$$

The rms jitter for an individual mirror is usually expressed as

$$\sigma_n = \sigma_r = (\sigma_x^2 + \sigma_y^2)^{1/2} \ , \tag{5.93}$$

where σ_x and σ_y are the rms jitters along orthogonal axes (often referred to as azimuth and elevation).

5.13.4 Diffraction Effects

Even an otherwise perfect beam diffracts as it passes from the exit aperture of the system to the target. Beam obscurations, beam shape, and energy distribution in the beam affect the far-field distribution. Diffraction effects are among the most difficult to budget, and are generally treated as individual uncorrelated effects. Exact analysis by means of diffractive optics computer codes is necessary whenever the beam phase and intensity structures become complicated. The diffraction BQ term is divided into shape (overall diffraction) and obscuration effects:

$$BQ_{\text{diff}} = BQ_{\text{shape}} BQ_{\text{obs}} \ . \tag{5.94}$$

The BQ_{shape} term should be omitted if the approximation of Eq. (5.69) is used.

For a uniform-intensity flat-phase beam, the edge intensity criterion defines an effective target spot size. Referring to Fig. 5.18, the 50% edge fluence point contains about 50% of the encircled energy. For a radius of $\lambda/2D$,

$$BQ_{shape} = \left(\frac{1}{0.456}\right)^{1/2} = 1.48 \ . \tag{5.95}$$

The lower the allowable edge fluence, the larger an effective "bucket" can be used, and the lower the diffraction BQ term. In general,

$$BQ_{shape} = \left[\frac{1}{1 - J_0^2(x) - J_1^2(x)}\right]^{1/2} \ , \tag{5.96}$$

where the J_n are the Bessel functions of the first kind (tabulated in Sec. 5.17), and

$$x = \frac{2\pi}{\lambda L} rz \ , \tag{5.97}$$

where r is the near-field radius, z is the far-field coordinate, and L is the propagation distance. For $z/L = \lambda/2D$, where D is the near-field diameter, $x = \pi/2$ and BQ = 1.48.

The effect of obscurations on BQ, exclusive of power loss, is

$$BQ_{obs} = \left(\frac{1}{1 - \varepsilon}\right)^{1/2} \ . \tag{5.98}$$

If the power loss term were included, the expression would be

$$BQ_{obs} = \left(\frac{1}{1 - \varepsilon^2}\right)^{1/2} \ , \tag{5.99}$$

but the equivalent term would need to be excluded from the power loss budget. Most treatments of aberration theory assume a uniform intensity and assume that obscuration power loss and geometry are considered together. Although rigorous, such an approach makes budgeting difficult, because power loss effects are distributed into the beam quality budget.

5.13.5 Propagation Effects

As the beam propagates through a medium to the target, it may suffer effects including turbulence, interactive effects such as blooming, or nonlinear phenomena such as stimulated Raman scattering. Propagation effects are unique to the system application. For space or vacuum applications they are minimal. The dominant BQ effects result from turbulence and blooming:

$$BQ_{prop} = BQ_{bloom}BQ_{turb} \ . \tag{5.100}$$

Because of nonlinearities, this simple relationship is valid for a particular power and diameter. The individual terms must be recalculated if power or diameter change significantly.

5.13.6 Other Effects

Other effects include polarization, output frequency stability, and aberrations specific to particular systems. As with propagation effects, these effects are sufficiently disparate to defy convenient summary and are treated individually:

$$BQ_{other} = BQ_{pol}BQ_{freq} \cdots .$$ (5.101)

Except for specific problems in particular systems, $BQ_{other} = 1.0$.

5.14 WAVELENGTH

The choice of a wavelength for a DE system is an extremely complicated problem. Available laser lines restrict choices, but lines become more abundant each year. Some directed energy systems are tunable over a range of wavelengths (for example, free-electron lasers). The temptation to choose a shorter wavelength because of a smaller beam spot must always be balanced against the many other wavelength effects. Polishing errors, thermal distortion, and coating OPD are independent of wavelength, and have proportionally larger influence as wavelength decreases. System jitter is also independent of wavelength. In addition, smaller diameter systems are generally lower in magnification, increasing the effective jitter. All of these terms affect the IoA in a quadratic fashion, inversely proportional to the quadratic wavelength effect [see Eq. (5.66)].

In a multiple-wavelength system, the power-weighted (pw) average wavelength should be used for all the approximate calculations in this chapter:

$$\lambda_{pw} = \frac{\lambda_1 P_1 + \lambda_2 P_2 + \cdots + \lambda_n P_n}{P} ,$$ (5.102)

where $P = \Sigma P_n$. Consider the spectrum

$P_1(6)$ $P = 0.25$ $\lambda = 2.7075$ μm
$P_1(7)$ $P = 0.30$ $\lambda = 2.7440$ μm
$P_2(6)$ $P = 0.20$ $\lambda = 2.8319$ μm
$P_2(7)$ $P = 0.25$ $\lambda = 2.8705$ μm .

The power-weighted wavelength is 2.7841 μm. This approach is valid as long as the wavelength spread is reasonably small. Aberrations resulting from dispersive elements must be calculated separately.

5.15 MARGIN

The conceptual design process requires margin at the system, subsystem, or component level. Once a detailed design starts, designers routinely employ

"design conservatism" to ensure adequate performance. A system engineer should hold *explicit margin,* which refers to designing a system to requirements somewhat better than desired performance. *Implicit margin* means designing a system with individual elements each slightly better than their requirements. Top-level performance margin for a DE system should be 50% at the conceptual design stage, 30% at the preliminary design, and 10% at the critical design. After allocation to the lowest levels of the budget, even 50% margin calls for only moderate component margins.

5.16   RANGE

The range of a DE system is as important as any other parameter because irradiance falls off as the square of the range. For illuminator/tracker or laser radar systems, returned power varies inversely as the fourth power of the range. Of critical importance for many systems is range accuracy. The target must be well within the DE system's depth of focus. Incorrect range can also lead to degradation of system performance because of the beam structure in the propagation direction as it nears the far field. Because many target coupling mechanisms depend on peak power, edge fluence, and average power (all three simultaneously), beam structure at the target plane is critical. Range accuracy needs to be about 10% of the depth of focus for many applications. The depth of focus is

$$\Delta z = \pm 2\lambda (f/\#)^2 , \tag{5.103}$$

where $f/\#$ is the system $f/\#$ (effective focal length divided by the clear aperture, or approximately range divided by diameter) and λ is the wavelength. The effective range criterion (10% of the depth of focus) is then

$$\Delta R = \pm 0.2\lambda \left(\frac{L}{D}\right)^2 . \tag{5.104}$$

For $\lambda = 1$ μm, $L = 1000$ km, and $D = 10$ m,

$$\Delta R = \pm 0.2(1 \times 10^{-6})(1 \times 10^{12})(1 \times 10^{-2}) = \pm 2 \times 10^3 \text{ m} . \tag{5.105}$$

This depth-of-focus criterion represents a Strehl ratio of 0.8 for a uniformly illuminated circular aperture. The effective depth of focus is smaller for aberrated beams.

5.17   DATA COMPENDIUM

5.17.1   Laser Wavelengths

Directed energy systems use those lasers that are capable of high continuous or pulsed power. Table 5.4 tabulates the output wavelengths of lasers from which high power has been obtained. Not all wavelength transitions are listed, because many are impractical for high-power use. Transitions labeled (n) in the wavelength column are indicated to show the extrema of the group and

Table 5.4   Output Wavelengths of Lasers

Lasing Medium	Transition	Wavelength (μm)
HF (ml)	$P_1(1)$	2.5508 (n)
	$P_1(4)$	2.6398
	$P_1(5)$	2.6727
	$P_1(6)$	2.70752
	$P_2(4)$	2.7604
	$P_1(7)$	2.7441
	$P_1(8)$	2.7826
	$P_2(5)$	2.7953
	$P_1(9)$	2.8231
	$P_2(6)$	2.8318
	$P_2(7)$	2.8706
	$P_2(8)$	2.9111
	$P_2(9)$	2.9539
	$P_2(16)$	3.3206 (n)
DF (ml)	$P_1(2)$	3.4933
	$P_1(4)$	3.5507
	$P_1(5)$	3.5811
	$P_1(6)$	3.6128
	$P_2(4)$	3.6665
	$P_1(7)$	3.6456
	$P_1(8)$	3.6798
	$P_2(5)$	3.6983
	$P_1(9)$	3.7155
	$P_2(6)$	3.7310
	$P_2(7)$	3.7651
	$P_2(8)$	3.8007
	$P_2(9)$	3.8375
	$P_2(17)$	4.1862 (n)
CO_2 (sl–ml)	$P12$ (00^01–10^00)	10.51312
	$P13$	10.54838
	$P14$	10.53209
	$P15$	10.56744
	$P16$	10.55140
	$P17$	10.58684
	$P18$	10.57105
	$P19$	10.60658
	$P20$	10.59104
	$P21$	10.62666
	$P22$	10.61139
	$P23$	10.64710
	$P24$	10.63210
	$P25$	10.66789
	$P26$	10.65316

Table 5.4   *(continued)*

Lasing Medium	Transition	Wavelength (μm)
	$P27$	10.68404
	$P28$	10.67459
	$P29$	10.71055
	$P30$	10.69639
	$P22$ $(00^01–02^00)$	9.56918
	$P24$	9.58623
	$P26$	9.60357
	$P28$	9.62122
Excimers		
KrF		248.4
		249.1
XeCl		307.0
		307.3
		307.65
		307.92
		308.17
		308.43
XeF		348.8
		351.1
		353.1
		354.0
Nd:glass	(depends on glass)	1.06
Nd:YAG $(Y_3Al_5O_{12})$		1.06
Oxygen iodine		1.315
HF overtone		1.32
CO		5.35 (4.9–5.5)

do not usually have significant power. The number of wavelengths available simultaneously is limited in multiple-wavelength lasers. Usually four to eight lines lase simultaneously. Lasers are identified as multiple wavelength (ml) or single wavelength (sl). The two categories are blurred because multiple wavelength lasers can be made single wavelength by the use of intracavity gratings. The labels indicate the mode in which the most power can be extracted.[f]

Another important laser is the free-electron laser, usually run in a MOPA configuration with the output wavelength determined by the *oscillator,* usually Nd:glass or Nd:YAG.

5.17.2   Typical Budget Values

Figure 5.27 shows a complete IoT budget for a typical high-energy laser system of arbitrary power. The exact values of budget parameters depend on individual system constraints. The system is assumed to have no adaptive optics correction.

Calculation of Irradiance on Aimpoint.   For a uniformly illuminated circular aperture with no aberrations (BQ = 1), and $\lambda = 1 \times 10^{-6}$ m, $D = 2$ m, $P = 1 \times 10^5$ W, and $L = 1 \times 10^6$ m, the brightness is

---

[f]The wavelength tables come from data found in Marvin J. Weber, *Handbook of Laser Science and Technology,* Vol. 2. Weber lists extensive tables of lines for various lasers, along with a bibliography of sources for the information, but does not indicate which lines are useful for high-power devices.
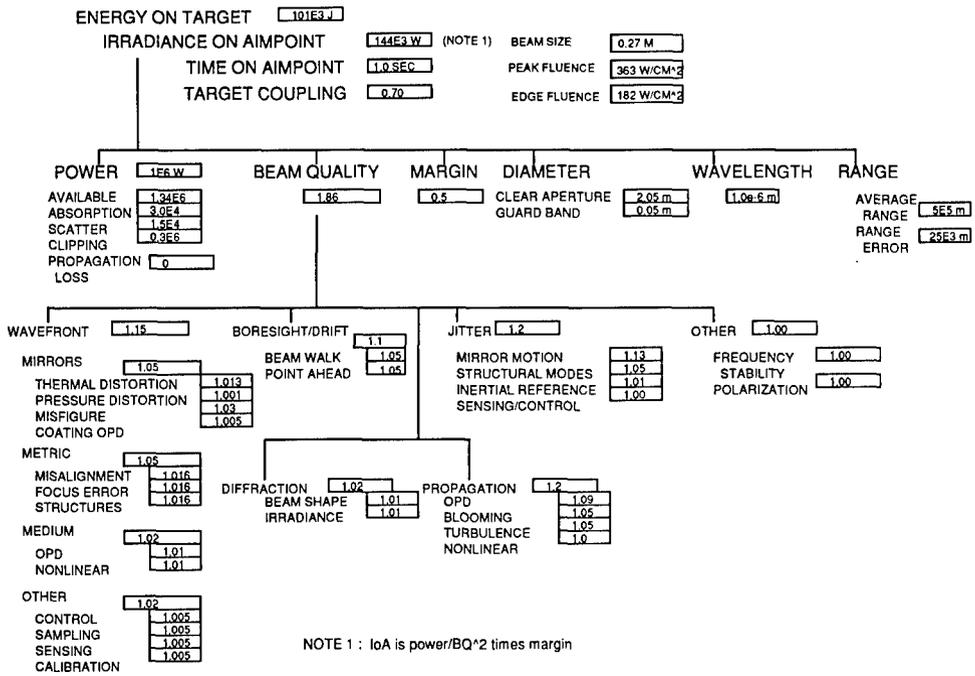
Fig. 5.27   Irradiance-on-target budget (no adaptive optics correction).

$$B = \frac{\pi P D^2}{4\lambda^2 BQ_2} = 3.14 \times 10^{17} \; .$$

(5.106)

The irradiance on aimpoint is

$$\text{IoA} = P[1 - J_0^2(x) - J_1^2(x)] \; ,$$

(5.107)

where

$$x = \frac{2\pi r z}{\lambda L} \; .$$

(5.108)

The intensity at the center of the far-field distribution is

$$I_0 = \frac{\pi P D^2}{4\lambda^2 BQ^2 L^2} \; .$$

(5.109)

The actual IoA for an aberrated beam must be found by integrating the far-field distribution over the defined spot radius. Except for simple, and single, aberrations, such solutions are not tractable analytically. Computer solutions or experimental data must be obtained. As the far-field spot becomes aberrated, two effects happen. Energy is redistributed into the wings of the diffraction

pattern, and the central lobe of the pattern decreases in intensity and expands. For small aberrations, the change in shape is not strong, and the IoA is approximately

$$\text{IoA} = \frac{P}{\text{BQ}^2}[1 - J_0^2(x) - J_1^2(x)] \ ,\tag{5.110}$$

where x is as in Eq. (5.108). This is a rule of thumb only and should not replace more detailed calculations.

5.17.3 Laser Line Absorption[32]

Typical absorption values for laser lines under varying conditions are given in Table 5.5.

Table 5.5   Selected Laser Line Attenuation by the Atmosphere
(from Ref. 18, pp. 5-66 to 5-69)

HF Laser Parameters			Atmospheric Absorption Coefficients (km^{-1})			
			$z = 0$ km, Sea Level			$z = 12$ km
Band	Rot. ID	$v(cm^{-1})$	k_{trop}	k_{mw}	k_{sw}	k_{mw}
1-0	P11	3436.12	2.21	0.221	0.0542	0.0000287
	P12	3381.50	0.496	0.0751	0.0231	0.000022
2-1	P8	3435.17	2.01	0.209	0.0512	0.0000267
3-2	P6	3373.46	0.364	0.0537	0.0168	0.000029
4-3	P8	3130.09	0.801	0.148	0.0554	0.000295
	P9	3083.83	1.12	0.211	0.0808	0.000806
5-4	P4	3150.67	0.498	0.126	0.0736	0.00229
6-5	P6	2921.74	0.586	0.0453	0.0103	0.000077
	P7	2880.70	0.0430	0.00424	0.00121	0.000006
	P8	2838.59	0.369	0.0654	0.0218	0.000044

DF Laser Parameters			Atmospheric Absorption Coefficients (km^{-1})			
			$z = 0$ km, Sea Level			$z = 12$ km
Band	Rot. ID	$v(cm^{-1})$	k_{trop}	k_{mw}	k_{sw}	k_{mw}
1-0	P1	2884.934	0.414	0.123	0.0772	0.00316
	P2	2862.652	0.0540	0.0115	0.00485	0.00316
	P3	2839.779	0.0386	0.00725	0.00266	0.000038
	P4	2816.362	0.0837	0.0190	0.0104	0.00108
	P5	2792.437	0.0471	0.0106	0.00496	0.000157
	P6	2767.914	0.0719	0.0184	0.00952	0.000672
	P7	2743.028	0.0352	0.00801	0.00352	0.000043
	P8	2717.536	0.114	0.0204	0.00718	0.000034

Table 5.5   *(continued)*

DF Laser Parameters			Atmospheric Absorption Coefficients (km^{-1})			
			z = 0 km, *Sea Level*			z = 12 km
Band	*Rot. ID*	$\nu(cm^{-1})$	k_{trop}	k_{mw}	k_{sw}	k_{mw}
	P9	2691.409	0.0248	0.00485	0.00252	0.000053
	P10	2665.20	0.0237	0.00752	0.00489	0.000307
	P11	2638.396	0.337	0.0664	0.0247	0.000187
	P12	2611.125	0.0133	0.00394	0.00302	0.000090
	P13	2584.91	0.0145	0.0102	0.00981	0.00390
	P14	2557.09	0.0176	0.0180	0.0185	0.00335
	P15	2527.06	0.0145	0.0155	0.0161	0.000565
	P16	2498.02	0.0261	0.0282	0.0295	0.00103
2-1	P3	2750.05	0.0401	0.00898	0.00403	0.000074
	P4	2727.38	0.0378	0.00653	0.00272	0.000033
	P5	2703.98	0.00528	0.00171	0.00118	0.0000307
	P6	2680.28	0.0600	0.0139	0.00611	0.000069
	P7	2655.97	0.0535	0.0134	0.00667	0.000733
	P8	2631.09	0.00950	0.00348	0.00293	0.000761
	P9	2605.87	0.0311	0.00776	0.00455	0.000110
	P10	2580.16	0.282	0.0295	0.0311	0.00180
	P11	2553.97	0.0144	0.0163	0.0177	0.000883
	P12	2527.47	0.0140	0.0152	0.0158	0.000554
	P13	2500.32	0.0240	0.0265	0.0278	0.000072
	P16	2417.27	0.0811	0.0901	0.0943	0.00330
3-2	P3	2662.17	0.354	0.00790	0.00361	0.000047
	P4	2640.04	0.0437	0.00914	0.00424	0.000075
	P5	2617.41	0.00490	0.00276	0.00253	0.000090
	P6	2594.23	0.0118	0.00557	0.00480	0.000152
	P7	2570.51	0.0507	0.0560	0.0613	0.00557
	P8	2546.37	0.0322	0.0356	0.0379	0.00228
	P9	2521.81	0.0150	0.0164	0.0171	0.00599
	P10	2496.61	0.0319	0.0298	0.0307	0.00107
	P11	2471.34	0.0509	0.0491	0.0508	0.00184
	P12	2445.29	0.0659	0.0728	0.0756	0.00266

Table 5.5 *(continued)*

CO$_2$ Laser Parameters			Atmospheric Absorption Coefficients (km^{-1})			
			z = 0 km, Sea Level			z = 12 km
Band	Rot. ID	ν(cm^{-1})	k_{trop}	k_{mw}	k_{sw}	k_{mw}
	P40	924.970	0.514	0.0359	0.0112	0.000812
	P38	927.004	0.521	0.0423	0.0154	0.00164
	P36	929.013	0.744	0.0581	0.0190	0.00211
	P34	930.997	0.538	0.0536	0.0227	0.00311
	P32	932.956	0.557	0.0650	0.0302	0.00520
	P30	934.890	0.572	0.0737	0.0360	0.00677
	P28	936.800	0.588	0.0852	0.0440	0.00887
	P26	938.684	0.583	0.0853	0.0447	0.00955
	P24	940.544	0.603	0.0955	0.0517	0.0118
	P22	942.380	0.606	0.1021	0.0569	0.0136
	P20	944.190	0.609	0.0958	0.0521	0.0125
	P18	945.976	0.635	0.1223	0.0717	0.0186
	P16	947.738	0.572	0.0747	0.0378	0.00897
	P14	949.476	0.607	0.1101	0.0642	0.0173
	P12	951.189	0.591	0.1058	0.0619	0.0171
	P10	952.877	0.596	0.1008	0.0580	0.0161
	P8	954.541	0.553	0.0817	0.0452	0.0123
	P6	956.181	0.513	0.0615	0.0314	0.00810
	P4	957.797	0.484	0.0498	0.0236	0.00573
	P2	959.388	0.978	0.0753	0.0282	0.00609
	R0	961.729	0.456	0.0347	0.0130	0.00234
	R2	963.260	0.461	0.0401	0.0170	0.00367
	R4	964.765	0.478	0.0502	0.0241	0.00590
	R6	966.247	0.519	0.0614	0.0308	0.00783
	R8	967.704	0.505	0.0663	0.0352	0.00931
	R10	969.136	0.510	0.0714	0.0389	0.0104
	R12	970.544	0.578	0.0788	0.0418	0.0109
	R14	971.927	0.556	0.0796	0.0427	0.0110
	R16	973.285	0.554	0.0799	0.0425	0.0106
	R18	974.618	0.522	0.0755	0.0405	0.0101
	R20	975.927	0.194	0.2140	0.0740	0.0109
	R22	977.210	0.674	0.0871	0.0398	0.00803
	R24	978.468	0.503	0.0641	0.0318	0.00699
	R26	979.701	0.484	0.0579	0.0280	0.00585
	R28	980.909	0.474	0.0529	0.0245	0.00471
	R30	982.091	0.552	0.0587	0.0240	0.00378
	R32	983.248	0.454	0.0436	0.0183	0.00324
	R34	984.379	0.455	0.0439	0.0158	0.00229
	R36	985.484	0.436	0.0357	0.0133	0.00176

5.17.4 Bessel Functions[33]

Table 5.6 shows the Bessel functions J_0 and J_1 for arguments typical of the range of values used in directed energy analyses.

<div align="center">Table 5.6 Bessel Functions</div>

x	J_0	J_1	x	J_0	J_1	x	J_0	J_1
0	1	0	2.1	0.166607	0.568292	4.2	-0.37656	-0.13865
0.1	0.997502	0.049938	2.2	0.110362	0.555963	4.3	-0.36101	-0.1719
0.2	0.990025	0.099501	2.3	0.05554	0.539873	4.4	-0.34226	-0.20278
0.3	0.977626	0.148319	2.4	0.002508	0.520185	4.5	-0.32054	-0.23106
0.4	0.960398	0.196027	2.5	-0.04838	0.497094	4.6	-0.29614	-0.25655
0.5	0.93847	0.242268	2.6	-0.0968	0.470818	4.7	-0.26933	-0.27908
0.6	0.912005	0.286701	2.7	-0.14245	0.441601	4.8	-0.24042	-0.2985
0.7	0.881201	0.328996	2.8	-0.18504	0.409709	4.9	-0.20974	-0.31469
0.8	0.846287	0.368842	2.9	-0.22431	0.375428	5	-0.1776	-0.32758
0.9	0.807524	0.40595	3.0	-0.26005	0.339059	5.5	-0.00684	-0.34144
1.0	0.765198	0.440051	3.1	-0.29206	0.300921	6	0.15065	-0.27668
1.1	0.719622	0.470902	3.2	-0.32019	0.261343	6.5	0.26009	-0.15384
1.2	0.671133	0.498289	3.3	-0.3443	0.220663	7	0.30008	-0.00468
1.3	0.620086	0.522023	3.4	-0.3643	0.179224	7.5	0.26634	0.13525
1.4	0.566855	0.541948	3.5	-0.38013	0.137374	8	0.17165	0.23464
1.5	0.511828	0.557937	3.6	-0.39177	0.095459	8.5	0.04194	0.27312
1.6	0.455402	0.569896	3.7	-0.39923	0.053822	9	-0.09033	0.24531
1.7	0.397985	0.577765	3.8	-0.40256	0.0128	9.5	-0.19393	0.16126
1.8	0.339986	0.581517	3.9	-0.40183	-0.02724	10	-0.24594	0.04347
1.9	0.281819	0.581157	4.0	-0.39715	-0.066			
2.0	0.223891	0.576725	4.1	-0.38867	-0.10327			

5.18 WORKED EXAMPLES

5.18.1 Diameter of a High-Energy Laser Illuminator System

Consider a 10.6-μm high-power cw space-based illuminator with a desired brightness of 1.0×10^{14} W/sr. What should be the diameter of the system?

Beginning with the brightness equation [Eq. (5.111)] and its diameter equivalent [Eq. (5.112)], develop a top-level performance budget by applying the principle of equal allocation:

$$B = \frac{\pi P D^2}{4\lambda^2 BQ^2} , \tag{5.111}$$

$$D = 2\lambda BQ \left(\frac{B}{\pi P}\right)^{1/2} . \tag{5.112}$$
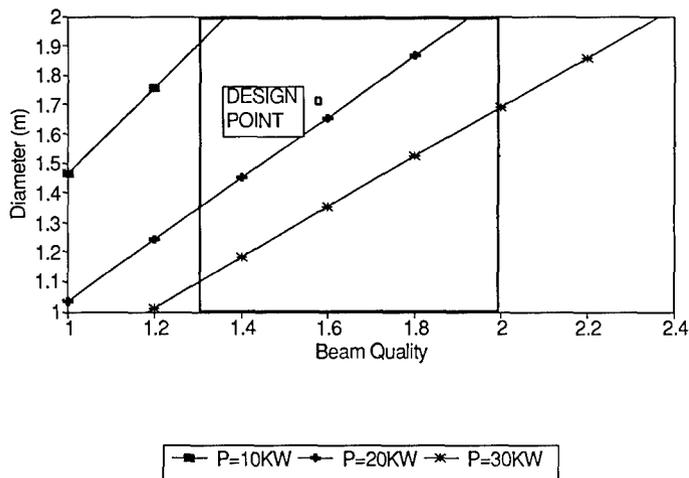
Fig. 5.28   Determination of illuminator diameter, $B = 1.5 \times 10^{14}$.

Each group of parameters at the same level of a budget is first given a value that has equal effect on brightness. Note that as the system design proceeds, the designer must equalize the difficulty of the design of the various elements (sometimes known as the *principle of equal pain*). Because the brightness and wavelength are fixed, the first level of design allocation is to determine the power, diameter, and effective beam quality. Note that power is net power leaving the beam expander.

Using the brightness equation [Eq. (5.112)], Fig. 5.28 shows the interrelationship among the three free parameters (for 50% margin above the specified brightness). Other system constraints and design experience determine allowable ranges of the parameters. These ranges are indicated by the horizontal and vertical maxima and minima lines for diameter and beam quality and the upper and lower of the power curves.

In this case, beam quality less than 1.3 is very difficult to achieve, whereas BQ over 2.0 is fairly simple to achieve. Laser packaging constraints determine the range of power. Mirror diameter is determined by a maximum packaging limit (2.0 m) and a recognition that a mirror less than 1.0 m is likely to overconstrain other parts of the system. Choosing round numbers near the center of the parameter space leads to the following approximate values (with 50% margin):

$$P = 19{,}000 \text{ W}, \quad D = 1.7 \text{ m}, \quad \text{BQ} = 1.6 .$$

There is no reason to calculate an exact set of values until other system constraints are considered.

What other constraints are typical of this type of system? Both the laser and the telescope are limited by the available volume. The volume depends on the $f/\#$ of the telescope and the power required from the laser. In addition,

the BQ depends on the $f/\#$ because faster systems have more aberrations. For this example, a simple set of arbitrary constraints is

$$V_t = \frac{\pi D^3}{4}(7.5)(f/\#) \ , \tag{5.113}$$

where V_t is the ratio of the volume required for the telescope to the total volume in percentages, D is the diameter in meters, and the $f/\#$ is dimensionless. The percent volume required for the laser is $V_I = 2kP$, where P is the power in kilowatts, and k is the volume per kilowatt, unity in this example. The beam quality is

$$BQ = \left(\frac{6}{f/\#}\right)^{1/2} \ . \tag{5.114}$$

For the first values picked,

$$BQ = 1.6, \quad \text{thus } f/\# = 2.34 \ ,$$

$$D = 1.7 \text{ m}, \quad \text{thus } V_t = 68\% \ ,$$

$$P = 19 \text{ kW}, \quad \text{thus } V_I = 38\% \ .$$

The total volume is 106%, exceeding available volume. To reduce the volume, the $f/\#$ must be decreased (power could be decreased, but laser volume is already small). Reducing the $f/\#$ to 2.2:

$$BQ = [6/(f/\#)]^{1/2} = 1.65 \ , \tag{5.115}$$

for

$$V_t = 13D^3 = 62\%,$$
$$D = 1.68,$$
$$B = 1.38 \times 10^{14} \text{ W/sr.}$$

To meet the volume constraint, we have given up some margin and have found an acceptable diameter. Note that the constraint equations vary with system design.

5.18.2 Required Accuracy of Measurement of Relative Target Velocity

Assume that a budget for boresight error includes a term for point ahead, and that term is set at 10% of the spot size on target. How accurately must the relative target velocity be known to meet the budget? Start with the system parameters to determine spot size on target. The example system has:

$$\text{output diameter} = D = 1 \text{ m} \ ,$$

$$\text{wavelength} = \lambda = 1.0 \times 10^{-6} \text{ m} \ ,$$

$$\text{edge fluence criterion} = \text{spot size} = 50\% \ ,$$

a diffraction-limited unobscured circular aperture, and relative target velocity of 1 km/s. From Fig. 5.18, the spot size (radius) corresponding to 50% edge fluence is about $0.5\lambda/D$, or 500 nrad, and 10% of the spot size is 50 nrad (radius), or 100 nrad (diameter). From Eq. (5.64), the point-ahead angle for a 1000 m/s relative velocity is 3330 nrad. The budget (100 nrad) is 3% of the point ahead, requiring a target relative velocity measurement of 3% (in this case, 30 m/s).

5.18.3   Resonator Wavefront Error

How accurately must the mirrors of a two-mirror resonator be fabricated if the overall resonator beam quality resulting from fabrication is to be 1.218? The resonator has a magnification of 3, a secondary mirror size of 5 cm, a length of 1 m, and operates at a wavelength of 1.06×10^{-6} m. First, calculate the Fresnel core of the resonator:

$$a_{core} = \sqrt{\lambda L} = [(1.06 \times 10^{-6})(1)]^{1/2} = 1.03 \times 10^{-3} \text{ m} . \qquad (5.116)$$

Next calculate the equivalent number of resonator passes:

$$N = \frac{\ln(a/2a_{core})}{\ln(M)} = \frac{\ln[0.05/2(1.03 \times 10^{-3})]}{\ln(3)} = 2.9 \approx 3 . \qquad (5.117)$$

The output BQ is related to the WFE by

$$BQ = \exp\left\{\frac{1}{2}[2\pi(WFE)]^2\right\} ,$$

$$WFE = \frac{(2 \ln BQ)^{1/2}}{2\pi} , \qquad (5.118)$$

$$WFE \approx 0.1 \text{ wave rms} .$$

For a two-mirror resonator,

$$WFE = 2\sqrt{N}(\delta_1^2 + \delta_2^2)^{1/2} , \qquad (5.119)$$

where the δ are the resonator mirror aberrations. At this resonator size, neither mirror is very difficult to fabricate, but to reduce the cost of the larger mirror, specify it at twice the error of the smaller mirror. Thus,

$$WFE = 2\sqrt{3}(\delta_1^2 + 4\delta_1^2)^{1/2} = 0.1 \text{ waves (rms, wavefront)} . \qquad (5.120)$$

Solving for δ_1,

$$2\sqrt{3}(5\delta_1^2)^{1/2} = 0.1, \quad \delta_1 = \left[\frac{1}{5}\left(\frac{0.1}{2\sqrt{3}}\right)^2\right]^{1/2} = 1.29 \times 10^{-2} . \qquad (5.121)$$

Thus,

$$\delta_1 = 1.29 \times 10^{-2} \, (\lambda/78) \; ,$$
$$\delta_2 = 2.58 \times 10^{-2} \, (\lambda/39) \; , \quad \text{rms } 1 \, \mu\text{m} \; . \tag{5.122}$$

The factor of 2 in Eq. (5.122) converts surface errors into wavefront error. Because the δ are rms, the P-V errors can be about five times larger. Mirror fabrication errors are rarely sinusoidal, so conversions between P-V and rms based on sinusoids are usually in error (i.e., P-V is not equal to $2\sqrt{2}$ rms). The ratio of P-V to rms is usually about 5 for fabrication errors and may be larger. Many optical shops specify mirrors in terms of P-V error at HeNe wavelength. The P-V errors at 633 μm are five times less because of conversion from rms and 1.68 times less because of specification at HeNe:

$$\text{SM} = \lambda/9.3, \quad \text{PM} = \lambda/4.7 \; . \tag{5.123}$$

The mirrors are inexpensive.

Acknowledgments

References

1. R. W. F. Gross and J. F. Bott, *Handbook of Chemical Lasers,* John Wiley & Sons, New York (1976).
2. M. Rokni, J. A. Mangano, J. H. Jacob, and J. C. Hsia, "Rare gas fluoride lasers," *IEEE Journal of Quantum Electronics* **QE-14,** 7 (1978).
3. O. R. Wood, "Pulsed molecular lasers," *Proceedings of the IEEE,* 380–397 (1974).
4. S. N. Suchard, "Lasers—where are we now and where are we going?" *Proceedings of the SPIE* **92,** 174–182 (1976).
5. W. R. Sooy, "Lasers and optics—an overivew," *Optical Engineering* **15**(2), 131–135 (1976).
6. John Soures, *Selected Papers on High Power Lasers,* SPIE Milestone Series, Vol. MS43, SPIE Optical Engineering Press, Bellingham, WA (1991).
7. Albert V. Jelalian, *Laser Radar Systems,* Artech House, Boston (1992).
8. C. G. Bachman, *Laser Radar Systems and Techniques,* Artech House, Boston (1981).
9. D. L. Begley and B. D. Seery, Eds., *Free Space Laser Communication Technology IV, Proceedings of the SPIE* **1635** (1992).
10. D. H. Douglas-Hamilton, A. Kantrowitz, and D. A. Reilly, *Laser-Assisted Propulsion Research, Progress in Astronautics and Aeronautics,* Vol. 61, p. 271, K. W. Billman, Ed., AIAA, New York (1978).
11. J. T. Kare, Ed., *Proceedings of the 1986 SDIO/DARPA Workshop on Laser Propulsion,* Lawrence Livermore National Laboratory, Livermore, CA (1987); also the *1987 Proceedings* (1990).
12. Edward E. Montgomery, *Proceedings of the SELENE FY91 Program Results and FY92 Program Kickoff,* NASA Marshall Space Flight Center (1991).
13. M. Born and E. Wolf, *Principles of Optics,* p. 462, Pergamon Press, Oxford (1975).
14. D. Schwamb, WJSA internal memorandum (Oct. 1991).
15. C. Lieto, "Standing wave resonator misalignment analysis," WJSA memorandum (1990); see also W. F. Krupke and W. R. Sooy, "Properties of an unstable confocal resonator CO_2 laser system," *IEEE Journal of Quantum Electronics* **QE-5**(12), 575–586 (1969).

16. D. Schwamb, et al., *OIST Final Report,* Rocketdyne (1984).

17. F. H. White, "P1(7) propagation," SAIC memorandum FHW-91-033 (1991); see also F. G. Gebhardt, "High power laser propagation," *Applied Optics* **15**, 6 (1976); and D. H. Leslie and P. S. Lebow, "Computed survey spectra of 2–5u atmospheric absorption," NRL Report 5168 (1983).

18. W. L. Wolfe and G. J. Zissis, *The Infrared Handbook,* Environmental Research Institute of Michigan, Ann Arbor (Revised 1985).

19. J. M. Elson, J. P. Rahn, and J. M. Bennett, "Relationship of the total integrated scatter from multi-layer-coated optics to angle of incidence, polarization, correlation length, and roughness cross-correlation properties," *Applied Optics* **22**(20), 3207 (1983).

20. M. C. Cielinski, "Near field propagator algorithm," WJSA Memorandum MC-88-08 (1988).

21. D. H. Leslie et al., "Cooled mirror technology assessment final report—task D," Contract DAAH01-83-D-A017 (May 1985).

22. Robert K. Tyson, *Principles of Adaptive Optics,* Academic Press, New York (1991).

23. V. N. Mahajan, "Zernike annular polynomials for imaging systems with annular pupils," *Journal of the Optical Society of America* **71**, 1 (1981).

24. M. Born and E. Wolf, *Principles of Optics,* 5th ed., p. 396, Pergamon Press, Oxford (1975).

25. For example, J. Cernius, "Infrared communication systems," in *The Infrared Handbook,* W. L. Wolfe, G. J. Zissis, Eds., pp. 23–41, Environmental Research Institute of Michigan, Ann Arbor (Revised 1985); see also L. M. Beyer, S. H. Cobb, and L. C. Clune, "Ensquared power for obscured circular pupils with off-center imaging," *Applied Optics* **30**, 25 (1991), which lists several more references (see bibliography).

26. Charles Yee, Personal communication; M. Bottema and R. A. Woodruff, "Third-order aberrations in Cassegrain-type telescopes and coma correction in servo-stabilizer images," *Applied Optics* **10**(2), 300–303 (1971); W. B. Wetherell, "The use of image quality criteria in designing a diffraction limited large space telescope," *Proceedings of the SPIE Seminar on Instrumentation in Astronomy,* pp. 45–79 (1972).

27. D. Malacara, *Optical Shop Testing,* John Wiley & Sons, New York (1978).

28. R. K. Tyson, *Principles of Adaptive Optics,* pp. 31–35, Academic Press, New York (1991).

29. V. I. Tatarski, *Wave Propagation in a Turbulent Media,* McGraw-Hill, New York (1961).

30. D. C. Smith, "High power laser propagation," *Proceedings of the IEEE* **65**, 12 (1977).

31. V. N. Mahajan, "Strehl Ratio for primary aberrations: some analytical results for circular and annular pupils," *Journal of the Optical Society of America* **72**(9), 1258 (1982).

32. Anthony J. LaRocca, "Atmospheric absorption," in *The Infrared Handbook,* W. L. Wolfe and G. J. Zissis, Eds., pp. 5-63–5-69, Environmental Research Institute of Michigan, Ann Arbor (Revised 1985).

33. M. Abramowitz and I. S. Stegun, *Handbook of Mathematical Functions,* pp. 355–434, Dover, New York (1972).

Bibliography

Acebal, R., et al., "Technical and theoretical support for the DARPA space laser program," DARPA Contract F29601-82-C-0058 Final Report (1983).

Alexander, W. B., "Specifying glass for laser applications," *Electro-Optical Systems Design* **7**(4), 12A-14, 16A-22 (April 1975).

Ameer, George A., et al., "Parameter analysis for ground based laser systems," DARPA Contract F30602-82-C-0035 Final Report (1985).

Anan'ev, Yu. A., "Unstable resonators and their applications (review)," *Soviet Journal of Quantum Electronics* **1**(6), 565 (1972).

Antoshkin, L. V., O. N. Emaleev, and V. P. Lukin, "Piezoelectric wavefront corrector," in *Pribory i Tekhnika Eksperimenta* **31**(5), 211–212 (1988). Translated in *Instruments and Experimental Techniques* **31**(5, pt. 2), 1349–1351 (1988).

Apollonov, V. V., S. A. Chyetkin, V. Yu. Khomich, and A. M. Prokhorov, "Adaptive force optical elements based on open pore structures," in *Proceedings of a Symposium, Laser Induced Damage in Optical Materials: 1982,* (NBS-SP-669) pp. 220–227, H. E. Bennett, A. H. Guenther, D. Milam, and B. E. Newman, Eds., NBS, Boulder, CO (1984).

Apollonov, V. V., et al., "Stability parameters of mirror surfaces used in laser reflectors," *Soviet Technical Physics Letters* **1**(6), 240 (1975).

Ballard, Stanley S., Stephen E. Brown, and James S. Browder, "Measurements of the thermal expansion of six optical materials, from room temperature to 250°C," *Applied Optics* **17**(7), 1152–1154 (1978).

Banks, D. S., "Toric cathoptrics," *Electronic Progress* **16**(2), 13–19 (1974).

Barakat, R. and E. Blackman, "The influence of random wavefront errors on the imaging characteristics of an optical system. II. Statistics of the transfer function," *Optica Acta* **20**(11), 901–915 (1973).

Barakat, R., "The influence of random wavefront errors on the imaging characteristics of an optical system," *Optica Acta* **18**(9), 683–694 (1971).

Basov, N. G., V. A. Danilychev, A. A. Ionin, and V. A. Sobolev, "Electron-beam-controlled carbon monoxide laser with an output energy up to 100 J," *Kvantovaya Elektronika,* Moskva **1**(11), 2527–2529 (1974). Translated in *Soviet Journal of Quantum Electronics* **4**(11), 1412–1413 (1975).

Beckers, J. M. and L. E. Goad, "Image reconstruction using adaptive optics," in *Instrumentation for Ground-Based Optical Astronomy Present and Future, The Ninth Santa Cruz Summer Workshop in Astronomy and Astrophysics,* pp. 315–336, Springer-Verlag, Berlin (1988).

Bennett, H. E., "Scattering characteristics of optical materials," *Optical Engineering* **17**(5), 480–488 (1978).

Bennett, H. and D. K. Burge, "Multilayer thickness uniformities required to meet wavefront error tolerances in laser mirrors," in *Laser Induced Damage in Optical Materials: 1980, Proceedings of a Symposium (NBS-SP-620),* pp. 356–368, H. E. Bennett, A. J. Glass, A. H. Guenther, B. E. Newman, Eds., NBS, Washington, DC (1981).

Berthold, J. W., III and S. F. Jacobs, "Ultraprecise thermal expansion measurements of seven low expansion materials," *Applied Optics* **15**(10), 2344–2347 (1976).

Berthold, J. W., III, S. F. Jacobs, and M. A. Norton, "Dimensional stability of fused silica, Invar, and several ultralow thermal expansion materials," *Applied Optics* **15**(8), 1898–1899 (1976).

Bittel, R. H., W. L. Casey, G. R. Doughty, and A. S. Roberts, III, "RF and laser space-based communication links: another perspective," *Proceedings of the SPIE* **996,** 20–27 (1988).

Borsare, E. C. and R. W. Dyer, "Beam control for space laser technology," Contract DAAH01-80-C-0664 Final Report, SDIO (1984).

Browne, J. C., "Applications of directed energy weapon technology to civilian problems," *Transactions of the American Nuclear Society* **59,** 5–6 (1989).

Buczek, C. J., R. J. Freiberg, and M. L. Skolnick, "Laser injection locking," *Proceedings of the IEEE* **61,** 1411 (1973).

Bystricky, K. M. and N. Harned, "Beam expanding telescope for alignment and system integration," *Proceedings of the SPIE* **330,** 33–38 (1982).

Cederquist, J. N., J. R. Fienup, C. C. Wackerman, S. R. Robinson, and D. Kryskowski, "Wavefront phase estimation from Fourier intensity measurements," *Journal of the Optical Society of America A (Optics and Image Science)* **6**(7), 1020–1026 (1989).

Cederquist, J., S. R. Robinson, D. Kryskowski, J. R. Fienup, and C. C. Wackerman, "Cramer-Rao lower bound on wavefront sensor error," *Optical Engineering* **25**(4), 586–592 (1986).

Chatterton, N. E. and Larry Stewart, "Space-based neutral particle beam concepts and definition study," SDC Contract DASG60-85-C-0069 (1989).

Copley, S. M., "Commercial applications of laser processing," *Thin Solid Films* **84**(4), 367 (1981).

Crockett, G. A. and D. A. Strange, "Computer model for evaluating synthetic aperture propagation," *Proceedings of the SPIE* **440,** 77–84 (1984).

Dabney, G., et al., "Beam control concept definition and technology integration synthesis report," MICOM Contract DAAH01-86-D-0008 Final Report (1989).

Dame, L., "New high resolution phase conjugated optical correctors for diffraction limited applications," *Proceedings of the SPIE* **679,** 177–180 (1986).

Davis, Alan N. and James R. Merritt, "Phased integrated laser optics technology—technology implementation study," USAF Contract F29601-85-C-0120 Final Report (1987).

Decker, D. L., J. M. Bennett, M. J. Soileau, J. O. Porteus, and H. E. Bennett, "Surface and optical studies of diamond-turned and other metal mirrors," *Optical Engineering* **17**(2), 160–166 (1978).

Dente, G. C., "Separating misalignment from misfigure in interferograms on off-axis aspheres,"

Proceedings of the SPIE **429,** 187–193 (1983).

Downie, J. D. and J. W. Goodman, "Optimal wavefront control for adaptive segmented mirrors," *Applied Optics* **28**(24), 5326–5332 (1989).

Dwyer, T. A. W., III, "Automatic decoupling of flexible spacecraft slewing maneuvers," in *Proceedings of the 1986 American Control Conference,* IEEE, New York (1986).

Eastman, J. M., J. A. Boles, and J. M. Soures, "Rochester's 3-TW Omega system," *Laser Focus,* p. 48 (Dec. 1978).

Ellerbroek, B. L., G. M. Cochran, and W. Moretti, "Image-plane phase sensing for phased array telescopes," *Proceedings of the SPIE* **440,** 85–90 (1984).

Gerry, E. T., et al., "Advanced weapons concept evaluation study," WJSA 84S-VA-193, W. J. Schafer Associates (20 Nov. 1984).

Glenn, P., "Set of orthonormal surface error descriptors for near-cylindrical optics," *Optical Engineering* **23**(4), 384–390 (1984).

Goldstein, G. R., "Free electron lasers as ground based space weapons," *AIP Conference Proceedings* **178,** 290–315 (1988).

Golnik, Gary, et al., "Analytical support for SBL beam control/ATP systems," USAF Contract F29601-86-C-0030 Task Report (1987).

Golnik, Gary, "LODE and SBL beam control technology support," U.S. Army Contract DAAH01-86-D-0008 DO 0003 Final Report (1986).

Golnik, Gary, et al., "Technology base analysis and assessment," SDIO Contract SDIO84-88-C-0056 Tasks 2, 14, 30 Final Reports (1989, 1990, 1992).

Golnik, Gary, et al., "LODE testing preparations," SDIO Contract MDA 903-85-C-0155 Final Task Report (1985).

Greenwood, D. P., "Power spectra of turbulence-induced wavefront errors," *Proceedings of the SPIE* **142,** 91–97 (1978).

Greenwood, D. P., "Shared apertures in tracking turbulence-induced wavefront tilt errors," *Conference on Laser and Electrooptical Systems,* p. 88, IEEE (1976).

Hardy, J. W., J. E. Lefebvre, and C. L. Koliopoulos, "Real-time atmospheric compensation," *Journal of the Optical Society of America* **67**(3), 360–369 (1977).

Hardy, J. W., "Instrumental limitations in adaptive optics for astronomy," *Proceedings of the SPIE* **1114,** 2–13 (1989).

Harvey, J. E. and G. M. Callahan, "Wavefront error compensation capabilities of multi-actuator deformable mirrors," *Proceedings of the SPIE* **141,** 50–57 (1978).

Hatheway, A. E., "Finite elements for evaluating optical system performance," *Proceedings of the SPIE* **518,** 145–149 (1985).

Hayes, C. L. and W. C. Davis, "High-power-laser adaptive phased arrays," *Applied Optics* **18**(24), 4106–4111 (1979).

Herendeen, R., N. Kreigh, and J. Wetzel, "Optical alignment equipment and methods for visible and IR imaging meteorological radiometer," *Proceedings of the SPIE* **1167,** 255–268 (1989).

Herrmann, J., "Least-squares wavefront errors of minimum norm," *Journal of the Optical Society of America* **70**(1), 28–35 (1980).

Holly, S. and L. James, Eds., *Adaptive Optical Components, Proceedings of the SPIE* **141** (1978).

Holly, S. and L. James, Eds., *Adaptive Optical Components II, Proceedings of the SPIE* **179** (1979).

Holmes, D. and P. Avizonis, "Integrated high power optical system analysis," *Journal of Defense Research* **7B**(1), 283 (Spring 1975).

Holmes, D. A. and P. V. Avizonis, "Approximate optical system model," *Applied Optics* **15**(4), 1075 (1976).

Horio, S. P., et al., "Strategic Defense System (SDS) phase 2 space-based laser (SBL) concepts study," USAF Contract F04701-82-C-0155 Final Report (1989).

Horio, S. P., et al., "Space-based laser concept and development definition," USAF Contract F04701-82-C-0155 Final Report (1988).

Hudgin, R., "Wavefront compensation error due to finite corrector-element size," *Journal of the Optical Society of America* **67**(3), 393–395 (1977).

Hulburd, B. and D. Sandler, "Segmented mirrors for atmospheric compensation," *Optical Engineering* **29**(10), 1186–1190 (1990).

Iorio-Fili, D., G. Misuri, and F. Scandone, "Theory of three-mirror telescopes with diffraction limited performance," *Optica Acta* **27**(8), 1035–1052 (1980).

Jacob, J. H. and J. A. Mangano, "Modeling the KrF laser discharge," *Applied Physics Letters* **28**(12), 724–726 (1976).

Jacobs, S. F., J. N. Bradford, and J. W. Berthold, III, "Ultraprecise measurement of thermal coefficients of expansion," *Applied Optics* **9**(11), 2477–2480 (1970).

Jacobs, S. F., M. Sargent, III, and M. O. Scully, *Adaptive Optics and Short Wavelength Sources,* Addison-Wesley, London (1978).

Johnson, R. L., J. T. Lewis, T. B. Smith, and J. K. Flemister, "A minicomputer-controlled diagnostic and display system for high-energy lasers," *Proceedings of the SPIE* **138**, 139–144 (1978).

Jones, E. D. and G. W. Gobeli, "Sandia four-beam laser system (fusion applications)," in *1973 IEEE/OSA Conference on Laser Engineering and Applications Digest of Technical Papers,* p. 92, IEEE, New York (1973).

Kappesser, Robert R. and Jeffrey B. Shellan, "Optical components requirement definition," MICOM Contract DAAH01-86-C-0009 Final Report (1989).

King, W. B., "Unobscured laser-beam-expander pointing system with tilted spherical mirrors," *Applied Optics* **13**(1), 21–22 (1974).

Klein, C. A., "Thermally induced distortion in high-energy laser systems," *Optical Engineering* **18**(6), 591–601 (1979).

Kline, Douglas C., "Technical synthesis requirements for space based laser large optics interfaces," MICOM Contract DAAH01-86-D-0009 Final Report (1989).

Kline, Douglas C., "Beam control systems analysis," MICOM Contract DAAH01-86-D-0008 Final Task Report (1988).

Knowlden, R. E., "Measurement in visible light of wavefront errors produced by dielectric-enhanced infrared reflectors: some calculations," *Proceedings of the SPIE* **288**, 78–82 (1981).

Kogelnik, Herwig and Tingye Li, "Laser Beams and Resonators," *Applied Optics* **5**(10), 1550 (1966).

Koryakovskiy, A. S., V. M. Marchenko, and A. M. Prokhorov, "Diffraction theory of the Talbot method and the diagnostics of wide-aperture wavefronts," in *Formation and Control of Optical Wavefronts,* pp. 43–126, P. P. Pashinin, Ed., Nova Science Publishers, Commack, NY (1989).

Krupke, W. F. and W. R. Sooy, "Properties of an Unstable Confocal Resonator CO_2 Laser System," *IEEE Journal of Quantum Electronics* **QE-5**(12), 575 (1969).

Legner, Hartmut H., et al., "Laser effects manual," DNA Contract DNA001-83-C-0186 Technical Report (1984).

Lemons, D. S. and M. E. Jones, "Theory and simulations of neutralization and focusing of ICF ion beams," *IEEE Transactions on Nuclear Science* **NS-32**(5, pt. 1), 2474–2476 (1985).

Lockheed Missiles and Space Co., "Space-based laser concept formulation and technology development planning," Contract F04701-82-C-0153 Final Report (1990).

Mahajan, V. N., "Included power for obscured circular pupils," *Applied Optics* **17**, 964 (1978).

Mahajan, V. N., "Imaging with noncentrally obscured circular pupils," *Journal of the Optical Society of America* **68**, 742 (1978).

Mahajan, V. N., "Zernike annular polynomials for imaging systems with annular pupils," *Journal of the Optical Society of America* **71**, 75 (1981).

Mahajan, V. N., "Zernike annular polynomials for imaging systems with annular pupils," *Journal of the Optical Society of America* **71**, 1408 (1981).

Marquet, L. C. and J. Hardy, "Historical review of adaptive optics technology," *Proceedings of the SPIE* **141**, 2–10 (1978).

Matoba, M., H. Nishimura, H. Toya, H. Fujita, K. Iba, S. Nakai, and C. Yamanaka, "High power E-beam controlled CO_2 laser system for laser fusion research," *Technology Reports of the Osaka University* **26**(1276–1307), 139–149 (1976).

Meinel, A. B., M. P. Meinel, P. K. Manhart, and E. B. Hochberg, "Active wavefront control challenges of the NASA large deployable reflector (LDR)," *Proceedings of the SPIE* **1114**, 406–409 (1989).

Meyer, T. W., R. L. Gullickson, B. J. Pierce, and D. R. Ponikvar, "The ground based free electron laser: challenges and solutions," *Nuclear Instruments & Methods in Physics Research, Section A (Accelerators, Spectrometers, Detectors and Associated Equipment)* **A285**(1–2), 369–373 (1989).

Miller, T. L. and R. D. Grigg, "Balanced thermal deflection approach for beam handling in medium power optical systems," *Proceedings of the SPIE* **518**, 150–154 (1985).

Miller, David E., et al., "ATP-FC and AMTB technology study," RADC Contract F30602-86-D-0083 Final Report (1989).

Monsler, Michael J. and Wayne R. Meier, "A ground-based laser system for strategic defense," MICOM Contract DAAH01-86-D-0008 Final Report (1989).

Monsler, M., "High power optics," *Optical Engineering* 15(2), 158–165 (1976).

Monsler, M., "Technical synthesis requirements for space based laser beam control interfaces," MICOM Contract DAAH01-86-D-0008 Final Report (1989).

Munitz, A., "Epitaxy and surface melting of aluminum alloys by high powered directed energy," *Metallurgical Transactions B (Process Metallurgy)* 11B(4), 563–573 (1980).

Myers, C. J. and R. C. Allen, "Development of an analytical mirror model addressing the problem of thermoelastic deformation," *Applied Optics* 24(13), 1933–1939 (1985).

Neal, D. R., J. R. Torczynski, and W. C. Sweatt, "Time-resolved wavefront measurements and analyses for a pulsed, nuclear-reactor-pumped laser gain region," *Optical Engineering* 29(11), 1404–1412 (1990).

Noll, R. J., "Effect of mid- and high-spatial frequencies on optical performance," *Optical Engineering* 18(2), 137–142 (1979).

Noll, R. J., "Phase estimates from slope-type wavefront sensors," *Journal of the Optical Society of America* 68(1), 139–140 (1978).

Noll, R. J., "Zernike polynomials and atmospheric turbulence," *Journal of the Optical Society of America* 66(3), 207–211 (1976).

O'Neill, F. and W. T. Whitney, "Tunable, high pressure infrared gas lasers," *Optics Communications* 18(1), 126 (1976).

Orza, J. M. and C. Domingo, Eds., *8th International Symposium on Gas Flow and Chemical Lasers, Proceedings of the SPIE* 1397 (1990).

Oughstun, K. E., "Intracavity compensation of quadratic phase aberrations," *Journal of the Optical Society of America* 72(11), 1529 (1982).

Pashinin, P. P., *Formation and Control of Optical Wavefronts,* Nova Science Publishers, Commack, NY (1989).

Pimentel, K. D., D. T. Gavel, and J. W. Roblee, "Preliminary findings for integrated modeling and simulation of directed energy weapons," *Proceedings of the 2nd European Simulation Congress,* pp. 581–586 (1986).

Pohle, R., et al., "Phased array system element definition (PHASED) study," USAF Contract F29601-82-C-0098 Final Report (1988).

Protz, R., "Technology of active optics," *Proceedings of the SPIE* 650, 109–117 (1986).

Rigrod, W. W., "Saturation effects in high-gain lasers," *Journal of Applied Physics* 36(8), 2487 (1965).

Riordan, J. T., "The role of lasers in the military," *Photonics Spectra* 18(8), 84 (1984).

Rivolta, C., "Airy disc diffraction pattern: comparison of some values of f/no. and obscuration ratio," *Applied Optics* 25, 2404 (1986).

Roberts, T. G. and J. F. Perkins, "Highly off-axis beam expanders with low aberrations," in *Gas Flow and Chemical Lasers, 1984, Proceedings of the Fifth International Symposium,* pp. 373–378, A. S. Kaye and A. C. Walker, Eds., Adam Hilger, Bristol, UK (1985).

Root, Robert G., "Rep pulse laser device scaleup evaluation," MICOM Contract DAAH01-86-D-0011 Final Report (1988).

Roychoudhuri, C., "PILOT two dimensional laser diode array," USAF Contract F29601-85-C-0140 Final Report (1989).

Rubin, I., et al., "Repetitively pulsed laser effects phenomenology program," USAF Contract F29601-86-C-0224 (1989).

Schelonka, L. P., "The fidelity of stimulated Brillouin scattering with weak aberrations," *Optics Communications* 64(3), 293–296 (1987).

Schuöcker, Dieter, Ed., *7th International Symposium on Gas Flow and Chemical Lasers, Proceedings of the SPIE* 1031 (1988).

Schoen, N. C., "Modal adaptive optics basis functions for obscured apertures," *Applied Optics* 23(10), 1525–1528 (1984).

Sepucha, Robert, "Concept definition and technology integration for directed energy weapon beam control systems," MICOM Contract DAAH01-86-D-0008 Final Report (1990).

Shellan, J. B., "Phased-array performance degradation due to mirror misfigures, piston errors, jitter, and polarization errors," *Journal of the Optical Society of America* 2(4), 555 (1985).

Shellan, J. B. and G. Zeiders, "Calculation of the near field phase in unstable resonators with mirror misfigures," *Applied Optics* **20**, 3172 (1981).

Shih, C. C., G. R. Karr, K. Smith, and F. Filingeri, "Investigations of the effects of flow medium homogeneity on laser beam jitters," in *Gas Flow and Chemical Lasers, 1984, Proceedings of the Fifth International Symposium,* pp. 379–384, A. S. Kaye and A. C. Walker, Eds., Adam Hilger, Bristol, UK (1985).

Papurt, D. M., J. H. Shapiro, and S. T. Lau, "Measured turbulence and speckle effects in laser radar target returns," *Proceedings of the SPIE* **415**, 166–178 (1983).

Durante, D. D. and C. G. Hunter, "Beam jitter transmission and correction," *Proceedings of the SPIE* **450**, 176–181 (1984).

Shirkey, R. C. and L. D. Duncan, "The electro-optical systems atmospheric effects library," *Proceedings of the SPIE* **302**, 19–23 (1981).

Siegman, A. E., "Effects of small-scale phase perturbations on laser oscillator beam quality," *IEEE Journal of Quantum Electronics* **QE-13**(5), 334 (1977).

Siegman, A. E., "Unstable resonators for laser applications," *Proceedings of the IEEE* **53**, 277 (March 1965).

Simmons, W. W., "Aberrations and focusability in large solid-state laser systems," *Proceedings of the SPIE* **293**, 27–35 (1981).

Small, J. G., W. O. Eckhardt, and F. Chilton, "Decoy discrimination using ground-based high power microwaves," *Proceedings of the SPIE* **1061**, 342–349 (1989).

Southwell, W. H., "Wavefront estimation from wavefront slope measurements," *Journal of the Optical Society of America* **70**(8), 998–1006 (1980).

Stamnes, J. J., H. Heier, and S. Ljunggren, "Encircled energy for systems with centrally obscured circular pupils," *Applied Optics* **21**, 1628 (1982).

Sterbenz, H. W., et al., "FY 1987 lethality and target hardening project final report," DNA Contract DNA001-85-C-0309 Final Report (1990).

Sterbenz, H. W., et al., "FY 1988 lethality and target hardening project final report," DNA Contract DNA001-89-C-0080 Final Report (1990).

Stolzman, D. E., "The perfect point spread function," in *Applied Optics and Optical Engineering,* R. R. Shannon and J. C. Wyant, Eds., Vol. 9, p. 113, Academic Press, New York (1983).

Stregack, J. A., "Handbook for space-based laser reference concept for directed energy weapons," MICOM Contract DAAH01-86-D-008 Final Task Report (1988).

Stregack, J. A., et al., "Beam control and FEL system update," MICOM Contract DAAH01-86-D-0008 Final Report (1989).

Thompson, K., "A graphic approach to the analysis of perturbed optical systems," *Proceedings of the SPIE* **237**, 127–134 (1980).

Triebes, Kenneth J., et al., "Bifocal integrated brassboard," AFWL Contract F29601-86-C-0236 Final Report (1990).

TRW Defense Systems Group, "Ground based laser concept definition for technology identification (GBL CDTI)," USAF Contract F04701-85-C-0082 Final Report (1989).

TRW Defense Systems Group, "GBL CDTI nearer-term GBL special study," USAF Contract F04701-85-C-0082 (1987).

Tschunko, H. F. A., "Imaging performance of annular apertures. 6. Limitations by optical surface deviations," *Applied Optics* **26**(24), 5333–5336 (1987).

Tyson, R. K., "Using the deformable mirror as a spatial filter: application to circular beams," *Applied Optics* **21**(5), 787–793 (1982).

Tyson, Robert K., "Three mirror optical path length corrector," in *Proceedings of the International Conference on Lasers,* p. 510–514, C. P. Wang, Ed., STS Press, McLean, VA (1986).

VanDevender, P. and J. Smiley, "Review of high power excimer lasers," *Proceedings of the SPIE* **1225**, 2–22 (1990).

Vieceli, James, et al., "Beam control performance requirements for boost phase intercept and midcourse target identification," MICOM Contract DAAH01-86-D-0008 Final Report (1989).

Vieceli, James, and Lawrence Zajac, "Directed energy weapon (DEW) technology assessment of requirements and cost analysis," MICOM Contract DAAH01-86-D-008 (1988).

Viswanathan, V. K., J. V. Parker, T. A. Nussmier, C. J. Swigert, W. King, A. S. Lau, and K. Price, "Use of adaptive optics element for wavefront error correction in the Gemini CO_2 laser fusion

system," *Proceedings of the SPIE* **190,** 251–257 (1979).

Vorontsov, M. A., G. M. Izakson, A. V. Kudryashov, G. A. Kosheleva, S. I. Nazarkin, Yu. F. Suslov, and V. I. Shmal'gauzen, "Adaptive cooled mirror for the resonator of an industrial laser," *Kvantovaya Elektronika,* Moskva **12**(7), 1337–1338 (1985). Translated in *Soviet Journal of Quantum Electronics* **15**(7), 888 (1985).

Walker, B. H., "Effect of small wavefront errors and aperture obscurations on system image quality," *Proceedings of the SPIE* **549,** 145–150 (1985).

Walton, Linwood G., et al., "Space based laser. Phase 3. Nearer term SBL update and Zenith Blue impact study," USAF Contract F04701-82-C-0153 Final Report (1989).

Walton, Linwood, G., et al., "Space based laser. Phase 3. Detailed nearer-term SBL," USAF Contract F04701-82-C-0153 Final Report (1989).

Washburne, T. P., M. J. Konopnicki, and T. D. Miller, "Far-field diffraction effects of large aperture laser systems due to polarization effects caused by multilayer dielectric mirror coatings," *Proceedings of the SPIE* **1166,** 340–345 (1990).

Washburne, T. P., M. J. Konopnicki, and T. D. Miller, "Polarization effects in large aperture laser systems with return wavefront sensors," *Proceedings of the SPIE* **1166,** 391–400 (1990).

Wells, W. H., "Modes of a tilted-mirror optical resonator for the far infrared," *IEEE Transactions on Quantum Electronics* **QE-2**(5), 94 (1966).

Wenhan, Jiang, Shufu Huang, Ning Ling, and Xubin Wu, "Hill-climbing wavefront correction system for large laser engineering," *Proceedings of the SPIE* **965,** 266–272 (1989).

Wetherell, W. B., "The use of image quality criteria in designing a diffraction limited large space telescope," *Proceedings of the SPIE* **28,** 45–79 (1972).

Wilson, D. H., "Structurally compatible surface vectors for wavefront aberration description," *Proceedings of the SPIE* **450,** 88–94 (1984).

Yoder, P. R., Jr., E. R. Schlesinger, and J. L. Chickvary, "Active annular-beam laser autocollimator system," *Applied Optics* **14**(8), 1890–1895 (1975).

Zajac, Lawrence, "Analysis of optical requirements for high energy laser deployment," MICOM Contract DAAH01-86-D-0009 Final Report (1989).

Zelazny, S. W., et al., "Modeling DF/HF cw lasers: an examination of key assumptions," *AIAA Journal* **16**(4), 297 (1978).

Holography

Emmett N. Leith
University of Michigan
Ann Arbor, Michigan

CONTENTS

6.1   INTRODUCTION

Holography is the process of recording both the phase and amplitude of a wave by interfering the wave with a reference wave, so that the phase is recorded in the form of interference fringes. The process was invented by Dennis Gabor[1-3] in 1947, who later received the Nobel prize for this work.

6.2   BASIC THEORY

6.2.1   Making the Hologram

A wave $u = |u| \exp(i\phi)$ (Fig. 6.1) is characterized by both amplitude and phase. The amplitude and phase structure may have been produced, for example, by the wave being transmitted or reflected from an object, and therefore represents information about the object.

The objective of holography is to record the wave field on a photosensitive medium (typically photographic film) with sufficient completeness that the wave field can later be regenerated from this record. To accomplish this objective, both the amplitude and the phase must be recorded. If the recording process has been carried out well enough, the regenerated, or reconstructed, wave field is indistinguishable from the original and therefore can be focused to form an image, just as could the original wave.

A complete record would of course also include the polarization, but this is usually unnecessary since in most cases a perfectly satisfactory image is obtained from a single polarization component. Ignoring the polarization simplifies not only the recording process but also the analysis, since a scalar theory is then adequate. The polarization of the original wave can be preserved and reconstituted when this is desirable.

The reference wave u_0 (Fig. 6.1) is usually a plane or spherical wave of uniform amplitude. By interfering with the signal wave u, it assists in the capture of u on a recording medium, even though the medium is insensitive to phase. The total wave field impinging on the recording surface P is $u_0 + u$.
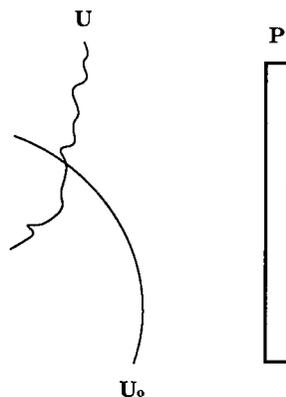
Fig. 6.1   Formation of a hologram. Interference between reference wave u_0 and signal wave u is recorded on a photographic plate.
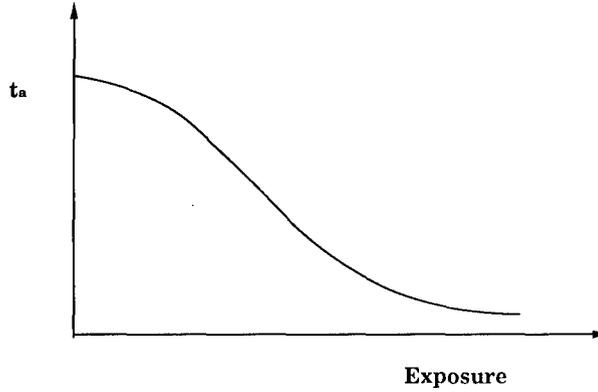
Fig. 6.2 A typical transmittance-exposure curve.

The recording medium responds to the irradiance $|u_0 + u|^2$ of the field. If the medium is photographic film, as is usual, the recorded signal becomes the amplitude transmittance of the developed film. The equation

$$t_a(x,y) = g[I(x,y)] = g[(u_0 + u)^*(u_0 + u)] \tag{6.1}$$

relates the transmittance t_a of the film to the irradiance of the exposing light. This relation is in general nonlinear, but a linear approximation is sufficient for most purposes. Gabor showed that a linear relation between t_a and I is optimum, and indeed such a linear relation can be realized as closely as we wish by choosing the ratio u_0/u sufficiently large; inspection of the transmittance exposure curve (Fig. 6.2) indicates that irradiance ratios as low as 2:1 or 3:1 may sometimes give reasonable linearity, although ratios of 5:1 may sometimes be more suitable. Equation (6.1) may therefore be written

$$t_a = k_0 - k_1 I = k_0 - k_1(u_0 u_0^* + uu^* + u_0 u^* + u_0^* u) , \tag{6.2}$$

where k_0 and k_1 are positive constants.

Equation (6.1) and in particular its linearized form, Eq. (6.2), are fundamental to holography. The central problem of holography is to regenerate the signal wave from the record, or hologram, these equations describe.

6.2.2   The Reconstruction Process

Let the hologram be illuminated by a wave that is identical (except for possibly a constant coefficient) to the reference wave u_0. The wave emerging from the hologram is then

$$u_e = u_0 t_a = k_0 u_0 - k_1 u_0^2 u_0 - k_1 u_0 uu^* - k_1 u_0^2 u^* - k_1 u_0 u_0^* u . \tag{6.3}$$

Imposing the condition that $u_0 u_0^* = a_0^2 = $ a constant, i.e., that u_0 is of uniform amplitude, the last term on the right side of Eq. (6.3) becomes $k_1 a_0^2 u$. This term, then, is a replica of the signal wave u and has all of the properties of

the original wave. In particular, this wave can form an image of the original object so exact that an observer sensing the wave would see it as if it were emanating from the original object.

For a perfect reconstruction, the term $k_1 u_0^2 u$ should be separable from the other terms. This separation had been a major problem in holography, and considerable effort was made in the early days of holography to solve it.

6.2.3 Carrier-Frequency (Off-Axis Reference Beam) Technique

The most successful method of separating the desired term from the spurious ones is that of introducing the spatial carrier, or off-axis reference beam.[4] Here the reference beam, typically a plane or spherical wave, impinges on the recording surface at an oblique angle to the signal wave u. The reference wave may be written as $a_0 \exp(i2\pi\alpha x)$, where α is a constant and the linear phase progression $2\pi\alpha x$ indicates that the wave is plane and strikes the plate obliquely in the x dimension. Certainly a spherical wave will do equally well, but is somewhat less convenient for analysis, although in the laboratory the spherical reference wave is more common. Also, let the signal beam be written $u = a \exp(i\phi)$. Equation (6.2) then becomes

$$t_a = k_0 - k_1 a_0^2 - k_1 a^2 - 2k_1 a_0 a \cos(2\pi\alpha x - \phi) , \qquad (6.4)$$

where the last two terms of Eq. (6.2) have been combined.

The last term of Eq. (6.4) is the one of interest. This equation lends itself to an interesting interpretation. The information $a \exp(i\phi)$, carried originally as amplitude and phase modulation of a monochromatic, spatially coherent light wave, has been transferred to a spatial carrier wave $\cos(2\pi\alpha x)$. The amplitude a modulates the amplitude of the spatial carrier, and the phase ϕ modulates its phase. To carry this interpretation further, the reference wave has functioned much like the local oscillator of a superheterodyne receiver, and the film, with its square-law characteristic, is analogous to a mixer.

The separability of the various terms of Eq. (6.4) depends on their spatial-frequency content. If for simplicity the signal u is assumed to have a uniform spatial-frequency content (along the x dimension) between the values $-W$ to W, the spectra of the various terms of Eq. (6.4) are as shown in Fig. 6.3. From
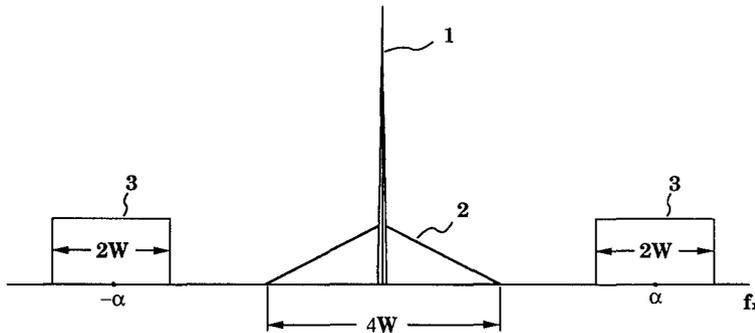
Fig. 6.3 Spatial-frequency spectrum of a hologram. Spectrum corresponding to term 1: $k_0 - k_1 a_0^2$; spectrum corresponding to term 2: $k_1 a^2$; spectrum corresponding to term 3: $2k_1 a_0 a \cos(2\pi\alpha x - \phi)$.
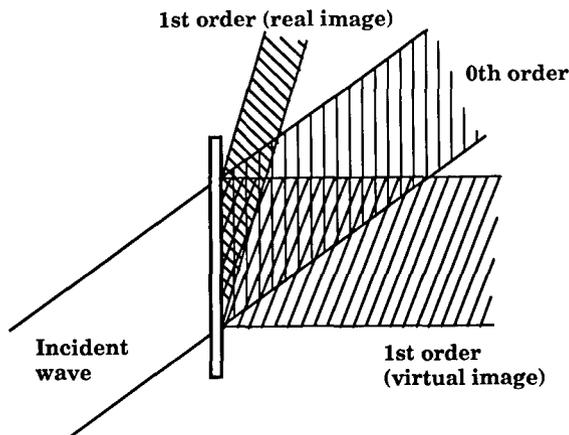
Fig. 6.4 The diffracted orders produced when the hologram is coherently illuminated.

this diagram it is apparent that the desired spectral separation can be achieved by choosing α sufficiently large.

In the readout process, the hologram is illuminated with a plane wave a_0 $\exp(i2\pi\alpha x)$. The hologram, by virtue of its fine-line structure, as embodied in the term $k_1 a_0 a \cos(2\pi\alpha x - \phi)$, behaves as a diffraction grating, producing a zero-order wave and a pair of first-order diffracted waves, as shown in Fig. 6.4.

If the wave emerging from the hologram is expanded in a Fourier series of plane waves, the angular spectrum of the emergent waves corresponds to the spatial-frequency spectrum of the recorded signal in accordance with the relation $f_1 = \sin(\theta_1)/\lambda$, where f_1 is a spatial-frequency component and θ_1 is the diffraction angle of the wave representing the component f_1.

The spatial-frequency spectrum of the recorded transmittance pattern is thus converted into an angular spectrum of diffracted waves. The zero-order spectrum is produced by the components $k_0 - k_1 a_0^2$ and $k_1 a^2$, and the two first-order spectra arise from the term $2k_1 a_0 a \cos(2\pi\alpha x - \phi)$. If this term is written in exponential form, we obtain two signals, one on the carrier $\exp(i2\pi\alpha x)$ and the other on the carrier $\exp(-i2\pi\alpha x)$. Each exponential term corresponds to one of the two first-order diffractions. By substituting in Eq. (6.3) for u_0 and u, using the relations $u_0 = a_0 \exp(i2\pi\alpha x)$ and $u = a \exp(i\phi)$, we obtain, for the two first-order diffractions,

$$-k_1 u_0 u_0^* u = -k_1 a_0^2 \exp(i\phi) , \tag{6.5}$$

$$-k_1 u_0^2 u^* = -k_1 a_0^2 a \exp(i2\pi\alpha x - \phi) . \tag{6.6}$$

As before, one of the orders [Eq. (6.5)] is identical, except in amplitude, to the signal wave from which the hologram was made, and this wave travels in such a direction that it eventually becomes separated from all other components. Thus, a complete optical reconstruction is obtained; the result is a virtual image, lying in the same position relative to the hologram that the original object had when the hologram was made.

The other sideband term [Eq. (6.6)] differs in two ways. First, it resides on a spatial carrier $\exp(i2\pi\alpha x)$; second, the phase term $\exp(-i\phi)$ is conjugate to that of the original wave. The carrier term $\exp(i2\pi\alpha x)$ indicates that this wave propagates in a different direction. The conjugate phase term gives rise to an image that forms on the side of the hologram plane away from the source and is therefore a real image. This image may sometimes be preferable to the primary, or virtual, image. For example, it can be photographed without a lens by placing a photographic plate in the plane of the image, or in whatever plane makes the best fit to a three-dimensional image; the entire holographic process, from object to final photographic record, is then performed without the use of imaging lenses.

The first two terms of Eq. (6.4), $k_0 - k_1 a_0^2$, represent a bias. Such a bias term is necessary if the function $-k_1 a_0 a \cos(2\pi\alpha x - \phi)$ is to be represented as the transmittance of photographic film, since this function has both positive and negative values, whereas photographic transmittance is inherently positive. The holographic process automatically generates this bias, and the reconstruction process carries it away in a direction such that it becomes separated from the reconstructed images.

The term $-k_1 a^2$ of Eq. (6.4) is an intermodulation term arising from the square-law nature of the recording process, which causes each spatial-frequency component of the signal to beat with all other components.

6.2.4   In-Line Holography

When $\alpha = 0$, the three diffracted waves of Fig. 6.4 become coincident and can no longer be separated. The real and virtual images must each be viewed with the other as a background. In addition, the remaining terms, the bias term $k_0 - k_1 a_0^2$ and the intermodulation term $k_1 a^2$, contaminate the holographic image.

This was the original configuration of Gabor.[1-3] Typically, a transparent object was used, and the reference beam was that portion of the light that was transmitted without scatter, and u was the scattered part. The transmittance is

$$t_a = k_0 - k_1 a_0^2 - k_1 a^2 - 2k_1 a_0 a \cos\phi \ . \tag{6.7}$$

This equation offers an interesting insight into the twin-image problem. The field u is a complex quantity, written in polar form as an amplitude part and a phase part, $u = |u| \exp(i\phi)$. Ordinary photographic processes record the amplitude $|u|$ but completely lose the phase part and half of the information is lost, assuming that the information is divided equally between the amplitude and the phase, an assumption that may or may not be the case.

Alternatively, a complex quantity can be written in rectangular form as a summation of a real part and an imaginary part, $u = a \cos\phi + i \sin\phi$. In the in-line method, the real part $a \cos\phi$ is captured, but the imaginary part is totally discarded. Hence again, only half of the information is recorded, so that the holographic process has not captured any more of the total field than has conventional photography; however, the mix is different. The missing part, $a \sin\phi$, results in the twin image.

Gabor proposed capturing the $a \sin\phi$ part by making two holograms, the second with the reference beam shifted by $\pi/2$ in phase, so that this hologram

would then record $a \sin\phi$. The two holograms would then be placed in two branches of a two-beam interferometer and combined with a $\pi/2$ phase shift. The resulting image contains the entire object wave $a(\cos\phi + i \sin\phi) = a \exp(i\phi)$. This method seems impractical, since to accurately maintain the quadratic relation between the two holograms, all of the interferometer components, including the hologram, would have to be optically flat to within much better than $\lambda/4$, and photographic film does not have such a degree of flatness. The carrier-frequency method achieves the same result in a simpler and more practical way, on a single hologram, without the need for stringent optical quality.

The holographic image, however, can be reasonably good despite these degrading effects, if proper precautions are observed. First, choosing a sufficiently strong coherent background ensures that the intermodulation term will be negligibly small. A background wave with intensity ten times that of the signal wave (i.e., a_0 about three to five times the mean value of $|a|$) is generally satisfactory.

For in-line holography, the object generally is constrained to be a transparency in which the major portion of the incident wave is transmitted without scattering. The coherent background is then a component of the holographic signal, and its presence in the holographic image is therefore proper. The object may be said to supply its own coherent background or reference wave. Such an in-line system is shown in Fig. 6.5(a). If the object does not include a coherent background wave, as for example in the case of transparent lettering against a dark background, the coherent background can be introduced by passing part of the coherent illumination around the object with an interferometer and recombining the two beams at the recording plane.

The twin-image problem can be reduced to tolerable proportions if the object is simple enough. For example, if the object is a single point scatterer, as in Fig. 6.5(b), and if we choose the real image for observation, the coherent background in the virtual image is of relatively low intensity, since its energy is spread over a relatively great area. As the number of scattering points increases, the relative intensity of the virtual image increases [Fig. 6.5(c)]. In the limit, when each possible resolution element in the object plane contains a scattering element and the object is of an extent comparable to the hologram aperture, the intensity due to the virtual image is the same as that of the real, or in-focus, image. In this case, the out-of-focus image is highly objectionable.

Inherent in the in-line method is a problem of restoring the proper bias level to the holographic image. The impingement of the signal $a_0 + a \exp(i\phi)$ on the recording plane results in the reconstructed image bias term $a_0(k_0 - k_1 a_0^2)$ and a virtual image term $-k_1 a_0^2 a \exp(i\phi)$. The former must be positive, since the transmittance t_a of the hologram is positive, just as was the original transparency. However, the latter term, the spatially varying part of the field, is reversed in sign. Hence, the holographic image is negative, just as in conventional photography. To produce a positive image, a second step is needed: the hologram must be copied, with due care that the overall process produces an acceptable bias level, so that the final recorded image has a bias level comparable to that of the original object.

In off-axis holography, the bias level of the signal, like the rest of the signal, is modulated onto a carrier and is reconstructed with proper polarity and

(a)

(b) .

(c)
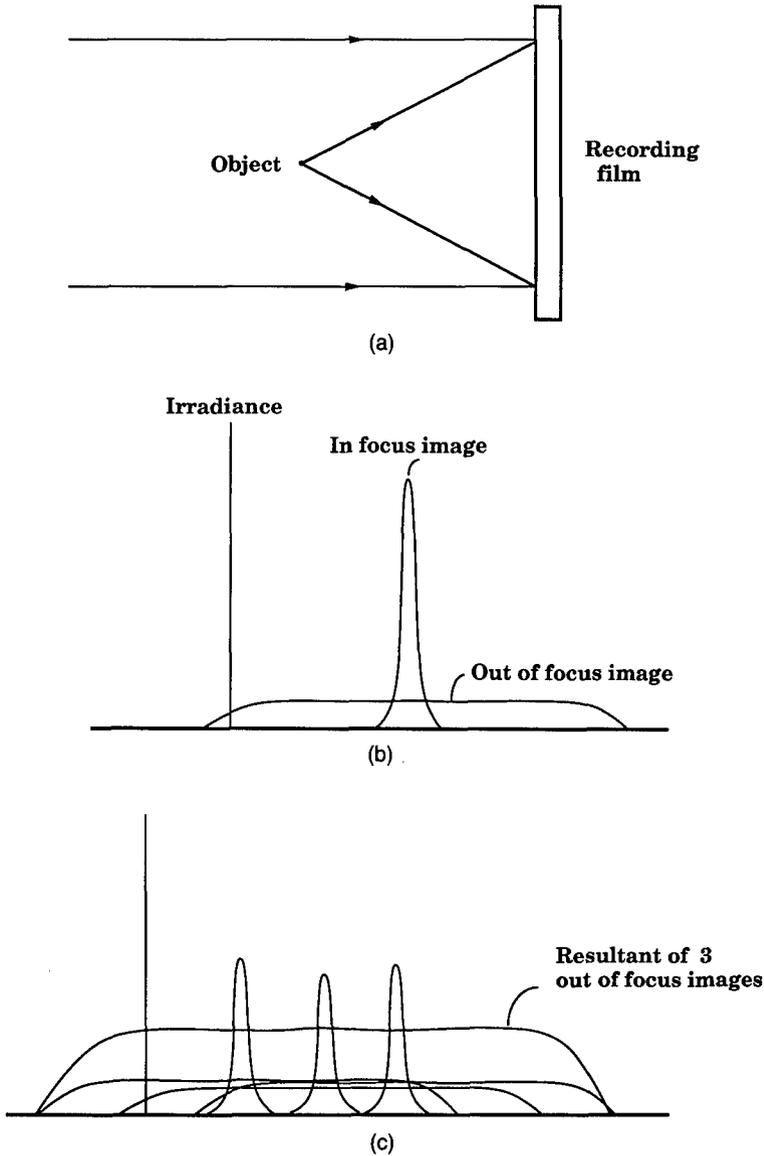
Fig. 6.5 In-line holography: (a) basic system; (b) superposition of in-focus and out-of-focus images, single-point object; (c) as in (b), but for a three-point object.

magnitude almost regardless of the recording characteristics; hence, the holographic image is positive, with very nearly the exact contrast of the object.

In-line holography lacks flexibilty, is restricted in the choice of object, and at best gives results inferior to those obtained by the off-axis method. Under the proper conditions, however, it gives reasonably good results. Also, it is simpler and requires less resolution in the recording medium than does the off-axis method.

6.2.5   Summary: Off-Axis Versus In-Line Holography

In most applications of holography, an off-axis reference beam is used. The advantages of this method over in-line holography are as follows:

1. Imagery is superior, since the various terms of the reconstruction are separated.
2. The method is more flexible, since there is little restriction as to choice of object.
3. High fidelity is achieved with smaller ratios of reference beam to object beam.
4. The holographic image is always a positive.
5. The method has less sensitivity to the transmittance-exposure characteristic of the film, first because the contrast of the image is highly insensitive to the contrastiness of the emulsion, and second because the error terms generated by nonlinearities are primarily higher harmonics, which produce higher order diffracted waves that propagate in a direction different from that of the image wave.

The disadvantages of off-axis holography are as follows:

1. Higher spatial frequencies must be recorded and, hence, higher resolution, slower film is required.
2. Since the reference beam follows a path different from that of the signal beam, the process is more sensitive to thermal, vibratory, and other instabilities, just as multiple-path interferometry is more sensitive to these effects than common-path interferometry. In some cases, the optical paths of signal and reference beams are nearly the same and this objection does not apply.
3. In all holography, instabilities tend to act as low-pass spatial filters. Some instabilities, such as lateral motion of the recording plate, act on the holographic signal after its modulation onto a carrier and therefore degrade the image more in off-axis holography.

It is sometimes assumed that off-axis holography, since it requires the recording of more fringes than does in-line holography, imposes more severe coherence requirements on the light source used for making the hologram. This assumption is incorrect. The use of interferometric techniques that produce achromatic and broad source fringes can reduce the coherence requirements of off-axis hologhapy exactly to those of in-line holography. The original method of off-axis holography used such an interferometer[4].

6.2.6   The Zone-Plate Viewpoint

The assumption that the reference and readout beams should be identical is unnecessary. The hologram produces a pair of images regardless of these restrictions. A difference between the two, either in curvature or in direction, alters the localization and magnification of the image and its aberrations. The first-order imaging properties of the image can be readily seen by approaching holography from the viewpoint of Fresnel zone plate theory.

The Fresnel zone plate is a well-known diffraction device with lens-like properties. Gabor, in his early papers on holography,[2,3] noted that the hologram
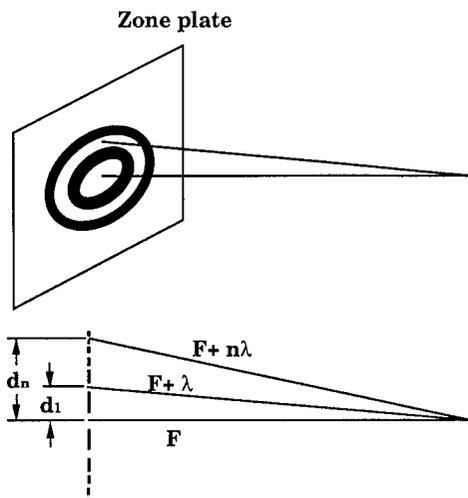
Fig. 6.6   Zone plate and the formation of the first-order positive image.

of a point source is a form of Fresnel zone plate, differing from the conventional zone plate primarily in having a smooth transition from transparent to opaque regions instead of the binary opaque-transparent characteristics of the latter. Rogers developed this viewpoint in some length [5]; the result was a simple and highly useful way of visualizing holographic imaging parameters.

The formation of a real image by a zone plate is shown in Fig. 6.6. The successive annular slit apertures of the zone plate are spaced so that, at some point along the zone-plate axis, contributions from successive annuli are retarded by one wavelength. By choosing the radii of these annuli in accordance with the relations shown in Fig. 6.6, we have

$$d_n = [(F + n\lambda)^2 - F^2]^{1/2} \cong (2Fn\lambda)^{1/2} ,$$

(6.8)

giving a focal length of

$$F = \frac{d_n^2}{2n\lambda} .$$

(6.9)

A virtual focus is found in a similar manner. Thus, the zone plate behaves simultaneously like a positive and a negative lens. There are also higher-order foci.

A point scatterer produces a hologram having amplitude transmittance

$$t_a = c_0 + c_1 \cos[K(x^2 + y^2)] = c_0 + c_1 \cos(Kr^2) ,$$

(6.10)

where c_0 and c_1 are constants. Points of maximum transmittance occur at $r = 0$, $(2\pi/K)^{1/2}$, $(2n\pi/K)^{1/2}$, and by analogy with Eqs. (6.8) and (6.9) the hologram has a focal length

$$F = \frac{\pi}{\lambda K} .$$ (6.11)

There are no higher order foci unless nonlinearities in the recording process cause the transmittance of the hologram to depart from the sinusoidal amplitude pattern.

In the formation of the holographic image, the zone plate focuses a portion of the incident light into a real image of the point source, and another portion into a virtual image of the source. The zone-plate lens obeys the basic lens equations for image-object plane, lateral object-image coordinates, and magnification.

Since an object can be approximated as a collection of point scatterers, we may similarly regard a hologram of an object as being composed of a summation of many zone-plate lenses, each one acting independently of the others to cause part of the incident light to be brought to a pair of point foci. The point images thus formed are laterally separated by an amount proportional to the displacement of the elementary zone plates, which in turn have separations proportional to the lateral coordinates of the object. Thus, the zone plates act in concert to form the holographic image.

By relating the focal length (or lengths) of the hologram, considered as a collection of zone plates, and using the lens maker's equation to find where the zone plates image the readout beam, the locations of the primary and conjugate images are readily obtained. Similarly, by considering two object points, with separation Δx, the separation Δx_h of the two zone plates they produce is readily found [Fig. 6.7(a)]. Similarly, Fig. 6.7(b) shows how to cal-

$$M_1 = \frac{\Delta X_h}{\Delta X}$$

(a)

$$M_2 = \frac{\Delta X_{im}}{\Delta X_h}$$

$$M_1 M_2 = \frac{\Delta X_h}{\Delta X} \frac{\Delta X_{im}}{\Delta X_h} = \frac{\Delta X_{im}}{\Delta X}$$

(b)
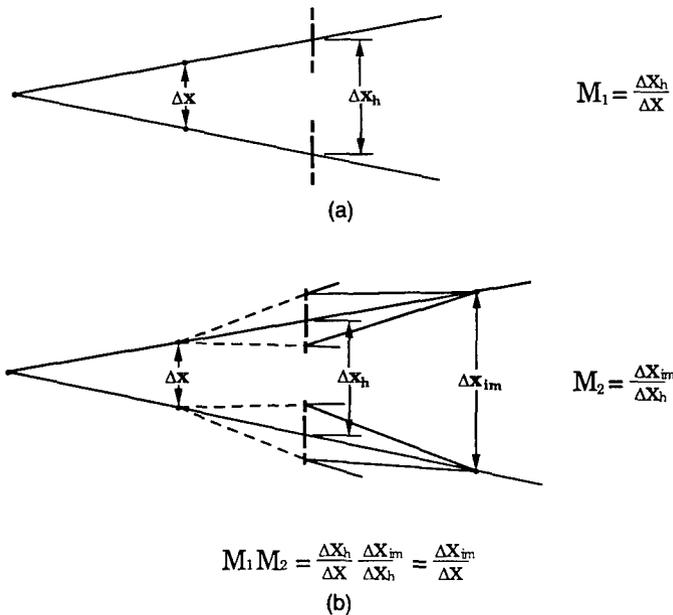
Fig. 6.7 Calculation of magnification using zone-plate theory: (a) forming the hologram and (b) the readout process.

culate the image separation Δx_{im}, both primary and conjugate. The magnification $\Delta x_{im}/\Delta x = (\Delta x_{im}/\Delta x_h)(\Delta x_h/\Delta x)$ is found from simple geometry to be

$$M = \left[\pm \left(\frac{z_p}{z_c}\right)\left(\frac{\lambda_1}{\lambda_2}\right)\left(\frac{1}{p}\right) - p\left(\frac{z_p}{z_0}\right) + p \right]^{-1}, \qquad (6.12)$$

where we have allowed for the hologram, after recording, to be scaled by a factor p; i.e., on the hologram, $x \rightarrow px$. In Eq. (6.12), z_0, z_p, and z_c are, respectively, the distance between the hologram and reference beam point-source planes, hologram and object planes, and hologram and image planes, λ_1 is the recording wavelength, and λ_2 is the readout wavelength.

The holographic image will, in general, have aberrations; these are just those of simple lenses: spherical aberration, coma, astigmatism, curvature of field, and distortion. The aberrations will be absent for two cases. The primary image will be aberration-free (for the usual case where $p = 1$) when the readout beam is identical to the reference beam and $\lambda_1 = \lambda_2$. The conjugate image will be aberration-free when the readout beam is conjugate to the reference beam, i.e., when it follows the same path, but counterpropagating. The magnification in each case is unity. Thus, magnification is always accompanied by aberration. For the more general case where $p = 1$, the primary image is aberration-free when the wavelength ratio is changed by the amount p and the readout beam is scaled by p. Thus, to obtain an aberration-free image with magnification of 2, the hologram should be expanded by a factor 2, the readout wavelength is doubled, and the readout beam satisfies the condition $z_c = 2z_0$.

6.3 HOLOGRAMS IN THREE-DIMENSIONAL RECORDING MEDIA

6.3.1 Bragg Diffraction

When the angle between the object and reference beams is sufficiently large, the fringes may have a spacing comparable to or less than the emulsion thickness. In such cases, the hologram must be analyzed as a 3-D structure instead of just as a surface.[6,7] The diffraction from such a structure is called Bragg diffraction. There are two basic regimes of Bragg diffraction: where the object and reference beams impinge from the same side of the recording medium and where they impinge from opposite sides.

A simple way to understand Bragg diffraction is to consider how a thick grating can evolve from a thin one. The thin grating of Fig. 6.8(a) obeys the basic diffraction-grating law

$$\sin\theta_i + \sin\theta_r = \frac{\lambda}{d} \qquad (6.13)$$

for the first diffracted order depicted in Fig. 6.8. If we allow the grating to become thick [Fig. 6.8(b)], the grating rulings become surfaces that reflect the light, so that now not only must the grating equation hold, but also the mirror law, $\theta_i = \theta_r$ must hold, which leads to the equation
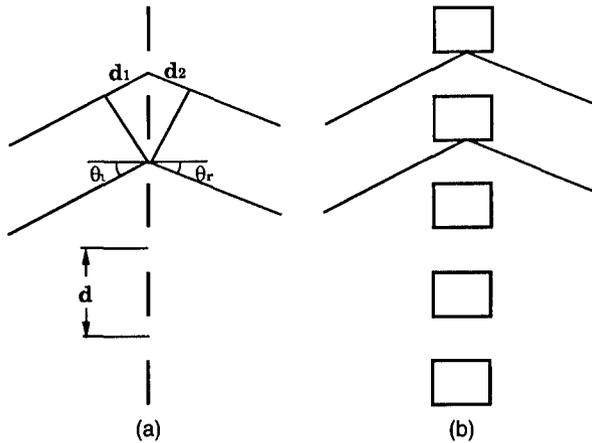
Fig. 6.8 (a) Thin grating diffraction and (b) thick grating diffraction.

$$2 \sin\theta_i = \frac{\lambda}{d} \ . \tag{6.14}$$

The principal difference between the above two equations is that the former allows for a wide range of incident angles θ_i, each producing a diffracted beam in a different direction, while the latter allows only one angle of incidence to produce diffraction, and if this angle (the Bragg angle) is not satisfied, diffraction does not occur.

6.3.2 Bandwidth of the Diffracted Light

When a beam of white light illuminates a hologram at the Bragg angle for the wavelength λ_0, the diffracted light will have a bandwidth $\Delta\lambda$, centered about λ_0. This bandwidth can be derived using the diagram of Fig. 6.9. A ray intercepts a fringe pattern recorded in a thick medium. Each fringe acts like a partially silvered mirror, reflecting a small portion of the light and letting the rest pass to the next surface. Thus, a diffracted ray is formed by constructive interference from many contributing surfaces. At the Bragg wavelength λ_0, the contributions from all the surfaces are in phase, giving a large resultant. A wave component slightly away from the Bragg angle has a small angle between the light diffracted from successive surfaces, giving a resultant that is somewhat reduced. At cutoff, the contributions sum to zero. Development of this view leads to the equation

$$\frac{\Delta\lambda}{\lambda_0} = \frac{1}{N} \ , \tag{6.15}$$

where N is the number of contributing surfaces.

More quantitatively, the regions of plane and volume holograms are defined in terms of Klein's Q-factor,

$$Q = \frac{2\pi\lambda T}{nd^2} \ , \tag{6.16}$$
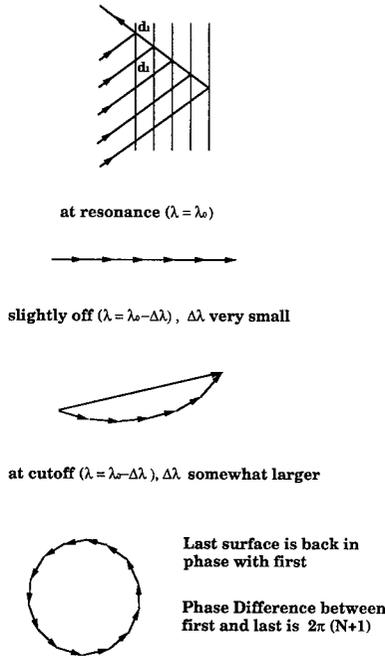
Fig. 6.9   Determining the bandwidth of Bragg-diffracted light.

where λ is the recording wavelength, T is the thickness of the hologram, n is the refractive index, and d is the grating spacing. If $Q \gg 10$, the hologram is clearly in the volume regime. If $Q \ll 1$, the hologram is in the thin region.

There is a nether region where the hologram is neither thick nor thin. This region is difficult to treat mathematically, but can sometimes be a useful one. Using this region, Angell[8] has made holographic gratings that yield high (about 40%) diffraction efficiency in both the +1 and −1 orders, a useful result for gratings used as beamsplitters; neither truly thick nor truly thin gratings could produce this result.

Volume holograms have various special properties. As noted, they have wavelength selectivity; the hologram, when illuminated with a beam of white light, produces a diffracted beam of narrow bandwidth. This property leads to white-light viewable holograms. The volume hologram also has angular selectivity, in that the readout beam must satisfy the Bragg angle to produce a diffracted beam. This property allows the storage of many holograms in the same recording medium, with readout of the various holographic images at different angles. Finally, volume holograms have the theoretical capability of 100% diffraction efficiency.[9] In practice, the achieved efficiency is often as high as 90%. Such high efficiency is impossible with thin holograms. The thin absorption hologram has a theoretical maximum efficiency of 6.26%, and if the hologram is bleached to form a pure phase hologram, the maximum possible

efficiency rises to 34%. The blazed phase grating can exhibit a much higher diffraction efficiency, but blazing the hologram is quite difficult and generally not practical. Thus, volume diffraction is of considerable importance in holography.

6.4 RECORDING MATERIALS FOR HOLOGRAPHY

The traditional recording medium for holography has been photographic film, as it remains today. In the meantime, other recording materials have come into widespread use and are challenging photographic film.

The photographic film selection for a holographic application depends on the resolution requirements, and these can vary considerably. For the basic Gabor in-line hologram of a simple transparency, the resolution requirements are exactly those of conventional photography, and conventional photographic films such as Panatomic-X are perfectly acceptable. At the other extreme, for display holograms in which a three-dimensional image is to be formed, with full parallax relations over a field of 30 deg or more, the resolution requirements can exceed 1000 cycles/mm. In the Denisyuk volume hologram, the resolution requirements are greatest of all, since fringe spacings of $\lambda/2$ must be recorded, leading to resolution requirements as high as 5000 cycles/mm. For display applications, special small-grain, extremely slow emulsions are available from Kodak (e.g., 649F, SO-254), Agfa-Gevaert (8E70, 10E70), and Ilford. Typically, the holograms are bleached, so that density variations are converted into phase variations by replacing the silver deposits with more transparent materials that produce phase delays. The diffraction efficiency is increased severalfold.

For many applications, especially holographic optical elements, dichromated gelatin is used instead of silver-halide emulsions. Typically, the dichromated gelatin is produced by fixing high-resolution plates to remove the silver and then ammonium-dichromate solution is diffused into the emulsion. The dichromated gelatin plates are light sensitive but are about 1000 times slower than even the fine-grain silver-halide films used in holography. The dichromated gelatin, after development, is a purely phase structure that yields very bright, low-scatter images. The resolution of dichromated gelatin is extremely high, in excess of 5000 cycles/mm. Since the dichromated gelatin is hygroscopic and the image is destroyed by water vapor, the dichromated plate holograms have cover plates cemented over them.

When holograms are to be made for mass production (see Sec. 6.9 on display holography), the hologram may be made on a photoresist, which produces a relief image from which nickel masters can be made for the production of mass-produced embossed holograms. The nickel master is used to press corresponding relief patterns onto thin sheets of plastic, which then become holograms and can be bonded onto paper or other backings.

Photopolymers are gaining popularity for holography. Photopolymer materials produced by Polaroid and by DuPont have yielded holograms of excellent quality having high diffraction efficiency in combination with low scatter.

With the improvements in electronic imaging, especially with low-noise CCD cameras with about 1000 × 1000 pixel elements, electronic holography is becoming important. The holograms can be made with a few milliseconds

exposure time, and the reconstruction process is carried out in a computer, with the reconstructed image being displayed on a monitor. Holograms thus can be made in rapid succession, and the camera surface is reusable almost without limit.

6.5 PHASE CONJUGATION

Holograms have the capability to form images through inhomogeneous media, including severely aberrated media and highly scattering media. This capability comes about in a variety of ways, and perhaps a half dozen distinctly different ways of holographic imaging through inhomogeneous media have been described. The earliest reported instance of holographic imaging through inhomogeneities is the holographic phase-conjugation method.[4]

The conjugate image had long been regarded as a problem, and considerable effort was expended in solving the problem. The earliest Leith and Upatnieks paper (1962)[4] not only proposed a very effective solution to the conjugate image problem, it also turned the situation around by offering a use for the conjugate image: image formation through inhomogeneities.

Figure 6.10 shows a hologram-forming system in which an irregular medium lies between the object and the hologram. This medium could be a turbulent gas, a lens with severe aberrations, or even a plate of diffusing glass.

For example, if the irregularity is a plate of frosted glass, as in the original reporting, the object cannot be seen from the recording position. The hologram can form an image only of what is visible from the hologram recording position. The hologram will, of course, form an image of the diffuser. Let the diffuser be described by a phase function $\exp[i\theta(x,y)]$, where $\theta(x,y)$ is a noise-like function that describes the phase retardation introduced by the diffuser. The real image formed by the hologram is $\exp[-i(\theta x,y)]$, since the real image is conjugate to the original object wave.

Placement of the original diffuser in the position where the holographic real image is formed results in the product $\exp(i\theta) \exp(-i\theta) = 1$. The diffuser and the real image superimpose and the phase of one is cancelled by the complementary phase of the other. The diffuser thus disappears and the object pre-
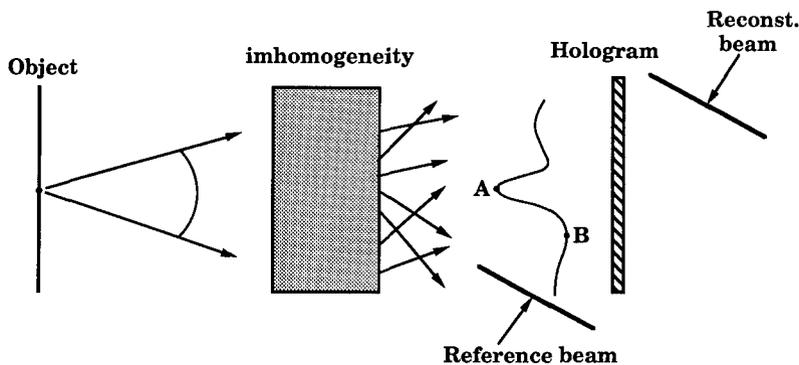
Fig. 6.10 Imaging through inhomogeneity with holography.

viously obscured is now revealed. Experimental results were reported in Refs. 10 and 11.

Although the results were interesting, little use was found for this technique for many years. One major reason is that the most useful applications, such as imaging through a turbulent atmosphere or flowing gases, require doing the process in real time, i.e., to shorter than the correlation time of the medium, which is typically about a millisecond.

In the 1970s, alternative means of producing the conjugate wave were found.[12] Four-wave mixing, using crystals, is analogous to holography and can be regarded as real-time holography in which the crystal is somewhat akin to photographic film, producing a hologram when both object and reference beams are present. The hologram disappears when one of the beams is removed.

Since the development of these essentially real-time materials, a new area known as phase conjugation has developed and has become an important area of optics.

In a variant of the phase conjugation process,[13] a nonobject beam (e.g., a plane wave) makes the first pass through the inhomogeneity, and the wavefront distortion is recorded holographically. Next, an object wave is passed through the hologram, where it acquires a phase distortion conjugate to that of the inhomogeneity-passed plane wave. This new wave then passes through the inhomogeneity. The expectation is that the second pass through the inhomogeneity compensates the phase errors produced by the first pass, producing a well-corrected image, and of course the object and image appear at opposite sides of the imaging system and of the inhomogeneity, which is a desirable arrangement. Although the inhomogeneity has been transilluminated twice, the object bearing beam has passed through only once, which is the justification for the appellation "one-way phase conjugation."

One might ask why it was necessary to pass the object wave initially through the inhomogeneity when a plane wave was all that was needed to produce the distorted wavefront for the conjugation process. The problem is that the two wavefronts, the distorted wavefront that makes the first pass and the distorted object wavefront that makes the second pass, are different and hence do not mutually compensate. If we decompose the object wave into a Fourier spectrum of plane-wave components, the one component that matches the first-pass plane wave will be perfectly compensated, but the other components will be imperfectly compensated. Thus, the technique gives only partial compensation.

This method was originally applied to the improvement of aberrations in a poor lens. A plane was passed through the lens and the hologram records the aberrations and becomes a phase corrector plate to be used with the lens.

6.6 HOLOGRAPHIC INTERFEROMETRY

6.6.1 The Concept

Holographic interferometry is one of the most important applications for holography. The concept is a fairly broad one with many aspects, but is essentially encompassed by the following statements. The wave generated by a hologram is a complete replication of the wave scattered from the original object, in both phase and amplitude. The holographic image can, therefore, be substituted for

the original object in any interferometric application. The holographic image can be interfered with the original object; alternatively, sequentially recorded holographic images can be interfered with each other. Since in the holographic process a wave field is coherently stored and released at a later time, we have the startling conclusion that, in effect, interference is now possible between two waves that existed at different times. Also, interference is possible between two waves that originally were at two different frequencies and were, therefore, mutually incoherent.

An early step toward holographic interferometry was made by Horman,[14] who observed that the image formed from a hologram could be substituted for the original object in interferometric measurements. By this means, a number of sequential interferometric tests can be made in a nonstationary object, with the object remaining unchanged throughout the tests. Proposed applications include flow patterns in a wind tunnel or a shock tube. With these dynamic situations holographically frozen, a sequence of tests involving schlieren photography, shadowgrams, and interferograms can be made with identical data.

One of the most significant forms of holographic interferometry is obtained by recording in sequence a number of holograms on a single plate. In the reconstruction, each separate exposure, or subhologram, produces an image; these images add coherently, thereby producing interference. This earliest form of hologram interferometry was described by Stetson and Powell,[15] who constructed holograms of vibrating objects.

Vibration measurement by holographic means is based on the recognition that two-beam holography is a form of interferometry, for which the usual stability requirements of interferometry apply. Elements of the object that move will cause a movement of the fringes produced by interference between those portions of the object and the reference beam. The fringes are therefore lost or attenuated in the recording; thus, the brightness of the holographic image is a function of the object movement during exposure.

Alternatively, we may think of the object as producing, at each incremental position, a separate installment, or subhologram. In the reconstruction process, each subhologram produces an image; the various images are slightly displaced, thereby producing intricate interference effects.

The object may have discrete motions during the exposure time, and in particular two separate exposures, each with the object in a slightly different position, may be made, in which case interference between the object at its two positions is obtained.

6.6.2   Theory

Let the object for the hologram be $s(x,y)$. When the object is in motion, we have

$$s(x,y) \rightarrow s[x_0 + x'(t), y_0 + y'(t), z'(t)] , \qquad (6.17)$$

where x_0, y_0, 0 are average coordinates of the object and x', y', z' are time-dependent variations about the position $(x,y,0)$. The virtual image term of Eq. (6.2) becomes

$$u_v = \frac{-k_1 u_0^*}{t_r} \int_0^{t_r} u(t)\, dt \ , \tag{6.18}$$

which, to a first-order approximation, can be written

$$u_v = \frac{K}{t_r} \int_0^{t_r} \int \int s(x_0 + x', y_0 + y', z') \exp\left[\frac{i2\pi r(t)}{\lambda}\right] d(x_0 + x')\, d(y_0 + y')\, dt \tag{6.19}$$

(see Fig. 6.11), where the bulk delay $2\pi z/\lambda$ between object and image has been neglected, K is a constant, and r has been expanded in the paraxial form

$$r = z' + \frac{1}{2z}(x - x_0 - x')^2 + \frac{1}{2z}(y - y_0 - y')^2 \ . \tag{6.20}$$

We see that the reconstructed wave front is indeed a summation of the wave fronts produced by the object at each of its positions, and therefore the hologram, as noted, reconstructs an image in each of the positions occupied by the object. These images, being coherent with one another, produce interference effects.

Of special interest is the case $x' = y' = 0$, producing

$$u_v = \frac{K}{t_r} \int \int \int_0^{t_r} s(x_0, y_0, z') \exp\left[i\left(\frac{2\pi z'}{\lambda} + \frac{p}{\lambda z}\right)\right.$$

$$\left. \times\, (x - x_0^2 + y - y_0)^2\right] dx_0\, dy_0 \ . \tag{6.21}$$

The reconstruction is of an effective object

$$s_{\text{eff}} = \frac{1}{t_r} \int_0^{t_r} s(x_0, y_0, z') \exp\left(\frac{i2\pi}{\lambda}\right) z'(t)\, dt \ , \tag{6.22}$$
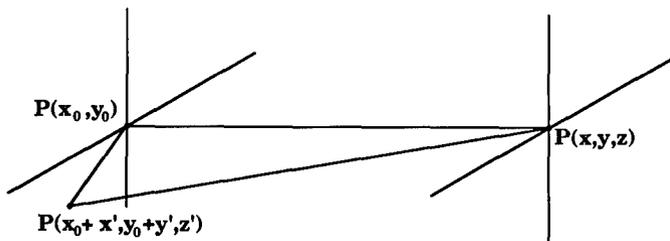
Fig. 6.11 Hologram interferometry. An object point has a rest position $P(x_0, y_0)$; vibratory motion gives it a displacement $P(x_0 + x', y_0 + y', z')$. Point $P(x,y,z)$ is a point on the hologram recording plane.

which, for small displacement z', can be written

$$s_{\text{eff}} = \frac{1}{t_r} s(x_0, y_0) \int_0^{t_r} \exp\left(\frac{i2\pi}{\lambda}\right) z'(t) \, dt \ . \tag{6.23}$$

6.6.3 Vibration (Time-Average) Holography

For a sinusoidally vibrating object, we have

$$z' = 2m(x_0, y_0) \cos(2\pi f_1 t) \ , \tag{6.24}$$

where m is the amplitude of the vibration (a function that may vary over the object) and f_1 is the frequency. Equation (6.23) then becomes

$$s_{\text{eff}} = J_0\left[\left(\frac{4\pi}{\lambda}\right) m(x_0, y_0)\right] s(x_0, y_0) \ . \tag{6.25}$$

The effective object then becomes the stationary object multiplied by a weighting function related to the amplitude of the vibration (but not its frequency). The zero-order Bessel function has a primary maximum at $m = 0$, along with periodic secondary maxima. The reconstructed image has superimposed on it contours of constant amplitude, corresponding to the zeros and secondary maxima of the Bessel function.

Examples of amplitude contours are shown in Fig. 6.12, which shows the holographic image of a vibrating surface that has been held stationary about its edge. The brightest areas are the nodes, and by counting the fringes lying between the node and a selected point on the surface, the amplitude of the vibration at that point is readily determined.

The fringes bear a resemblance to the Chladni figures produced by placing sand on a vibrating plate; the sand then migrates to the nodes. The holographic process, however, indicates not only the nulls, but numerous other constant amplitude loci as well.

The holographic method has the following advantages. It is applicable to any object, regardless of the complexity of its shape, and to any surface, whether

Fig. 6.12 Vibration contours on vibrating plate for various vibrating amplitudes (from Ref. 15).

smooth or diffusely reflecting. The object under analysis need not be disturbed; for example, no sensing mechanisms need touch the vibrating structure.

Equation (6.23) is independent of the frequency of the vibration. The intensity of the reconstructed image is determined only by the relative total duration that the object spends at each location and is independent of the way this distribution occurs as a function of time. This is obvious if we consider the holographic image as a summation of subimages produced by the sequence of subholograms, since the final result is certainly independent of the order in which the increments are recorded. This independence is explicitly stated by rewriting Eq. (6.23) as

$$s_{\text{eff}} = s(x_0, y_0) \int_{-\infty}^{\infty} \exp\left(\frac{i2\pi}{\lambda}\right) p[\xi(t)] \, d\xi \ , \tag{6.26}$$

where $p(\xi)$ is a weighting function describing the duration of the object in each incremental region $d(z')$. In particular, as noted by Kozma (in Stetson and Powell[15]), if the function of time z' is a stationary and ergodic random process, then Eq. (6.26) is equal to Eq. (6.23) and is an ensemble average, where $p[\xi(t)]$ is the probability density function of the random process $z'(t)$. This integral is the Fourier transform of the probability density function $p[\xi(t)]$ and is the characteristic function of $p[\xi(t)]$ evaluated at $(2\pi/\lambda)m(x_0, y_0)$.

6.6.4 Double-Exposure Holography

A special case of considerable importance is that of two discrete exposures; two related wave fields are then summed coherently and interference between the two is observed. For example, let the object position be $z' = 0$ for $t = 0$ to $(1/2)t_r$ and $z' = z'(x_0, y_0)$ for $t = (1/2)t_r$ to t_r. Then Eq. (6.23) becomes

$$s_{\text{eff}} = \frac{1}{2} s(x_0, y_0) \left\{ 1 + \exp\left[\left(\frac{i2\pi}{\lambda}\right) z'(x_0, y_0) \right] \right\} \ , \tag{6.27}$$

or

$$|s_{\text{eff}}|^2 = \frac{1}{2} |s(x_0, y_0)|^2 \left[1 + \cos\left(\frac{2\pi}{\lambda}\right) z'(x_0, y_0) \right] \ . \tag{6.28}$$

Equation (6.28) represents the interference between the image at its two different positions. As an example, one might suppose that between exposures the object had been distorted in some manner, for example, by the application of stress, and from the fringe pattern on the holographic image one could determine how the object had been distorted.

It is evident that the two exposures may be carried out with light of different polarizations or of different wavelengths while retaining interference between the two images.

6.6.5 Real-Time Holographic Interferometry

Another kind of hologram interferometry occurs when the hologram is interfered against the original object. The hologram is replaced at the position where

it was exposed and the experimental setup is undisturbed. The virtual image then coincides with the original object, and the observer looking through the hologram sees the two in superposition. Again, the waves add coherently and any mismatch between object and image is manifest as a fringe pattern.

This method has the powerful advantage of providing the interferometric analysis on a real-time basis. Along with this, there is the disadvantage that careful repositioning of the hologram is required when it is replaced in its original position. Failure to achieve accurate repositioning results in residual fringes, and in practice some residual fringing seems unavoidable. There are ways to develop the hologram *in situ* so that repositioning is unnecessary. For example, the recording plate can be placed in a tank and development and fixing can then be carried out without moving the plate. Alternatively, non-photographic film materials and techniques may be used, such as thermoplastic film or electronic holography. In the latter case the hologram is made on the detector of an electronic camera and the reconstruction is carried out in a computer. The result is then displayed on a monitor. In cases where high resolution is not needed, this method is very attractive.

Since its development in the mid-1960s, holographic interferometry in its various forms has been used successfully as a science and engineering tool.

6.7 COMPLEX SPATIAL FILTERING

6.7.1 Construction of Holographic Complex Spatial Filters

One of the earliest applications of holography was the construction of complex spatial filters in which any arbitrary phase and amplitude characteristic can be realized.

Figure 6.13 shows a spatial filtering system. An object distribution s_0 is to be imaged onto the image plane, and a mask H is placed at the common focal plane of the two imaging lenses. Since under coherent illumination the light distribution in this plane is Fourier-transform related to the object distribution, this plane is called the frequency plane and a mask placed here is a spatial filter. The filter modifies the spatial frequency content of the object spatial-frequency spectrum, forming at the image plane a modified image s_i. A stop
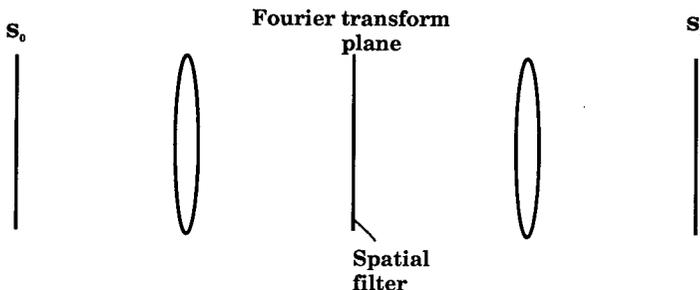
Fig. 6.13 A spatial filtering system.

on axis, for example, removes the zero spatial content of the object, producing what in microscopy is termed dark-field illumination. Another common and important spatial filter is the Zernike phase-contrast filter, which by shifting the zero spatial-frequency component of s_0 by $\pi/2$ converts phase objects into amplitude images, which are then rendered visible.

So that optical spatial filters have the same broad range of filtering capabilities that are attainable with electrical filters, it would be desirable to produce masks of arbitrary form $H = |H| \exp(i\phi)$, where $|H|$ and ϕ are arbitrary functions of the spatial frequencies f_x and f_y. In concept the way to produce such a filter is to produce a mask $|H|$ containing the desired transmittance function, and a phase mask with the phase delay $\exp(i\phi)$, produced as thickness variations. The two masks are combined and placed in the frequency plane of the optical system, where the combination becomes a complex spatial filter. In this way, an arbitrary transfer function can be generated. The construction of such a filter is not practical, since it is extremely difficult to fashion a phase plate that has, over its entire surface, very precisely the proper phase delay, except in very simple cases, such as the previously noted Zernike phase plate.

Holography provides exactly this capability. The desired transfer function $H = |H| \exp(i\phi)$ can be synthesized as a real function with transmittance $t_a = a_0 + |H| \cos(2\pi df_x + \phi)$, where d is a constant, which is, of course, just an off-axis hologram. Since the mask is to be used in the frequency plane of an optical system, we use spatial-frequency variables f_x, f_y. Since the relation between spatial frequency f_x and a physical coordinate x is $f_x = x/\lambda F$, where F is the focal length of the lens and λ is the wavelength of the light, we could write the mask in physical coordinates as $a_0 + |H(x/\lambda F, y/\lambda F)| \cos[(2\pi dx/\lambda F) + \phi(x/\lambda F, y/\lambda F)]$.

This mask could be constructed by computer and computer graphics, thus making it a computer-generated hologram. Alternatively, if the function h, the inverse transform of H, is available as a photographic transparency, then H could be formed as a Fourier-transform hologram of h, a process that leads to a matched filter for the signal h.

The hologram is placed in the frequency plane of an imaging system in which an object s_0 is imaged. The field impinging on the mask t_a is $S_0(f_x, f_y)$, the Fourier transform of s_0. The field emerging from the mask is

$$s_0 t_a = s_0 a_0 + s_0 |H| \cos[2\pi df_x + \phi(f_x, f_y)] \ , \tag{6.29}$$

and the image is the inverse transform of this, or

$$u_{\text{image}} = \mathcal{F}^{-1}\{S_0 t_a\} = aS_0 + \frac{1}{2}[S_0 H \exp(i2\pi df_x) + S_0 H^* \exp(-i2\pi df_x)] \ . \tag{6.30}$$

After performing the indicated inverse transformation, there results

$$u_{\text{image}} = a_0 s_0 + \frac{1}{2}\Gamma(x + d, y) + \frac{1}{2}\Omega(x - d, y) \ , \tag{6.31}$$
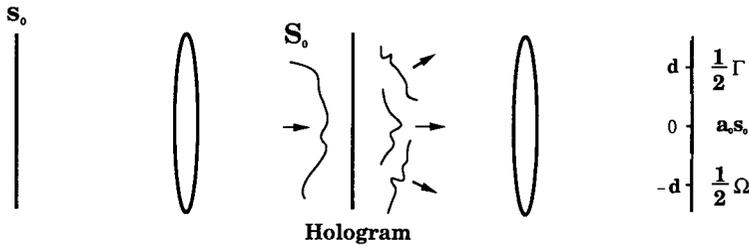
Fig. 6.14   Use of a hologram as a matched filter.

where $\Gamma = s_0 * h$, the convolution of s_0 and h, and is centered at $(-d, 0)$ and Ω is the cross correlation of s_0 and h, centered at $(d, 0)$. The filtering process is depicted in Fig. 6.14. The hologram, behaving like a diffraction grating, has produced three diffracted orders, corresponding to the zeroth diffracted order, the primary image, and the conjugate image of a conventional hologram. The undiffracted beam leads simply to an unaltered image of the object, the primary image yields a convolution with h, and the conjugate image yields the cross correlation of s_0 with h. Thus, we generate not only the required filtering operation $S_0 \rightarrow S_0 H$, but also the additional filtering operation $S_0 \rightarrow S_0 H^*$. Provided d is larger than the lineal dimension of the object along the x axis, the three terms will completely separate.

If the function h is available as a photographic transparency, the spatial filter can be made as a conventional Fourier-transform hologram, whereby a lens forms H, the Fourier transform of h, and an off-axis reference beam then overlays H, thus forming a Fourier-transform hologram. Alternatively, if h is not conveniently available, the hologram can be made by computer.

6.7.2   The Matched Filter

One of the best examples of the above principles is the realization of spatial matched filters. The matched filter is used to detect signals in noise. It can be shown that the filter that best discriminates against noise, in the sense of maximizing the peak-signal-to-rms-noise ratio when the noise is white and additive, is the matched filter.

A filter is matched to a signal g when it has the form G^*; that is, the matched filter has a transfer function that is the complex conjugate of the signal spectrum. The output of the filter is the autocorrelation function of g. Since this tends to be very narrow, the filter thus concentrates the output signal into a small area (Fig. 6.15). In this small area, the signal buildup is great and the signal thus tends to rise above the noise. The noise is relatively unaffected, since it does not correlate with the signal; however, it is diminished somewhat.

The use of a Fourier-transform hologram as a complex filter with full amplitude and phase control and in particular as a matched filter was developed in the period 1961–1962 at the University of Michigan Willow Run Laboratories (now ERIM). Such filters were required for the optical processing of synthetic aperture radar data, and the above ideas were conceived and devel-
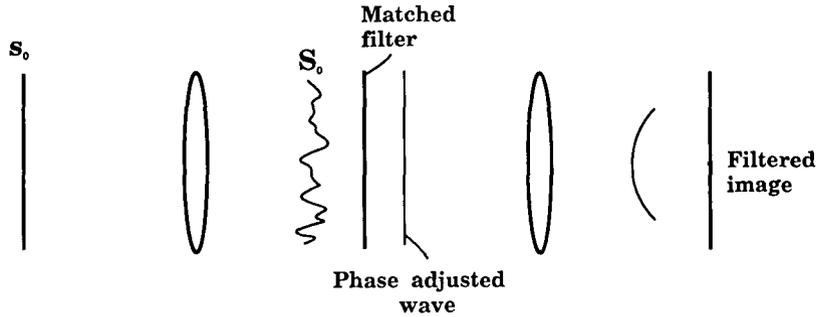
Fig. 6.15 The operation of a matched filter.

oped by a number of researchers. The basic idea of using a Fourier-transform hologram of an object as a complex spatial filter, in particular as a matched filter for that object, was suggested by C. Palermo. The earliest experimental results were produced in 1961 by Kozma and Kelley,[16] who made spatial matched filters for compressing radar pulses of the shift-register type. In 1962, VanderLugt[17] extended the application area to character and pattern recognition and produced some very sophisticated filters, i.e., filters of very high space-bandwidth product.

Since then, holographic spatial filter theory has developed considerably and over the years has figured significantly in optical-processing research. Although this branch of holography has not been developed commercially to the same extent as other branches of holography, such as holographic interferometry and display holography, it remains a powerful technique.

6.8 HOLOGRAPHIC OPTICAL ELEMENTS (HOEs)

Holograms have been used as a new kind of optical element, serving as beam-splitters, lenses, phase corrector plates, etc., but operating on principles of diffraction instead of refraction. The hologram typically does not replace a conventional element in a major application, but instead fills a niche where conventional techniques are inadequate. Since the diffraction process on which holography is based is highly dispersive, the holographic optical element typically is used in situations where the illumination is monochromatic.

A hologram of a point source is a Fresnel zone plate lens. A bleached hologram can achieve diffraction efficiency in the 90% region. Thus, the hologram can be used as a lens. Since the Seidel aberrations are severe, the holographic lens typically operates with a narrow field, so that only spherical aberration is a problem, and the spherical aberration can be corrected. For example, if a hologram is made by the interference of a diverging spherical wave with a collimated reference beam, then illumination of the hologram with a duplicate of the spherical yields an aberration-free collimated beam. The hologram thus can be a high-quality collimator. Alternatively, it could be used to focus a collimated beam to a high-quality point image.

Morris and George[18] have developed hybrid refractive-diffractive optical systems for forming an achromatic Fourier transform. They have been able to carry out spatial matched filtering in broadband light. The problem is that

when a Fourier transform is produced with conventional lenses, it has a scale factor proportional to λ, and this wavelength effect is extremely difficult to eliminate using refractive elements alone. However, a composite lens system using refractive and diffractive elements in cascade can reduce this chromatic aberration by several orders of magnitude.

The HOE has the powerful advantage that it can be aspherical. Conventional refractive elements can be made with flat, spherical, or cylindrical surfaces; other surfaces are difficult to produce. A hologram can be made of any wave front that can be generated. The HOE will, on readout, generate such a wave front. If the wave front for making the HOE cannot be generated, there is the possibility of generating the required hologram by computer. In principle, any wave front can be generated this way, leading to some quite sophisticated optical elements. A hologram thus can be made that corresponds to a refractive element of almost any arbitrary surface profile.

So-called holographic diffraction gratings have become commercially important, being quite competitive with conventional ruled gratings. It is, however, questionable whether a recorded interference pattern of two simple plane waves should qualify as a hologram. Nevertheless, much of the technology developed in holography applies to the construction of the grating.

Binary optics is a highly successful form of holographic optical element, in which binary (or sometimes multiphase) relief structures are deposited on the surface of conventional optical elements using ion etching or other techniques developed for integrated circuit manufacture.[19] The result is a single structure that is both refractive and diffractive, and the diffractive part is typically of the form of a phase corrector plate that reduces the aberrations of the refractive part. A variety of applications are being developed, particulary in the IR region of the spectrum.

6.9   DISPLAY HOLOGRAPHY

In recent years, display holography has emerged as the most important application of holography. The rise of display holography as a significant commercial activity was dependent on two factors: white-light viewability and mass-production technology. There are basically two kinds of white-light viewable holograms: the Denisyuk[6] and the Benton.[20]

6.9.1   Denisyuk Hologram

The construction of the Denisyuk hologram is shown in Fig. 6.16. A laser beam passes through a photographic plate, thereby exposing it. The transmitted light is reflected from a nearby object and exposes the plate a second time. A hologram is thus formed, with the light serving as a reference beam on the first pass through the plate, and as the object beam on the second pass. The laser should have a coherence length exceeding the object-hologram distance. The resulting hologram is thus of the volume-reflection type, with the fringes essentially parallel to the surface of the plate. The hologram is white-light viewable.

It is advantageous to have the image as close as possible to the hologram, since under white-light illumination resolution deteriorates with increasing hologram-image distance. The optimum arrangement is to have the image
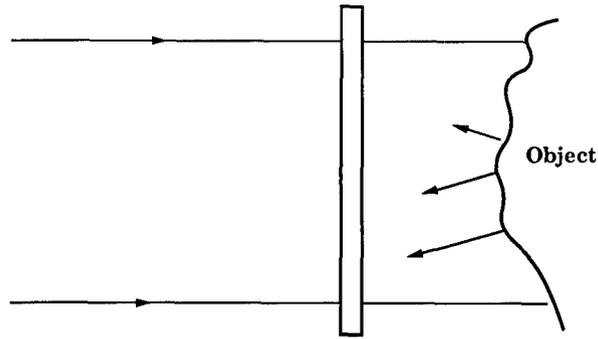
Fig. 6.16 Construction of a Denisyuk hologram.

straddle the plate, with a portion in front of the plate and a portion behind; thus, on average, the image-hologram distance is minimized.

To achieve this result, a two-step process is commonly used, with the first hologram being of the laser-transmission off-axis type. This hologram is read out using a readout beam conjugate to the reference beam, i.e., a beam exactly retracing the reference beam path, but in the opposite direction. The result is a real image, unaberrated and undistorted. This image is used as the object for a second hologram, made in the Denisyuk manner, with the recording plate straddling the image. The result is a holographic image that lies partly in front of the plate and partly behind it. If the depth is restricted to about a centimeter from the hologram surface, the readout process is very forgiving and the source can be fairly broad, e.g., a standard frosted light bulb at a distance of about 3 m, and the image will still appear sharp to the eye.

The resolution degradation due to the source size and to the spectral width of the diffracted light is readily seen by recognizing that each point of the object forms a zone-plate component of the hologram (here, of course, a reflection zone plate). Upon readout, the zone plate, being a lens, forms an image of the illuminating source; this source image is at the same time the image point corresponding to the object point that produced the zone plate. Thus, the hologram image can be regarded as multiple images of the source that produced the readout beam. This overall process (Fig. 6.17) shows how the readout beam is focused to form the reconstructed image. The size of the point-spread function is, by this viewpoint, just the size of the illuminating source, multiplied by the magnification the zone plate produces in imaging the source. This magnification is just the ratio of the image-forming distance to the distance of the illuminating source from the hologram. For image points close to the hologram, the demagnification is considerable, the image point-spread function is small, and the image will appear sharp to the eye. Image points forming further away are larger and hence appear blurred.

The effect of the spectral width $\Delta\lambda$ of the reconstruction beam is determined by similar considerations. The zone plate has a focal length inversely proportional to λ, so that different wavelength components form images at different distances, leading to a large longitudinal chromatic aberration, with each object point being focused into a line segment with length proportional to the
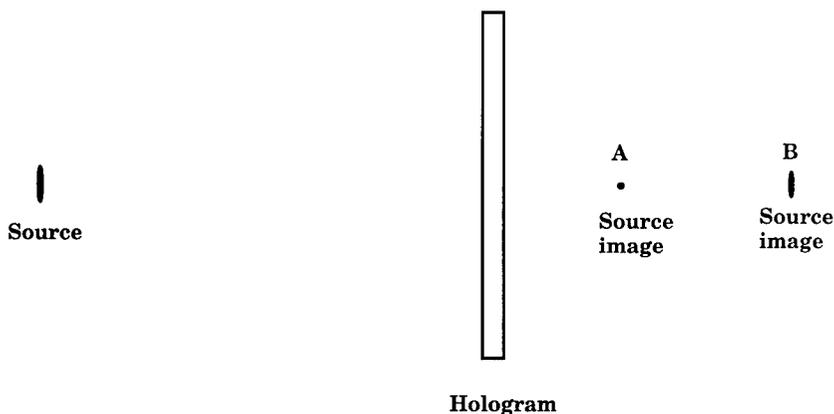
Hologram

Fig. 6.17   How source size affects resolution in a hologram. Each zone plate component of the hologram forms an image of the source. Image point A is formed from a zone plate of short focal length, produced by an object point close to the hologram. Image point A is thus a highly demagnified image of the source and is thus sharp even though the light source is rather large. Image point B, produced by a zone plate of longer focal length, is a less demagnified image of the source and is thus less sharp.

spectral width $\Delta\lambda$. The length is also proportional to the mean hologram-image distance, so that image points further from the hologram undergo greater longitudinal dispersion.

6.9.2   Benton Hologram

A second, quite different, kind of white-light viewable hologram was described by Benton in 1969.[20] In the Benton hologram, emulsion thickness with the attendant wavelength selectivity is not a factor. Instead, the white-light viewing capability is a consequence of the sacrifice of vertical parallax in the image. The Benton technique is a two-step process in which the first step is to make a conventional, off-axis transmission hologram, called the H_1 hologram. This hologram is then illuminated with a conjugate beam to form a real image, which is then used as the object for the second hologram H_2. The two-step process is shown in Fig. 6.18. In the second step, a horizontal slit is placed in front of the first hologram. The effect of the slit is to eliminate all vertical parallax. The second hologram can be formed by placing the recording plate in the vicinity of the real image, either in front of it, behind it, or straddling it. The reference beam is brought in from above or below, thus creating a fringe pattern that diffracts light in the vertical plane.

Illuminating the H_2 hologram with a wavelength λ_1 results in a sharp image, as well as an image of the slit [Fig. 6.18(c)] forming at a distance approximately that of the H_1–H_2 separation. When the observer places his eyes in the slit image, he sees the usual holographic image, except that he has only one vertical view. If a two-wavelength beam illuminates the hologram, each wavelength component forms a slit image, with vertical separation, due to the grating-like character of the hologram. The observer can place his eye at either slit image and see the holographic image, but in a different color. Finally, if the
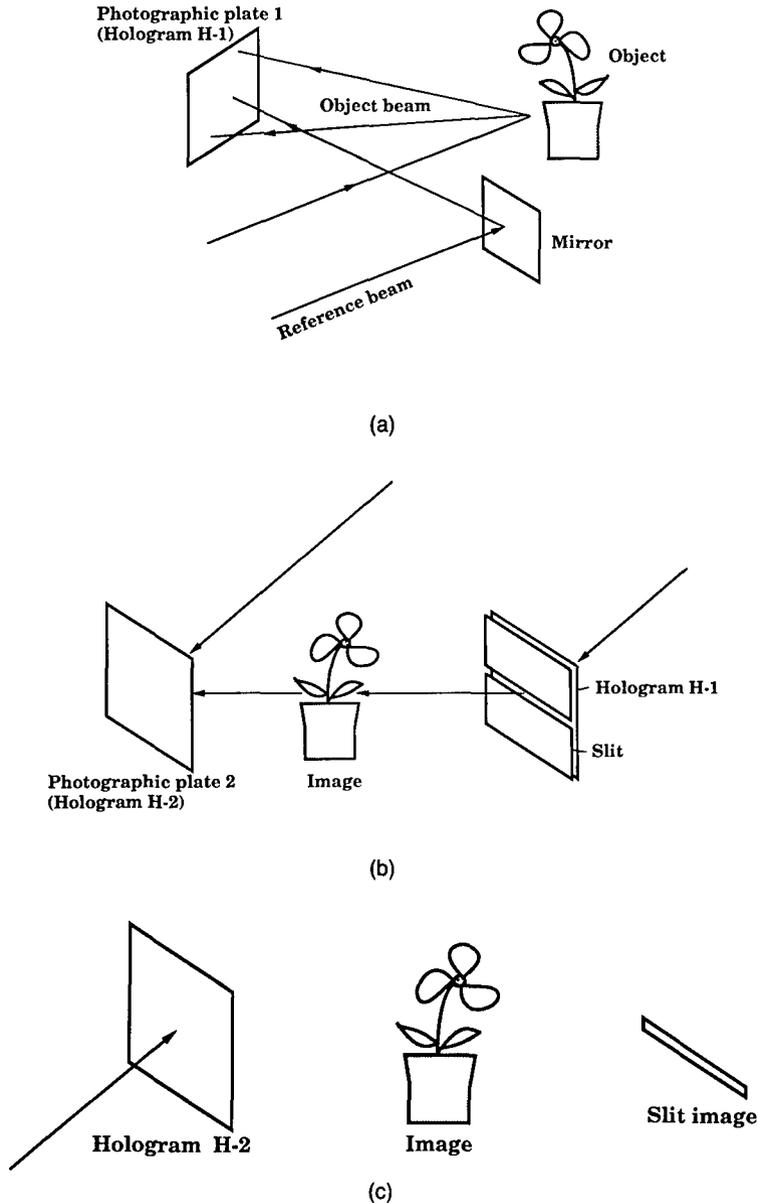
Fig. 6.18   The steps of making and viewing a rainbow hologram: (a) making the first hologram, (b) making the second hologram from the first, and (c) viewing the second hologram.

source is white, the slit image becomes a rainbow-like spectral distribution, and wherever in the spectrum the observer places his eye, he will see the holographic image in sharp focus, in a color that depends on his viewing position.

Thus, the hologram is white-light viewable, with the color of the image varying as the observer changes his viewing position vertically. In short, white-light viewing occurs because only one color of light is diffracted in any one direction. It was the limiting slit at H_1 that produced this situation, and it

also eliminates vertical parallax. Hence, the process is a trade-off of vertical resolution for white-light viewability.

As with the Denisyuk hologram, the white-light viewing capability of the Benton hologram entails a loss of resolution with increasing image-hologram distance and hence the image depth is usually kept small. An optimum slit width exists, and the resolution will be degraded if the slit is either too wide or too narrow.

6.9.3   Design Considerations for the Benton Hologram

In designing the Benton hologram, the principal parameters are separation of the H_1 and H_2 holograms, the angle of the reference beam in the H_2 hologram, and the width of the slit covering the H_1 hologram. The H_1–H_2 distance should be the intended viewing distance of the H_2 hologram, since the slit image should be at the viewing position, because in this slit image plane the holographic image will be uniform in color from top to bottom. The reference beam angle is typically in the range of 45 to 60 deg; small angles give less angular separation of the color components, leading to increased blur and a smaller vertical viewing range. The slit width is the most complicated consideration; too narrow a slit leads to low resolution and excessive speckles in the image, while too wide a slit also leads to a blurred image.

The basic principle involved in determining the optimum slit size is that the Benton hologram is really not a hologram at all in the vertical dimension. In the vertical dimension, the image is only as sharp as the defocused image had been at the H_2 hologram recording plane when the hologram was made. The situation is akin to conventional photography, where the recorded image is just what appears at the recording surface. For a 3-D object, the portion of the image that coincides with the recording plane will be very sharp, whereas parts of the image that form away from the recording plane will be blurred by being out of focus at the recording plane. The photographer optimizes the recorded image by adjusting the $f/\#$ of the imaging lens; stopping down the lens gives increased depth of field and therefore sharper resolution at image positions away from the plane of focus, but also at the expense of loss of resolution in the plane of the recording film.

This situation applies also to the hologram, with respect to the vertical dimenson. The problem of optimizing the slit width is complicated by the fact that in producing the commercial Benton hologram, the object is designed so that the details that should be rendered most sharply lie at the plane that will be focused onto the H_2 hologram; such a design procedure allows a larger slit than would otherwise be allowed.

6.9.4   The Composite Hologram

A third class of commercial display hologram is one that is synthesized from a set of many individual photographs of an object scene, each picture taken from a slightly different position. Such holograms are broadly called composite holograms, subtypes of which include integral and multiplex holograms.

A typical arrangement for the construction of an integral hologram is shown in Fig. 6.19. Figure 6.19(a) shows an object photographed from a large number of positions, often in the hundreds. The recording camera can be moved lat-
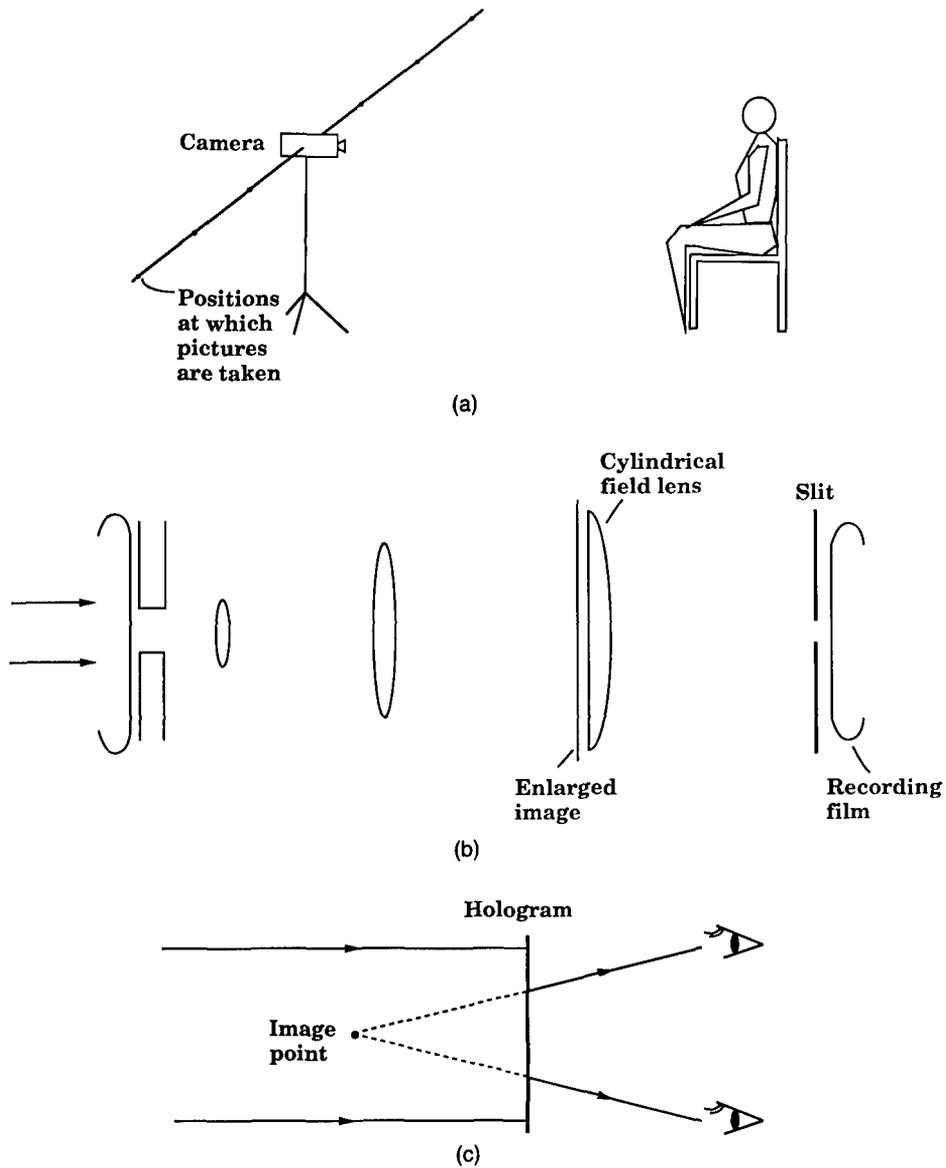
Fig. 6.19 Construction and viewing of a composite hologram: (a) taking the pictures, (b) forming the hologram, and (c) viewing the hologram.

erally, occupying all the indicated positions in Fig. 6.19(a), with a picture taken at each position. These many pictures are then synthesized into a hologram, as in Fig. 6.19(b). A picture is projected, using laser light, to a size desired in the hologram image, and a hologram is made using this image as an object. The hologram is recorded along a narrow vertical slit, and to direct light from each image point onto the slit a cylindrical lens may be overlaid on the image, as shown, or a plate of ground glass can be placed at the image plane. The former is more efficient, since a much larger proportion of the image-

forming light is directed onto the slit. The reference beam is introduced from above or below. The recording film is advanced one slit width, and the next picture is converted into a similar hologram. Each frame of the picture sequence thus is converted into a long thin hologram, about a millimeter in width. The result is a hologram that is a composite of many holograms, one for each picture frame. The hologram is viewed as in Fig. 6.19(c), which shows a virtual image produced by each individual hologram. The eyes see each image point through different holograms, and since any two holograms constitute a stereo pair, the observer sees an image in 3D. Lateral movement of the observer results in the image point being seen through a different pair of holograms, hence parallax between object points at different depths is obtained, although in small discrete steps rather than continuously. The hologram thus behaves approximately like a conventional hologram.

Since the hologram has holographic properties only along the x dimension, with no multiple views in the y dimension, it is white-light viewable for exactly the same reason that the Benton hologram is white-light viewable.

The hologram is typically used as the H_1 master to form a second hologram. A real image is formed by the H_1 hologram, from which an H_2 hologram is formed by placing the recording plate so as to straddle the image. The H_2 hologram will be sharper under illumination with a broad spectrum source, since on average the image is closer to the hologram than for the H_1 hologram.

The multiplex hologram is a variant of this process, in which the camera is held stationary and the object is on a rotating platform. The resulting hologram is then wrapped into a cylinder and a light source is placed within the cylinder, either above or below the hologram. The cylindrical shape is needed, since the rotation of the object is equivalent to the camera moving in an arc about a stationary object; the hologram curvature then corresponds to the curve through which the recording camera had equivalently moved.

The composite hologram (integral or multiplex) yields an image that is somewhat less sharp than the image of other holograms, since it is made from a finite number of discrete images. However, the technique enables a hologram to be made from any object, no matter how large, and the hologram can form an image of whatever size is desired, since the first step is conventional photography, and the overall process is thus freed from the unity magnification constraint of conventional holography.

6.10   SUMMARY

Holography today is a well-developed technology, with many diverse applications and with continuing active research. In addition to its well-recognized application areas, such as for display and for nondestructive testing, it is being applied in numerous specialized situations. For example, holograms have been proposed for interconnect applications in the research area of digital optical computing and have been used as gating devices in short-pulse laser technology. The range of applications for holography likely will continue to increase.

References

1. D. Gabor, "A new microscope principle," *Nature* **161,** 777–778 (1948).

2. D. Gabor, "Microscopy by reconstructed wavefronts," *Proceedings Royal Society* **A197,** 454–487 (1949).

3. D. Gabor, "Microscopy by reconstructed wavefronts II," *Proceedings Physical Society* **B64,** 449–469 (1951).

4. E. Leith and J. Upatnieks, "Reconstructed wavefronts and communication theory," *Journal Optical Society of America* **52,** 1123–1130 (1962).

5. G. L. Rogers, "Gabor diffraction microscopy: the hologram as a generalized zone plate," *Nature* **166,** 237 (1950)

6. Y. Denisyuk, "Photographic reconstruction of the optical properties of an object in its own scattered radiation field," *Soviet Physics-Doklady* **7,** 543–545 (1962).

7. P. J. van Heerden, "A new optical method of storing and retrieving information," *Applied Optics* **2,** 387–392 (1963).

8. D. Angell, "Incoherent spatial filtering with the grating interferometer," Ph.D. Thesis, p. 94, University of Michigan (1983). (Available from University Microfilms, Ann Arbor, MI.)

9. H. Kogelnik, "Coupled wave theory for thick holographic gratings," *Bell System Technical Journal* **48,** 2909–2923 (1969).

10. E. Leith and J. Upatnieks, "Holograms, their properties and uses," *SPIE Journal* **4,** 3–6 (1965).

11. H. Kogelnik, "Holographic image projection through inhomogeneous media," *Bell System Technical Journal,* 2451–2455 (1965).

12. R. Fisher, Ed., *Optical Phase Conjugation,* Academic Press, New York (1983).

13. J. Upatnieks, A. VanderLugt, and E. Leith, "Correction of lens aberrations by means of holograms," *Applied Optics* **4,** 589–593 (1966).

14. M. Horman, "An application of wavefront reconstruction to interferometry," *Applied Optics* **4,** 333–336 (1965).

15. K. Stetson and R. Powell, "Interferometric vibration analysis by wavefront reconstruction," *Journal Optical Society of America* **55,** 1593–1597 (1965).

16. A. Kozma and D. Kelly, "Spatial filtering for detection of signals submerged in noise," *Journal Optical Society of America* **54,** 1395 (1964).

17. A. VanderLugt, "Signal detection by complex spatial filtering," *IEEE Transactions on Information Theory* **IT-10,** 139–145 (1964).

18. M. Morris and N. George, "Frequency plane filtering with an achromatic optical transform," *Optics Letters* **5,** 446–448 (1980).

19. J. Veldkamp and G. J. Swanson, "Developments in fabrication of binary optical elements," in *Proceedings of the SPIE* **437,** 54–59 (1980).

20. S. Benton, "Hologram reconstruction with extended incoherent sources," *Journal Optical Society of America* **59,** 1545 (1969).

System Design Considerations for a Visually Coupled System

Brian H. Tsou
Armstrong Laboratory
Wright Patterson Air Force Base, Ohio

CONTENTS

7.1   INTRODUCTION

What is a visually coupled system (VCS)? At the VCS Development and Application Symposium held in 1972 at Brooks Air Force Base (AFB) in Texas, it was defined as ". . . a special subsystem which integrates the natural visual and motor skills of an operator into the system he is controlling. The VCS operator visually searches for, finds, and tracks an object of interest. His line of sight is measured and used to aim sensors and/or weapons toward the object. Information related to his visual/motor task from the sensors, weapons, or central data sources is fed back to him through special displays so as to enhance his task performance."[1] Figure 7.1 illustrates a use for such a system. The pilot looks at the target, the turreted forward-looking infrared (FLIR) automatically follows his line of sight (LOS) and points at the intended targets. With the magnified view presented in his display, he positively identifies the target and simultaneously slews the weapon, using his LOS again, to acquire the target. He then "pickles"[2] the ordnance and egresses.

A decade later, the U.S. Air Force completed its development of a Visually Coupled Airborne Simulation System (VCASS)[3] at Wright Patterson AFB, Ohio. Also, a Fiber Optic Helmet-Mounted Display (FOHMD)[4] was developed at Williams AFB, Arizona, and the U.S. Army built and fielded the Integrated Helmet and Display Sighting System (IHADSS) for the AH-64 Apache helicopters. During the 1980s, a great deal of R&D effort within the military and industry was devoted to increasing subsystem performance and, at the same time, to reducing the size, weight, and cost. Today, a high level of interest
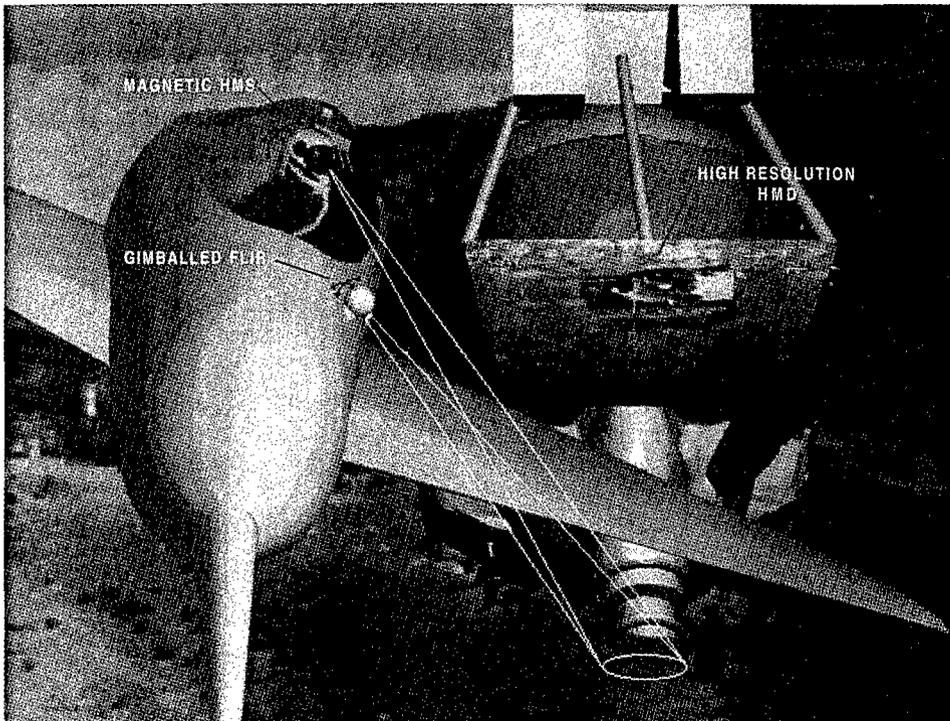
Fig. 7.1   Visually coupled system and subsystems.

exists for fielding VCSs in programs such as the Army's RAH-66 Comanche, the Navy's V-22 Osprey, and the Air Force's F/A-16 Fighting Falcon.

7.2   SYSTEM OVERVIEW

This chapter provides an introduction to some of the design parameters of a VCS and shows how these can be quickly estimated from simple formulas. Sample calculations are provided where appropriate.

A VCS could be essential in performing interdiction missions. It can provide the pilot with FLIR displays to aid in low-level or terrain-following flights, in the detection of hot targets at long stand-off ranges, and in air-to-ground attacks at night or in adverse conditions. As depicted in Fig. 7.1, a VCS uses a helmet-mounted sight (HMS) to locate the head LOS and the FLIR is gimbal-mounted and uses the HMS output to follow the pilot's LOS. The helmet-mounted display (HMD) feeds the FLIR imagery back to the pilot.

A brief overview follows of the three major components involved in VCS. Detailed engineering descriptions for each subsystem are beyond the scope of this chapter; only the system design parameters are described.

Forward-Looking Infrared.   A FLIR sees objects by detecting their thermal (IR) radiation. Therefore, the FLIR's effective detection range depends mainly on how "hot" the target is compared to the surrounding background and on how much of that "heat" gets through the atmosphere to the FLIR. Under weather conditions such as fog, rain, or snow, the FLIR can be degraded or, essentially, "blind." However, the longer wavelengths of IR radiation may enable a FLIR to see through dust, smoke, or other deceptive countermeasures often deployed on the battlefield. A turret-mounted FLIR that follows the pilot's LOS can be used to aid the pilot's situation awareness during terrain-following flights. Also, the zoom capability in a FLIR's optics can improve the pilot's targeting capability and enhance mission success.

Helmet-Mounted Display.   A HMD is a display that mounts on the pilot's helmet and projects, at infinity, a virtual image to the pilot. A miniature cathode-ray tube (CRT), about 1 in. in diameter, has been commonly used as the image source for the HMD to display the FLIR imagery and symbology. Future mission scenarios may demand even higher resolution, brightness, and ruggedness requirements than a CRT is able to deliver. Other possible image sources, such as small-sized and lightweight liquid crystal displays are promising but have other design constraints such as addressability and image quality.

Helmet-Mounted Sight.   Current and proposed HMS systems use different techniques to determine the LOS. The HMS fielded in the Apache helicopter uses IR beams and an array of detectors to measure the head position. For other applications, the magnetic HMS is used because of its larger motion box, within which accurate head orientation and position can be rapidly determined. A disadvantage of the magnetic HMS is the extensive mapping that is required to compensate for the electromagnetic scattering generated by metal structures commonly found inside a cockpit.

7.3   SUBSYSTEM DESCRIPTIONS

When target identification is the primary concern, as depicted in Fig. 7.1 for interdiction missions, the system resolution requirement is necessarily high. Once the system resolution is defined, it is used to estimate the needed performance level of the FLIR and HMD (Fig. 7.2). As is shown later, this high system resolution demands a dual field-of-view (FOV) FLIR design that consists of a wide FOV and a highly magnified narrow FOV. Because a one-to-one relationship between the real and imaged world is to be maintained for pilotage, the HMD FOV must match the FLIR's wide FOV. Furthermore, to support targeting, the HMD's resolution must support the high-resolution narrow FOV FLIR and the HMD contrast must provide the necessary gray levels required for target recognition.

Despite major improvements in CRT performance, HMD resolution is still limited by the resolution of the CRT. For the purpose of designing a high-resolution HMD, we must determine the maximum CRT resolution performance. Next, we need to know the maximum available CRT luminance at that resolution (this is tube dependent and can be found in the manufacturer's data sheet). The CRT luminance and the required display contrast define the minimum display optics transmission efficiency that allows us to choose a suitable
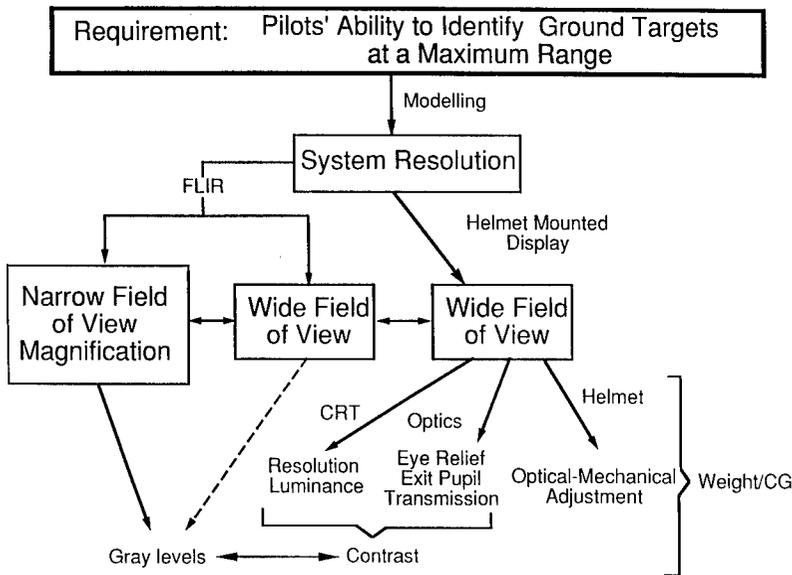
Fig. 7.2   VCS design parameters and their interrelationships.

optics design. The exit pupil and eye relief requirements can now be calculated based on physiology and any particularity of the chosen display optics.

The optics design may also dictate different implementations for the opto-mechanical adjustments to compensate for the head feature variabilities that exist among the aviator population. Finally, the total weight borne on the head and the helmet's center of gravity (CG) may affect or even limit the usability of such a display in an aircraft because of the force and inertia exerted on the neck during high g turns or bail out. Ultimately, the "helmet" must be engineered as a stable and reliable "fire-control" platform without causing too much discomfort to the pilot. No doubt, the helmet's fit and comfort may make or break the pilot's acceptance of such a display. A good helmet-fitting procedure must be devised.

Aspects of these relationships and their effects on either FLIR or HMD design are discussed further in the following sections. The last part of this section briefly discusses the HMS.

7.3.1   Resolution

To calculate the required resolutions for the FLIR and HMD subsystems, we must first determine the minimum overall system resolution. The minimum system resolution is set by the maximum allowable target recognition range, which depends on many factors that are not always predictable (such as the atmospheric condition). Nevertheless, this range can be approximated by means of an operational analysis technique that models the mission requirement, aircraft maneuverability, weapon effectiveness, etc. An IR recognition model is then used to find the total VCS resolution that would be required to achieve such a recognition range. For our purpose, we can estimate the system resolution by first calculating the resolvable frequency and then using the rule of thumb that the resolvable frequency is 65 to 75% of the system's cutoff frequency (or system resolution). For example, resolvable frequency, according to Johnson's criteria for recognition,[5] is proportional to the number of cycles (i.e., TV line pairs or pixel pairs) across the target and inversely proportional to the target's angular subtense in milliradians. The angular subtense is a function of the recognition range. For simplicity, we assume that the recognition range is the same as the mission effective range determined solely on operational constraints. Therefore,

$$f_{\text{resolv}} = \frac{N}{w/R} \, , \tag{7.1}$$

where

f_{resolv} = resolvable frequency in cycles per milliradians
N = number of cycles or pixel pairs across the target
w = width of a target in meters
R = range in kilometers at which the pilot should recognize, at a specific probability, the target.

Therefore, using Eq. (7.1) and some reasonable estimates of the listed parameters, we can calculate the required resolvable frequency. If $N = 4$ cycles,

w = 4 m, and R = 4.5 km, then f_{resolv} equals 4.5 cycles/mrad. Using the previous rule of thumb, the required system resolution is at least 6 or 7 cycles/ mrad. For our example, we assume that the resolvable frequency lies at 70% of the cutoff frequency, which gives us a system resolution of 6.43 cycles/mrad. Note that our resolvable frequency has been calculated using target width and, therefore, corresponds to the horizontal system resolution. The most recent release of the C2NVEO Advanced FLIR Systems Performance Model[6] uses both the horizontal and vertical resolution to predict target recognition range. However, for our example, the horizontal resolution alone suffices for illustrative purposes.

Now that we have an estimate of the overall system resolution, we can use this to estimate the required subsystem parameters as outlined in Fig. 7.2.

7.3.2   Forward-Looking Infrared

The FLIR and HMD performance must be considered together as a system, and the design process should iterate until a proper balance between each subsystem performance is obtained. The design of a FLIR is a complex subject and is discussed in detail elsewhere in this Handbook.

7.3.2.1   Pilotage: Wide Field of View.

For pilotage, a wide FOV FLIR image that is in a 1:1 registration with the outside world is needed to avoid spatial disorientation. A typical pilotage or navigation FLIR for fixed-wing aircraft has a 30-deg horizontal FOV. Whether 30 deg is wide enough for pilotage is still being debated, however, a 30- to 35-deg Falcon Eye[7] was flown by pilots in a simulated night attack mission and was reported to be sufficient. This FOV does not need the very high resolution (6.43 cycles/mrad) used for targeting but still requires high image quality even for pilotage.[8] A typical 30-deg FOV FLIR may have a horizontal cutoff frequency around 0.8 cycle/mrad. If we assume that the size of the detector limits the FLIR's resolution, then, because the discrete detector's modulation transfer function (MTF) is a sinc function, the cutoff frequency ($f_{c\text{-FLIR}}$) is proportional to the reciprocal of the detector's instantaneous FOV (IFOV), or

$$f_{c\text{-FLIR}} = \frac{1}{\text{IFOV}} \cdot \tag{7.2}$$

Therefore, the 0.8 cycle/mrad system resolution required to support pilotage corresponds to an IFOV of 1.25 mrad.

Figure 7.3 shows a simplified FLIR[9] for our discussion, where w_e is the width of a detector element, h_t is the height of the detector array, D is the absolute lens aperture, f_{FLIR} is the lens focal length, IFOV is each detector's instantaneous FOV in the horizontal dimension, and VFOV is the total vertical FOV. It is clear from Fig. 7.3 that the width of a detector within the array w_e determines the IFOV:

$$\text{IFOV} = 2 \tan^{-1}\left(\frac{w_e/2}{f_{\text{FLIR}}}\right) \cdot \tag{7.3}$$
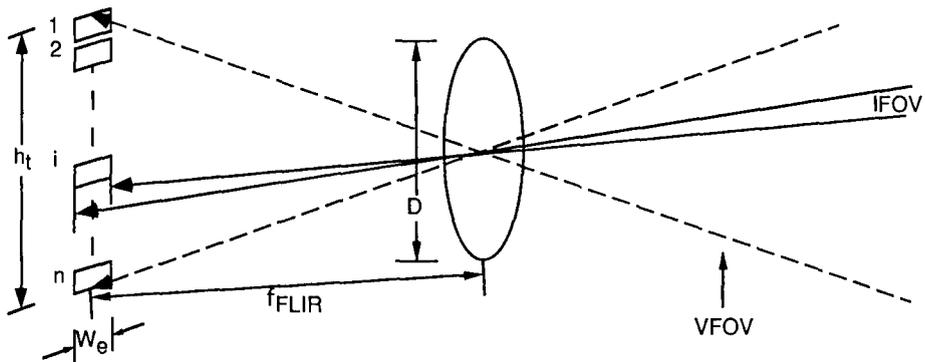
Fig. 7.3   FLIR optical geometry.

Assuming a video format of a 4:3 aspect ratio, for a 30-deg horizontal FOV sensor, the VFOV is 22.5 deg, which sets the minimum height of the detector array. Because

$$\text{VFOV} = 2 \tan^{-1}\left(\frac{h_t/2}{f_{\text{FLIR}}}\right) , \tag{7.4a}$$

then

$$h_t = 2[\tan(\text{VFOV}/2)\, f_{\text{FLIR}}] . \tag{7.4b}$$

Note that a trade-off exists between horizontal resolution and vertical FOV. As shown in Eqs. (7.3) and (7.4a), to improve the sensor's horizontal resolution (smaller IFOV) with a fixed detector size (w_e and h_t), we must increase f_{FLIR}. However, this results in a decreased VFOV of the sensor and, thus, the vertical FOV of the HMD if a 1:1 registration is to be maintained. Reducing the HMD vertical FOV may cause the pilot to increase his vertical head movements. In general, vertical head movements are more fatiguing than horizontal head movements, and it is possible a sensor with reduced vertical FOV may over time degrade pilot performance. A laboratory study[10] conducted to determine the optimum vertical FOV indicates that changing the FOV horizontal-to-vertical proportions made a difference only at the smallest FOV; as for larger (>20-deg) FOV, trade-offs between horizontal and vertical FOV did not seem to be as critical as achieving an optimum overall FOV area (FOV vertical size multiplied by horizontal size). However, general guidelines for choosing the right overall FOV and resolution combination for a specific mission are not available (see Refs. 8 and 11 for a review of this topic).

7.3.2.2   Targeting: Narrow Field of View and Magnification.   The very high system resolution of 6.43 cycles/mrad required to support the target recognition range corresponds to an IFOV of less than 0.16 mrad [Eq. (7.2)]. To achieve this IFOV for targeting with the same detector (w_e in Fig. 7.3) as for the pilotage, a lens with a much longer focal length (f_{FLIR}) is required [Eq. (7.3)].

The longer focal length produces a magnified image with a narrower FOV. In our example, we need $8\times$ magnification to increase the resolution from the 0.8 cycle/mrad (IFOV of 1.25 cycles/mrad) pilotage resolution to the 6.43 cycles/mrad (IFOV of 0.16 cycle/mrad) targeting resolution. Therefore, the targeting focal length needs to be approximately eight times longer (1.25/0.16).

In general, high magnification and narrow FOV are not advisable because of head LOS jitter and pilot disorientation. To prevent disorientation, pilots only take "snap" views of the highly magnified target and do not stay in the narrow FOV for long. A simulator study[12] has shown that a 7:1 change from wide (20-deg) to narrow (3-deg) FOV is not excessive. However, whether a head-stabilized display in conjunction with another manual augmentation, such as a thumb wheel to control the sensor LOS, can tolerate a magnification factor higher than $7\times$ is still under debate.

The absolute lens aperture (D in Fig. 7.3) is another important parameter, because it affects the size of the optical head and its gimballed mounting fixture. If the optical head is too large, it may degrade the aircraft's aerodynamic performance. Also, if the gimbal fixture has to be made too sturdy it may result in a slew rate too slow to follow the natural head movements. The optimum optical head size is, therefore, an aircraft-dependent issue that cannot be fully addressed here. However, the minimum acceptable D required to support the FLIR's targeting resolution or $f_{c\text{-FLIR}}$ of 6.43 cycles/mrad can be estimated using practical guidelines that assume the optics are diffraction limited and have 40% modulation at $f_{c\text{-FLIR}}$, where the detector's MTF falls to zero, i.e., the cutoff frequency for the FLIR optics is 6.43 divided by 0.6 or 10.72 cycles/mrad. Using the diffraction-limited system performance relationship[13]:

$$f_{c\text{-FLIR(optics)}} = \frac{1}{(f_{\text{FLIR}}/D)\lambda} \, , \tag{7.5a}$$

where λ is wavelength in millimeters and $f_{c\text{-FLIR(optics)}}$ is the cutoff frequency of the FLIR optics in cycles per millimeter. Solving Eq. (7.5a) for D and converting the frequency from cycles per millimeter to cycles per milliradians yields:

$$D = 1000 f_{c\text{-FLIR(optics)}}\lambda \, . \tag{7.5b}$$

Assuming that $\lambda = 10$ μm $= 0.01$ mm and $f_{c\text{-FLIR(optics)}} = 10.72$ cycles/mrad, $D = 107.2$ mm or the lens aperture should be slightly larger than 4 in.

7.3.3   Helmet-Mounted Display

Because a one-to-one relationship between the real and imaged world is to be maintained for pilotage, the HMD FOV must match the FLIR's wide FOV. At the same time, the HMD's resolution must support the targeting FLIR's high resolution.

7.3.3.1   Field of View/Resolution.   An HMD with a 30-deg FOV and a resolution of 0.8 cycle/mrad is sufficient to support the required 6.43 cycles/mrad FLIR's narrow FOV resolution. The reason for this is that the $8\times$ magnification of the FLIR limits the narrow FOV to about 3.75 deg. Therefore, the 30-deg

HMD optics will act like an $8\times$ magnifier and improve the HMD angular resolution to an equivalent 6.43 cycles/mrad (0.8 cycle/mrad $\times$ 8). Thus, the FLIR narrow FOV imagery that is superimposed on the outside scene through the HMD combiner as seen by the pilot is a magnified view of the target and is not in 1:1 registration with the outside world.

7.3.3.2 Gray Levels/Contrast.   The contrast ratio or contrast of the HMD is defined as

$$C_{\text{HMD}} = \frac{B_{\text{HMD}} + B_{\text{bkgd}}}{B_{\text{bkgd}}} \; , \tag{7.6a}$$

where, as shown in Fig. 7.4, B_{bkgd} is the outside scene luminance attenuated by the combiner and B_{HMD} is the HMD image source luminance (B_{CRT}) times the HMD optical transmission (t_{optics}). Equation (7.6a) can be rewritten as

$$C_{\text{HMD}} = \frac{B_{\text{CRT}} \times t_{\text{optics}} + B_{\text{bkgd}}}{B_{\text{bkgd}}}$$

or

$$C_{\text{HMD}} = \frac{B_{\text{CRT}} \times t_{\text{optics}}}{B_{\text{bkgd}}} + 1 \; . \tag{7.6b}$$
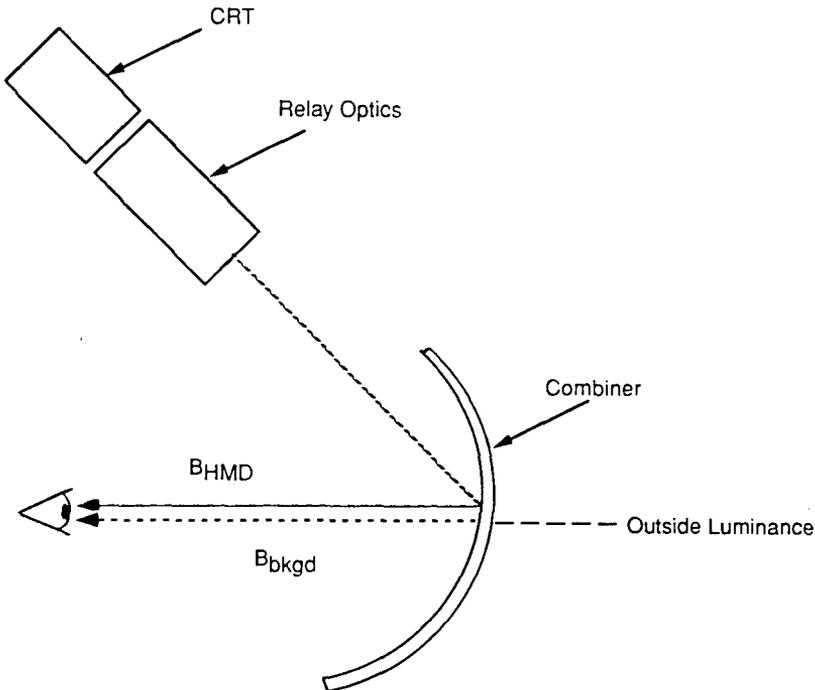
Fig. 7.4   HMD contrast and its components.

Enough contrast is required to provide adequate gray levels to support recognition. In general, eight gray levels are adequate for a simple search task with high operator expectation, but recognition tasks require 16 to 32 gray levels.[14] Therefore, for maximum performance, we assume that C_{HMD} equals 32.

For design considerations, we can assume the worst case in which B_{bkgd} is the same as the outside luminance (combiner see-through transmission is 100%). Typically, the sky illumination at sunset under clear weather is 30 fc. Assuming 100% reflectance, also for the worst case, from terrain features, the ambient luminance or B_{bkgd} would be 30 fL. If $C_{HMD} = 32$ and $B_{bkgd} = 30$ fL, then, from Eq. (7.6b), we can see that $B_{HMD} = 930$ fL.

Since it is well known that our ability to resolve details (visual acuity) is a function of light levels,[15] the system designer should, as a good practice, verify that the display resolution is not above the pilot's acuity. In our example, the eye is adapted to 100 fL (about 10% of the maximum display brightness of 960 fL) and has an acuity of 20/20. Since, by definition,

$$20/20 = 30 \text{ cycles/deg} = 1.72 \text{ cycles/mrad} ,$$

the eye is much better than the 0.8 cycle/mrad display resolution, and the system is usable.

7.3.3.3 Miniature Cathode-Ray Tube.

Despite the CRT's limited resolution, low brightness, high-voltage potential difference (typically in the 10-kV range), weight, and size, it is still the image source of choice for an HMD. This is due to its unique ability to address an infinite number of lines or pixels simply by applying the appropriate positioning signals to the X-Y deflection elements using only four connections. Significant improvements have been made in miniature CRT technology regarding its optical and electron beam performance.[16] However, in HMD applications, the CRT is still the resolution-limiting factor. From the preceding sections, we have established that a display resolution of at least 0.8 cycle/mrad is required to match the FLIR resolution and support target recognition. Since the display resolution is limited by the CRT, the CRT's resolution must be as good as 0.8 cycle/mrad, which is related to the spot size. The CRT spot size (linewidth) is limited by the electron beam diameter and the spreading of the light generated in the phosphor screen. On the other hand, the light output depends on the electron density in the beam at any given accelerating voltage of the electrons. Figure 7.5 shows the trade-off between linewidth and brightness for two popular miniature CRTs manufactured by Hughes Aircraft Company.[16] We use this plot to estimate CRT light output at the desired resolution of 0.8 cycle/mrad.

Luminance. The CRT beam spot luminance profile can be approximated by a Gaussian distribution. Therefore, the linewidth of the CRT can be calculated using the Gaussian point spread function (PSF) and its Fourier transform, the MTF. The MTF measures the CRT performance as a function of spatial frequency and can be expressed mathematically by:

$$g(f) = \exp[-2(\pi a f)^2] , \tag{7.7a}$$
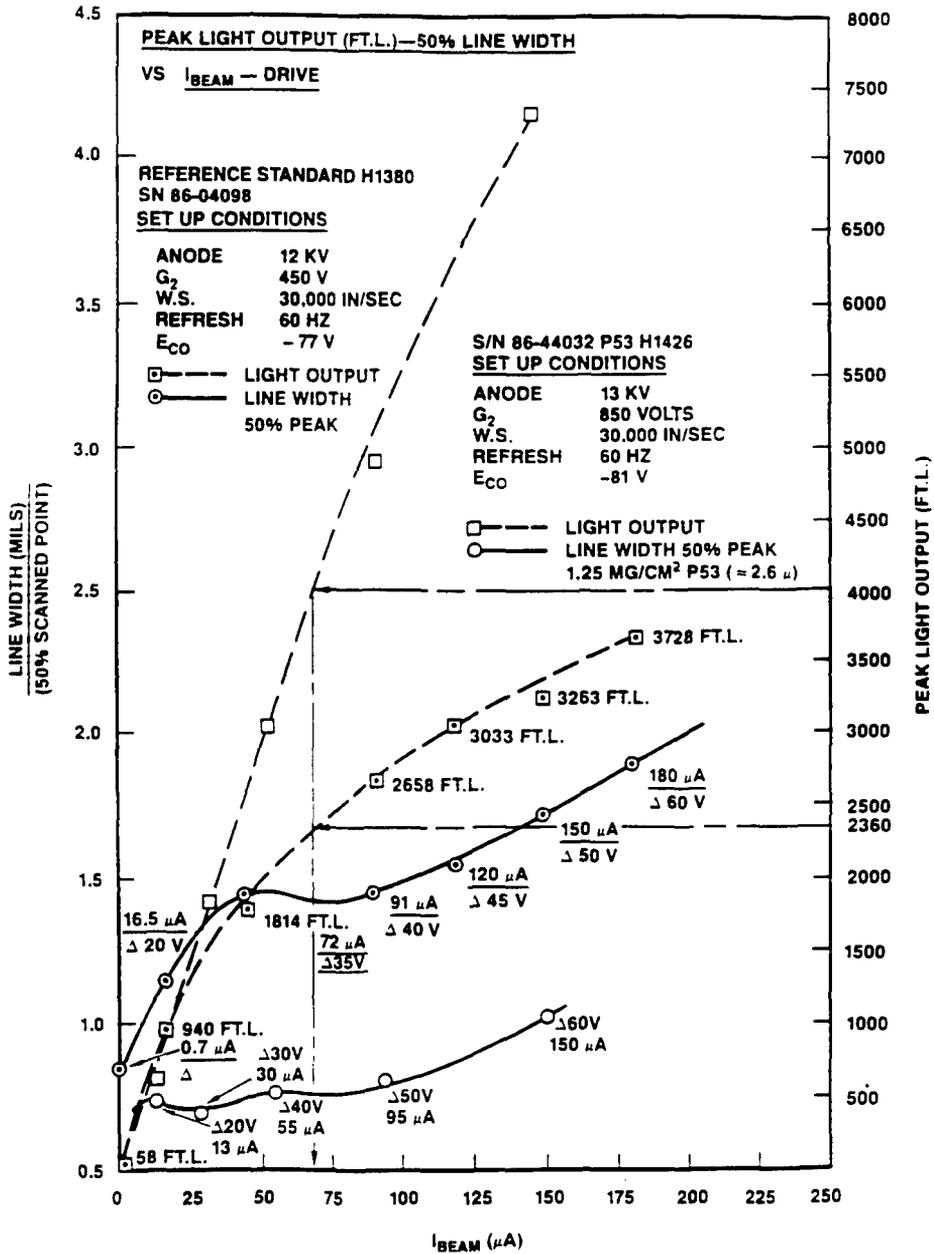
Fig. 7.5   Trade-off between linewidth and brightness for two miniature CRTs.[16]

where f is the spatial frequency in cycles per milliradians (Fig. 7.6) and a is the standard deviation. By solving Eq. (7.7a) for a we can then determine the bandwidth needed for the required resolution:

$$a = 0.34154 \left(\frac{\{-\log_{10}[g(f)]\}^{-1/2}}{f} \right) . \qquad (7.7b)$$

If we assume a 10% modulation at the CRT's limiting resolution, $f_{c\text{-HMD}} = 0.8$ and $g(f_{c\text{-HMD}}) = 0.1$, then from Eq. (7.7b), $a = 0.43$ mrad.

Equation (7.8a) represents the CRT PSF (Fig. 7.7), which is an inverse Fourier transform of Eq. (7.7a):

$$f(x) = \frac{1}{2\pi a^2} \exp\left(-\frac{x^2}{2a^2} \right) , \qquad (7.8a)$$

where x is the angular distance in milliradians from the center of the CRT beam spot. Normalizing $f(x)$ to a peak luminance of 1.0 at $x = 0$ gives

$$f_n(x) = \exp\left(-\frac{x^2}{2a^2} \right) . \qquad (7.8b)$$
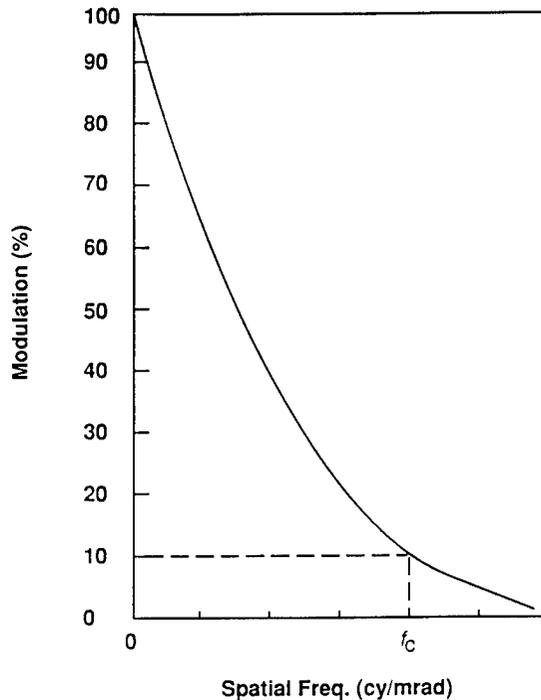
Fig. 7.6   CRT MTF.
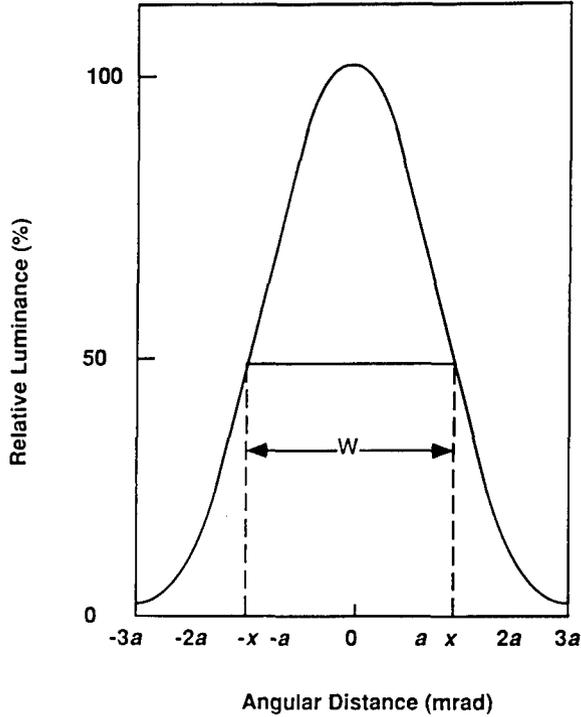
Fig. 7.7 CRT Gaussian PSF.

The 50% linewidth (W) is found by setting $f_n(x) = 0.5$ and solving for x:

$$W = 2x = 2.35482a .\tag{7.8c}$$

Substituting $a = 0.43$ mrad into Eq. (7.8c), we obtain $W = 1.01$ mrad.

To determine the peak luminance using Fig. 7.5, we must first convert the 50% linewidth from milliradians to mils:

$$W_{\text{in mils}} = 0.04 \times W_{\text{in mrad}} \times f_{\text{HMD}} ,$$

where f_{HMD} is the HMD optics focal length in millimeters. For a display that uses distortion-free f-tangent theta mapping:

$$f_{\text{HMD}} = \frac{F/2}{\tan\theta} ,\tag{7.9}$$

where F is the CRT format size in millimeters and θ is half of the FOV. For a 19-mm CRT format and a 30-deg FOV,

$$f_{\text{HMD}} = 35.45 \text{ mm} .$$

Then,

$W = 1.43$ mils .

From Fig. 7.5, for a 50% linewidth of 1.43 mil, the H1380 (19-mm format IHADSS tube) outputs a peak luminance of more than 1800 fL. Since $B_{CRT} \times t_{optics}$ has to be at least 930 fL in order to obtain an acceptable contrast, we can see that an optical transmission of at least 50% is required (if an H1380 tube is used).

CRT format size is another design trade-off consideration. The larger CRT can be made brighter; as a reference, H1426 (23-mm format) reaches 7350 fL for a 1-mil linewidth. The larger format also relaxes the optics design requirement (e.g., larger $f/\#$). However, the larger format CRT is heavier than the smaller CRT (a 23-mm format weighs ~ 6.5 oz versus the 19-mm format, which weighs ~ 5.0 oz).

7.3.3.4 Optics.

Many HMD optical designs exist.[17-22] Almost all of them take the form of an erecting eyepiece. The erecting eyepiece consists of an eyepiece and a relay system with a field lens imaging the aperture of the relay to form the exit pupil at the eye.[23] Different mission scenarios may require different optical designs for optimization.[24]

Transmission.   We need to pick a design that supports an optical transmission (t_{optics}) of 50% or more. Based on the Air Force/Army Virtual Panoramic Display[16] experience, both the Hughes Holographic[19] and the Farrand Off-aperture[23,25] HMDs are capable of supporting this transmission requirement (predicted CRT to eye transmissions are 85 and 90%, respectively). Both use an off-axis visor-projected approach to achieve high transmission; therefore, the optical distortion is more severe than that of an on-axis design. However, most of the distortion can be corrected by properly designed CRT drive electronics.[26] Moreover, both provide high see-through transmission from the outside scene and the viewing sensation is excellent—there is no obstruction within the pilot's downward FOV to see cockpit controls and displays. Thus, either is an acceptable HMD design capable of high light transmission (at least 50% in our example).

Exit Pupil.   The entrance pupil of the eye is an image of the iris and determines the amount of light that can enter the eye. When viewing a pupil forming display, it is essential that the eye pupil be fully illuminated by the display pupil to achieve maximum brightness [Fig. 7.8(a)]; otherwise, vignetting can occur. Therefore, the minimum exit pupil size (d_{ex}) should include the eye pupil (d'_{eye}), plus allow for eye movements that scan across the FOV (d_t). Also, an additional allowance must be made to compensate for any unavoidable helmet slippage:

$d_{ex} = d'_{eye} + d_t + d_{\text{helmet slippage}}$.

To obtain d_{ex} we must calculate each term in turn.

Figure 7.8(b) shows the eye pupil geometry at the edge of the FOV. It is clear that the actual eye pupil length as measured at the exit pupil plane is a cosine projection of the real pupil diameter. At that position,
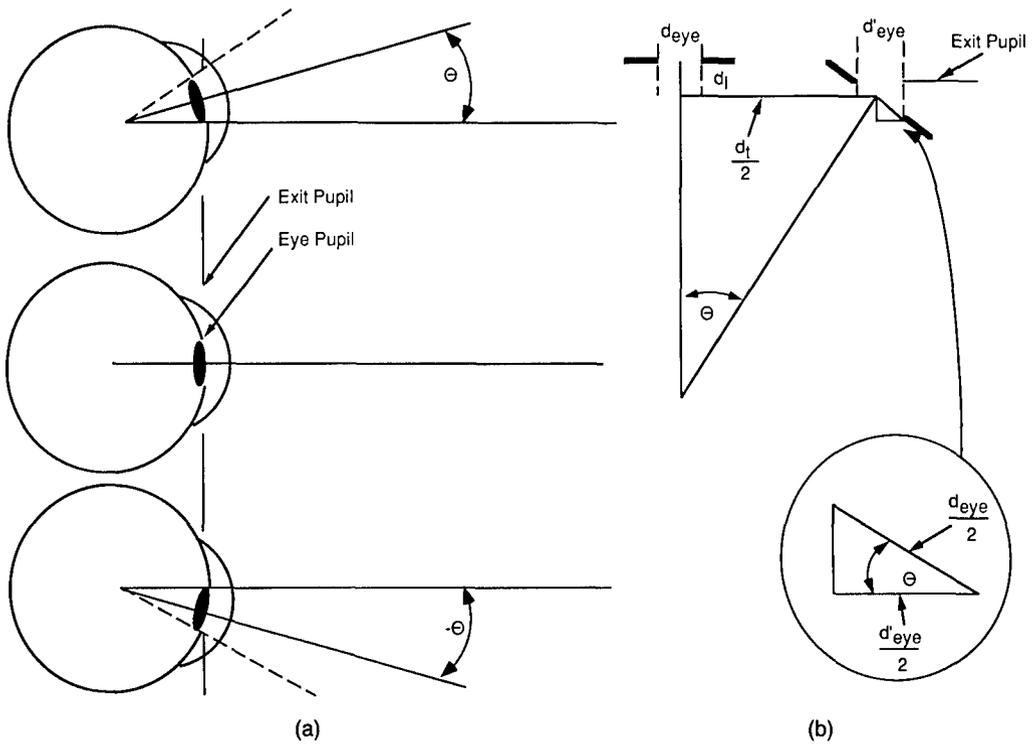
Fig. 7.8 Exit pupil geometry.

$$d'_{eye} = d_{eye} \cos\theta \ , \tag{7.10a}$$

where d_{eye} is the eye pupil diameter in millimeters. With an adapting luminance of 100 fL, the eye pupil diameter[27] is 3 mm. Therefore, d'_{eye} is 2.90 mm. Figure 7.8(b) also traces the chord as the eye pupil scans to the edge of the FOV. The length of the chord is the lateral translation of the eye pupil caused by eye movement and is given by:

$$d_t = 2r \sin\theta \ , \tag{7.10b}$$

where r is the radius of eye rotation. The longitudinal translation is

$$d_l = r(1 - \cos\theta) \ . \tag{7.10c}$$

If we assume a center of rotation of 10 mm behind the entrance pupil[28] and a FOV of 30 deg, then d_t equals 5.18 mm and d_l is 0.34 mm.

Finally, based on laboratory observations, the best fitted helmet slippage can be estimated conservatively to be ± 3 mm. Thus, for a good display, the exit pupil size, as a minimum, is

$$d_{ex} = 2.90 + 5.18 + 6 = 14 \text{ mm} \ .$$
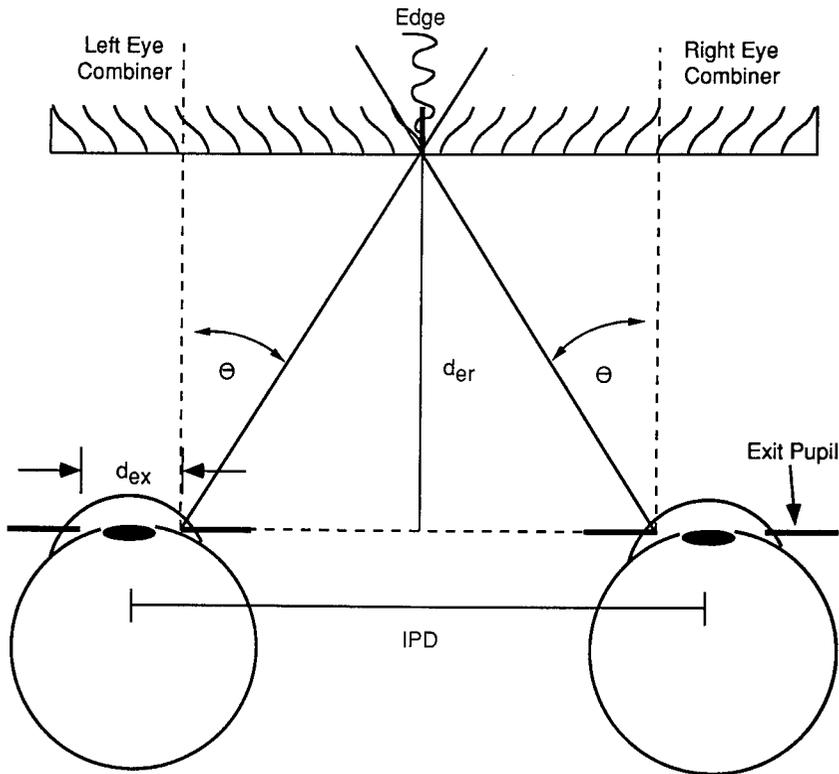
Fig. 7.9 Eye relief geometry.

Obviously, the larger the exit pupil, the more forgiving the display. However, whereas a larger exit pupil is desirable, it also translates to larger relay optics and increases the overall display weight.[23]

Eye Relief. Eye relief is defined as the distance between the exit pupil and the first optical element (usually the combiner). Ideally, the eye pupil is placed at the exit pupil. The eye relief should be long enough to accommodate other life-support gear that the pilot may wear beneath the HMD, e.g., oxygen mask, chemical and biological mask, laser protective shield, or just corrective spectacles. In the case of simple eyeglasses, since the distance from the corneal apex to the back surface of the glasses is less than 20 mm, an eye relief of not less than 20 mm is needed. But there is a limit as to how long the eye relief can be; for a given FOV and exit pupil, the size of the two combiners in front of the eyes increases quickly with increasing eye relief (weight and CG problem!), and eventually they get in the way of each other, as illustrated in Fig. 7.9. As shown,

$$d_{er} = \frac{(\text{IPD} - d_{ex})/2}{\tan\theta} \, , \qquad\qquad (7.11)$$

where d_{er} is the eye relief in millimeters and IPD is the interpupillary distance separation in millimeters. For example, using a minimum IPD of 59 mm, an exit pupil of 14 mm, and a FOV of 30 deg, the maximum eye relief is 84 mm. Otherwise, to go beyond 84 mm the inside of two combiners would have to be shaved off and a cusp in the middle (the loss of some nasal fields from both eyes) would become apparent. This has been known to exist in some HMDs with wide FOVs or large exit pupils.

7.3.3.5 Optomechanical Adjustments.

Like every real-world optical system, optomechanical adjustments are necessary for HMDs. They are needed to compensate for normal human anthropometric variabilities.

Interpupillary Distance Adjustment.   The IPD distribution as compiled on army personnel[29] (may not be valid for the aviator population) indicates that the 5th to 95th percentile range covers 58.8 to 71.0 mm. HMD designs such as the Off-aperture HMD[23,25] that employ two independent, physically separable optical channels can provide an IPD adjustment with minimal effort. However, for some designs, such as the Holographic HMD,[19] an IPD adjustment is difficult because the combiner is one continuous piece. For these designs, the IPD range can be covered by enlarging the exit pupil beyond the minimum size. For example, if an IPD range of 59 to 71 mm is to be covered, with a minimum 14-mm exit pupil size provided for everyone within that range, the HMD should be designed for an exit pupil size of at least 20 mm [(71 − 59)/2 + 14] with the nominal IPD set at 65 mm.

Vertical Adjustment.   The variability[24] (5th to 95th percentile) of the distance between the eye and the top of head is about 1 in. A 1-in. range is too much to be covered entirely by enlarging the exit pupil; but a small vertically adjustable attachment can probably be incorporated onto the helmet with a minimum weight cost, and can be further compensated for by better helmet fitting.

Longitudinal Adjustment.   The variability[24] (5th to 95th percentile) of the distance between the eye and the back of head is about 0.8 in., some of which can be best taken care of by a better helmet fit. However, some variability is already compensated for by the "extended" eye relief. As shown in Fig. 7.10, the distance in millimeters between the nominal exit pupil (design eye position) and the back 50% vignetting point is

$$d_{50} = \frac{(d_{ex}/2)}{\tan\theta} \ .$$
(7.12)

In our example, d_{ex} = 14 mm and θ = 15 deg, therefore, d_{50} = 26 mm or 1 in. For this example, the entire variability of 0.8 in. can be covered. Unfortunately, except for the nominal eye position, any eye movement will result in some degree of vignetting. Perhaps a large exit pupil and better helmet fit can accommodate all users.

Rotational Adjustment.   The physiological eye rest position in the dark is 6 deg below the upper jaw plane.[30] Based on 16 observers, the standard deviation is 11 and 13 deg for the right and left eye, respectively. In another HMD application-specific study[31] based on 10 subjects, the seatback position (or head
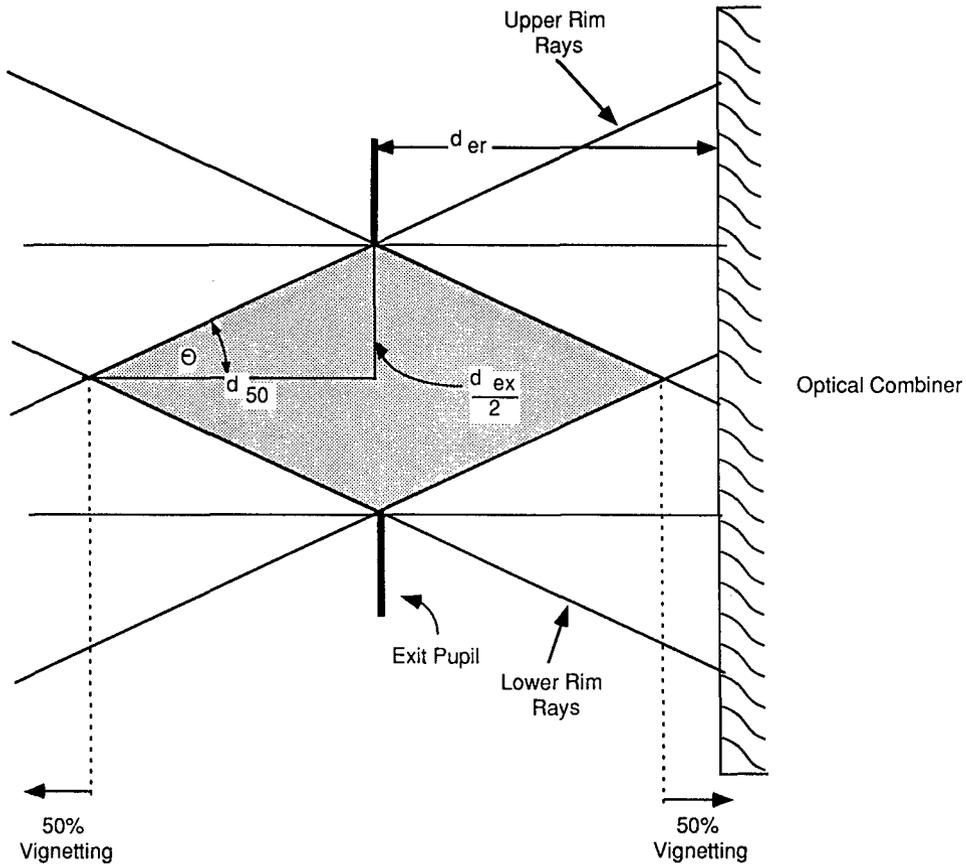
Fig. 7.10 Extended eye relief geometry.

tilt) had small but significant effects on the eye LOS. With the head upright, the LOS is at -0.5 deg, and drops to -5 to -12 deg with declined seatbacks. This study also showed that the eye LOS does not vary with target distance in a see-through display such as an HMD. To provide a 10-deg adjustment on the helmet is not trivial by any mechanical means. Therefore, until more studies can be performed to collect the user population data, a better helmet fit and large exit pupil may alleviate some of the rotational errors without the penalty of adding more weight on the helmet with a mechanical adjustment.

7.3.3.6 Weight and Center of Gravity. As yet, there is no published safe weight and CG requirement for cockpit usage. The 4.5-lb limit[32–34] seems to be deemed as the not-to-exceed safe limit for ejection compatibility and crash worthiness. There is also a concern that extra weight and unbalanced CG will cause fatigue and adversely affect the pilot's head aiming and tracking performance.

7.3.3.7 Future Development. Figure 7.11 shows an artist's conception of a high-transmission binocular HMD that is being prototyped for research. It is an off-axis design whose combiner is a simple toroidal mirror[25] rather than a complex surface,[35] so that its quality is limited not by the difficulty in achieving its profile, but in the smoothness of the pattern/molding process. The visor combiner will be made of polycarbonate plastics to satisfy safety concerns. Table 7.1 gives the list of performance goals to be achieved. The first prototype will be completed in 1993 and should provide a baseline for the type of HMD that has been illustrated for discussion throughout this section.

Fig. 7.11 Artist's conception of a binocular HMD under development.

Table 7.1   Performance Goals

Binocular FOV	22.5 V × 30 H deg
Resolution ($f_{c\text{-HMD}}$)	0.8 cycle/mrad
See-through transmission	>60%
CRT-eye transmission (t_{optics})	50%
Exit pupil (d_{ex})	15 V × 21 H mm
Eye relief (d_{er})	105 mm
Interpupillary distance (IPD)	62 to 74 mm
CRT format (F)	19-mm diameter, active
Number of CRTs	2

Helmet Compatibility
    Weight
    Center of gravity (CG)
	HGU-53/P
	<4 lb
	within 1 in. of head CG

Binocular alignment allowance Light imbalance	<6 arc min <5%
Distortion	<2%
Display focus tolerance	at infinity

7.3.4   Helmet-Mounted Sight

The performance requirement for the helmet-mounted sight (HMS) depends on the capability of human head dynamics. In the past, suitable HMS hardware was not available for field studies. This situation is about to change. A militarized HMS[36] capable of supporting both laboratory and airborne research with the following performance parameters is now within reach:

- *Motion box:* as large as the entire cockpit volume, even when the tracker is next to a conducting metal surface abundant in the cockpit
- *Update rate:* as high as 240 Hz
- *Accuracy:* as good as 1 to 2 mrad within the aircraft forward ± 30-deg cone to improve targeting and weapon delivery
- *Resolution:* as much as 14 bits (0.4 mrad) to eliminate undesirable discrete display jumps caused by head jitter.

Furthermore, the new HMS is also programmable and can easily implement various adaptive "filters" to minimize uncommanded head LOS movement caused by vibrational biodynamic interference. It is difficult to envision the need for an HMS to operate any better than the above, because these performances are approaching the human capability itself.[37,38]

Integrating the HMS into an aircraft is difficult because it involves locating the optimum placement for the HMS source to reduce cockpit-induced field distortion. The new HMS works in concert with an automated cockpit mapping fixture that improves the overall system accuracy by eliminating the past laborious error-prone process. However, the VCS engineer also needs to be aware that the CRT causes dynamic scattering and field distortion that is not easily predictable and correctable. Therefore, the CRT placement on the helmet

is quite critical in this regard. Unfortunately, complete analytical modeling for the optimum CRT location is not possible. A practical guideline[36] to follow is to avoid a location where the CRT has the greatest opportunity to block the HMS emitter's LOS to the receiver and degrade HMS performance.

7.4   OTHER DESIGN ISSUES AND CONCERNS

There are many issues that are integral parts of the VCS design considerations but are not included in Sec. 7.3. Some of them are highlighted here. One important issue is the image display lag time. It is obvious that for the VCS to work as intended, the lag time between the real world and the imaged display should be minimized, perhaps to less than 80 ms (see Ref. 39). Lag is introduced when (1) the HMS cannot calculate or update the new LOS solution fast enough, (2) the sensor gimbal slew rate is too slow to follow the head LOS, and (3) the sensor image processing (or update rate) is too low.

How fast should the update rate be? One laboratory study[40] demonstrated that, for a line-drawing display with a 60-Hz refresh and update rate, the maximum acceptable display velocity is a mere 7 deg/s. Typical (50th percentile) head velocities of 30 deg/s and maximum (90th percentile) velocities of 200 deg/s have been reported during helicopter maneuvers.[41] This would require the sensor update rate to be several hundred hertz, which is beyond the state of the art for VCS technology. The proper trade-off between sensor update rate and smooth motion percept awaits more research. To minimize lags and errors between head and sensor LOSs, a 300 deg/s slew rate and 5000 deg/s^2 acceleration[42] for the sensor gimbal mechanism working with either a 240-Hz (see Ref. 36) or a 120-Hz (see Ref. 43) HMS may be required. Moreover, a dedicated high-speed data bus can be installed between the HMS and sensor gimbal to reduce data transport delay further and improve VCS performance.

Another concern involves partial overlap binocular FOV. For example, an engineering approach to enlarge the FOV, yet maintain resolution and weight, is to overlap the two ocular FOVs partially so that only a portion of the central FOV is binocular (a panoramic display). Some evidence has been accumulated to suggest that any panoramic display, when binocular overlap is less than 100%, seems to elicit abnormal head movements.[44] A field study[45] based on simple car maneuverability showed no significant performance degradation using various binocular overlap levels (minimum 30 deg) and configurations (divergent or convergent). Another study[46] found, for tactical pilot training, better target detection with increasing binocular overlap. As a compromise for a wider FOV a 38-deg binocular overlap was eventually chosen for the trainer. Therefore, more quantitative data measured under mission specific conditions are needed to better understand the psychophysical nature of this trade-off and its relationship to perception.

Other human factors concerns are complex and interrelated and include display optics tolerances[13,47–50] and human binocular performance.[34,44,51–53]

7.5   SUMMARY

A visually coupled system is a complex technology that requires an interdisciplinary team approach to resolve all the system design and engineering

issues. Nevertheless, many design parameters can be systematically estimated based on simple known visual and optical criteria. All our calculations are done at the cutoff frequency for worst or best case conditions. In reality, these numbers are too optimistic.

There may be other interesting and worthy design parameters that are not covered here. This chapter is meant to be a brief "walk through," emphasizing the basic interrelationships between design parameters. The reader is encouraged to explore other aspects of this technology in the referenced materials.

Acknowledgments

References

1. J. A. Birt and H. L. Task, "A symposium on visually coupled systems: development and application," Report AMD-TR-73-1, Air Force Aerospace Medical Division, Brooks AFB, TX (1973).

2. P. C. Smith, *Close Air Support: An Illustrated History, 1914 to the Present,* p. 152, Orion Books, New York (1990).

3. R. Buchroeder, "An optical analysis of the Farrand VCASS helmet-mounted display," Report AFAMRL-TR-83-072, Air Force Aerospace Medical Research Laboratory, Wright-Patterson AFB, OH (1983).

4. CAE Electronics, "Wide-field-of-view, helmet-mounted infinity display system development," Report AD-A149641, Air Force Human Resources Laboratory, Brooks AFB, TX (1984).

5. L. N. Biberman, Ed., *Perception of Displayed Information,* p. 185, Plenum Press, New York (1977).

6. L .B. Scott and L. R. Condiff, "C2NVEO advanced FLIR systems performance model," in *Infrared Imaging Systems: Design, Analysis, Modeling, and Testing,* G. Holst, Ed., *Proceedings of the SPIE* **1309,** 168–180 (1990).

7. B. L. Butterfield, "F-16 helmet-mounted display flight evaluations," in *Helmet-mounted Displays II,* R. Lewandowski, Ed., *Proceedings of the SPIE* **1290,** 226–235 (1990).

8. R .H. Vollmerhausen and C. J. Nash, "Design criteria for helicopter night pilotage sensors," in *Proceedings of the IRIS Symposium on Passive Sensors,* Johns Hopkins University Applied Physics Laboratory (1990).

9. D. F. Barbe and S. B. Campana, "Imaging arrays using the charge-coupled concept," in *Advances in Image Pickup and Display,* Vol. 3, B. Kazan, Ed., p. 239, Academic Press, New York (1977).

10. B. H. Tsou and B. M. Rogers-Adams, "The effect of aspect ratio on helmet-mounted display field-of-view," in *Proceedings of the Aerospace Medical and Life Support Systems (Working Party 61) Aeromedical Aspects of Vision Symposium,* November 6–15, 1990, Toronto, Canada, Vol. IV, pp. 136–146 (1990).

11. J. S. Tatro and R. R. Taylor, "Helmet mounted displays and the emerging attack rotorcraft counterair mission," in *Helmet Mounted Displays,* J. Carollo, Ed., *Proceedings of the SPIE* **1116,** 37–46 (1989).

12. W. D. Pearson, Jr., B. C. King, Jr., and R. L. Hilgendorf, "Single seat attack-night simulation study Phase I, Part 1: Target detection/classification," Report ASD-TR-77-30, Air Force Aeronautical Systems Division, Wright-Patterson AFB, OH (1977).

13. W. J. Smith, *Modern Optical Engineering: The Design of Optical Systems,* 2nd ed., p. 356, McGraw-Hill, New York (1990).

14. W. L. Carel, J. A. Herman and L. A. Olzak, "Design criteria for imaging sensor displays," Report AD-A055411, Office of Naval Research, Washington, DC (1978).

15. H. W. Leibowitz and D. A. Owens, "Can normal outdoor activities be carried out during civil twilight?" *Applied Optics* **30**(24), 3501–3503 (1991).

16. D. F. Kocian, "Design considerations for virtual panoramic display (VPD) helmet systems," in *Proceedings of the AGARD Symposium on the Man-Machine Interface in Tactical Aircraft Design and Combat Automation,* AGARD-CP-425, pp. 22/1–32, AGARD, Nevilly sur Seine, France (1988).

17. D. J. Rotier, "Optical approaches to the helmet mounted display," in *Helmet Mounted Displays,* J. Carollo, Ed., *Proceedings of the SPIE* **1116,** 14–18 (1989).

18. J. G. Droessler and D. J. Rotier, "Tilted cat helmet mounted display," in *Helmet Mounted Displays,* J. Carollo, Ed., *Proceedings of the SPIE* **1116,** 19–26 (1989).

19. R. A. Lohmann and A. Z. Weisz, "Helmet-mounted displays for helicopter pilotage: design configuration tradeoffs, analyses, and test," in *Helmet Mounted Displays,* J. Carollo, Ed., *Proceedings of the SPIE* **1116,** 27–32 (1989).

20. A. A. Cameron, "The development of the combiner eyepiece night vision goggle," in *Helmet-mounted Displays II,* R. Lewandowski, Ed., *Proceedings of the SPIE* **1290,** 16–29 (1990).

21. J. E. Melzer and E. W. Larkin, "An integrated approach to helmet display system design," in *Display System Optics,* A. Cox, R. Hartman, Eds., *Proceedings of the SPIE* **778,** 83–88 (1987).

22. D. Naor, O. Arnon, and A. Avnur, "A lightweight innovative helmet airborne display and sight (HADAS)," in *Display System Optics,* A. Cox, R. Hartman, Eds., *Proceedings of the SPIE* **778,** 89–96 (1987).

23. M. Shenker, "Optical design criteria for binocular helmet-mounted displays," in *Display System Optics,* A. Cox, R. Hartman, Eds., *Proceedings of the SPIE* **778,** 70–78 (1987).

24. R. A. Buchroeder, Instructor, Helmet-Mounted Displays Short Course at the SPIE Technical Symposium on Display System Optics, A. Cox, R. Hartman, Eds., May 21–22, 1987, Orlando, Florida.

25. Farrand Optical Company, "Helmet mounted display optical design," Report AD-A098154, Armstrong Aerospace Medical Research Laboratory, Wright-Patterson AFB, OH (1984).

26. D. F. Kocian, "Visually coupled systems (VCS): The virtual panoramic display 'systems'," in *Proceedings of the Fifth Annual Space Operations, Applications, and Research (SOAR, 1991) Symposium,* July 9–11, 1991, Houston, Texas, pp. 548–561, Report CP-3127, NASA (1991).

27. R. J. Farrell and J. M. Booth, "Design handbook for imagery interpretation equipment," Boeing Aerospace, Seattle, WA (1984).

28. A. G. Bennett and R. B. Rabbetts, *Clinical Visual Optics,* 2nd ed., pp. 169 and 257, Butterworth, London (1989).

29. C .C. Gordon, T. Churchill, C. E. Clauser, B. Bradtmiller, J .T. McConville, I. Tebbetts, and R. A. Walker, "1988 anthropometric survey of U.S. Army personnel: summary statistics interim report," Report NATICK-TR-89-027, Army Natick Research, Development and Engineering Center, Natick, MA (1989).

30. J. Levy, "Physiological position of rest and phoria," *American Journal of Ophthalmology* **68**(4), 706–713 (1968).

31. D. L. McMullin, "Effects of image distance and seat inclination on the line-of-sight angle," MS thesis, Virginia Polytechnic Institute and State University (1990).

32. D. A. Fulghum, "Navy orders contractors to stop work on ejectable night vision helmets," *Aviation Week & Space Technology* **133**(23), 67–68 (Dec. 3, 1990).

33. D. H. Glaister, "Head injury and protection," in *Aviation Medicine,* 2nd ed., pp. 174–184, J. Ernsting, P. King, Eds., Butterworths, London (1988).

34. C. E. Rash, R. W. Verona, and J. S. Crowley, "Human factors and safety considerations of night vision systems flight using thermal imaging systems," Report USAARL-90-10, U.S. Army Aeromedical Research Laboratory, Fort Rucker, AL (1990).

35. R. J. Withrington, "Optical design of a holographic visor helmet-mounted display," in *Computer-Aided Optical Design, Proceedings of the SPIE* **147,** 161–170 (1978).

36. D. F. Kocian, "Visually coupled system (VCS): Preparing the engineering research framework," in *Proceedings of the 11th Annual IEEE/AESS Symposium: The Cockpit of the 21st Century—Will High Tech Payoff?,* November 28, 1990, Dayton, Ohio, pp. 29–38 (1990).

37. R. W. Verona, J. C. Johnson, and H. D. Jones, "Head aiming/tracking accuracy in a helicopter environment," Report A058444, U.S. Army Aeromedical Research Laboratory, Fort Rucker, AL (1979).

38. J. D. Grossman, "Tracking with the vertical plane of an aircraft," Report NWC TP 5678, Naval Weapons Center, China Lake, CA (1974).

39. R. H. Y. So and M. J. Griffin, "Effects of time delays on head tracking performance and the benefits of lag compensation by image deflection," in *Proceedings of the AIAA Flight Simulation Technologies Conference,* August 12–14, 1991, New Orleans, Louisiana, pp. 124–130 (1991).

40. G. V. Kellogg and C. A. Wagner, "Effects of update and refresh rates on flight simulation visual displays," Report TM-100415, National Aeronautics and Space Administration (1988).

41. R. W. Verona, C. E. Rash, W. R. Holt and J. K. Crosley, "Head movements during contour flight," Report A181203, U.S. Army Aeromedical Research Laboratory, Fort Rucker, AL (1986).

42. D. Allen, B. Tsou, G. Gieske, J. Bien, M. Shipley, and J. Walker, "System performance of a servo-optical projection system (SOPS)," in *Proceedings of the Image IV Conference,* June 23–26, 1987, Phoenix, Arizona, pp. 121–127 (1987).

43. J. C. Krieg, A. G. Rodgers, H. R. Jones, and M. R. Schneider, "A 4-msec low-latency 120-Hz electromagnetic tracker for virtual reality applications," in *Proceedings of the Society for Information Display International Symposium,* May 17–22, Boston, Massachusetts, pp. 77–79, SID, Playa del Rey, CA (1992).

44. B. McLean and S. Smith, "Developing a wide field-of-view HMD for simulators," in *Display System Optics,* A. Cox, R. Hartman, Eds., *Proceedings of the SPIE* **778,** 79–82 (1987).

45. B. H. Tsou, B. M. Rogers-Adams, and C. D. Goodyear, "The evaluation of partial binocular overlap on car maneuverability: a pilot study," in *Proceedings of the Fifth Annual Space Operations, Applications, and Research (SOAR, 1991) Symposium,* July 9–11, 1991, Houston, Texas, pp. 562–568, Report CP-3127, NASA (1991).

46. R. E. Barrette, "Wide-field-of-view full-color high-resolution helmet-mounted display," in *Proceedings of the Society for Information Display International Symposium,* May 17–22, 1992, Boston, Massachusetts, pp. 69–72, SID, Playa del Rey, CA (1992).

47. H. C. Self, "Optical tolerances for alignment and image differences for binocular helmet-mounted displays," Report AD-A174536, Armstrong Aerospace Medical Research Laboratory, Wright-Patterson AFB, OH (1986).

48. M. Shenker, "Survey paper on biocular optical systems," in *Technical Digest of the Optical Society of America Annual Meeting,* San Francisco, California (1972).

49. P. Rogers, "Visual optics, short course notes," in *OPTCON'90,* November 4–9, 1990, Boston, Massachusetts (1990).

50. M. Shenker and P. Weissman, "Aberrational effects in binocular helmet-mounted displays," in *Proceedings of the Society for Information Display International Symposium,* May 6–10, 1991, Anaheim, CA, pp. 318–320, SID, Playa del Rey, CA (1991).

51. S. G. Hart and M. S. Brickner, "Helmet-mounted pilot night vision systems: human factors issues," in *Spatial Displays and Spatial Instruments,* S. R. Ellis, M. K. Kaiser, A. Grunwald, Eds., Report CP-10032, National Aeronautics and Space Administration, Moffett Field, CA (1989).

52. H. L. Task, "Optical and visual considerations in the specification and design of helmet-mounted displays," in *Proceedings of the Society for Information Display International Symposium,* May 6–10, 1991, Anaheim, California, pp. 297–300, SID, Playa del Rey, CA (1991).

53. R. Jones, "Binocular fusion factors in visual displays," in *Proceedings of the Society for Information Display International Symposium,* May 17–22, 1992, Boston, Massachusetts, pp. 287–289, SID, Playa del Rey, CA (1992).

Author's note: Additional human factors references became available after the chapter was submitted for publication.

On Interpupillary Distance Adjustment:

Donelson, S. M., and Gordon, C. C., "1988 anthropometric survey of US Army personnel: pilot summary statistics," Report Natick-TR-91-040, US Army Natick Research, Development and Engineering Center, Natick, MA (1991).

On Rotational Adjustment:

Jampel, R. S., and Dian, X. S., "The primary position of the eyes, the resetting saccade, and the transverse visual head plane," *Invest. Ophthalmol. Vis. Sci.* **33**(8), 2501–2510 (1992).

On Weight and Center of Gravity:

Knox III, F. S., Buhrman, J. R., and Perry, C. E., "Biomechanics of ejection safety for night vision systems," in *Proceedings of the Night Vision Conference and Exhibition,* September 2–3, 1992, London UK, The Shepard Press, Ltd., UK (1993).

On Other Issues:

Chen, J. S., "A study on the effects of low update rate on visual displays," in *Proceedings of the Society for Information Display International Symposium,* May 16–21 1993, Seattle, Washington, pp. 510–513, SID, Playa del Rey, CA (1993).

Grigsby, S. S., and Tsou, B. H., "Visual factors in the design of partial overlap binocular helmet-mounted displays," in *Proceedings of the Society for Information Display International Symposium,* May 16–21 1993, Seattle, Washington, pp. 185–187, SID, Playa del Rey, CA (1993).

Jones, R., "Proximal accommodations in virtual displays," in *Proceedings of the Society for Information Display International Symposium,* May 16–21 1993, Seattle, Washington, pp. 195–198, SID, Playa del Rey, CA (1993).

Index to Volume 8

Cumulative Index

Volumes 1 through 8

Volume numbers are shown in boldface type followed by page numbers.