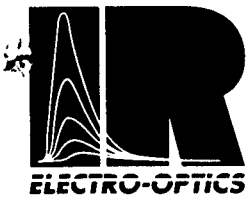
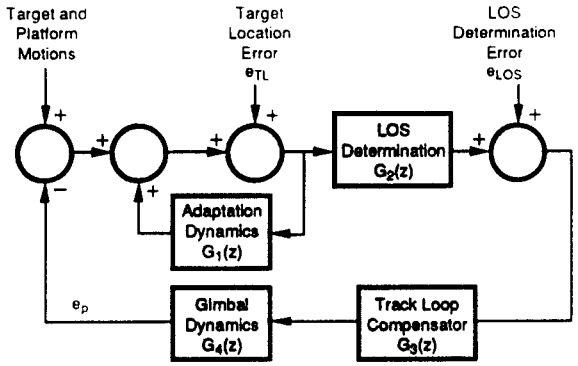


The Infrared & Electro-Optical Systems Handbook

VOLUME 4

Electro-Optical Systems Design, Analysis, and Testing

Michael C. Dudzik, *Editor*



Electro-Optical Systems Design, Analysis, and Testing

VOLUME 4

The Infrared and Electro-Optical Systems Handbook

The Infrared and Electro-Optical Systems Handbook

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

Chapter 3. Optomechanical System Design, Daniel Vukobratovich

Chapter 4. Infrared Imaging System Testing, Gerald C. Holst

Chapter 5. Tracking and Control Systems, Robert E. Nasburg

Chapter 6. Signature Prediction and Modeling, John A. Conant, Malcolm A. LeCompte

■ **VOLUME 5.** **Passive Electro-Optical Systems,**
Stephen B. Campana, *Editor*

Chapter 1. Infrared Line Scanning Systems, William L. McCracken

Chapter 2. Forward-Looking Infrared Systems, George S. Hopper

Chapter 3. Staring-Sensor Systems, Michael J. Cantella

Chapter 4. Infrared Search and Track Systems, Joseph S. Accetta

■ **VOLUME 6.** **Active Electro-Optical Systems,** Clifton S. Fox, *Editor*

Chapter 1. Laser Radar, Gary W. Kamerman

Chapter 2. Laser Rangefinders, Robert W. Byren

Chapter 3. Millimeter-Wave Radar, Elmer L. Johansen

Chapter 4. Fiber Optic Systems, Norris E. Lewis, Michael B. Miller

■ **VOLUME 7.** **Countermeasure Systems,** David Pollock, *Editor*

Chapter 1. Warning Systems, Donald W. Wilmot, William R. Owens, Robert J. Shelton

Chapter 2. Camouflage, Suppression, and Screening Systems, David E. Schmieder, Grayson W. Walker

Chapter 3. Active Infrared Countermeasures, Charles J. Tranchita, Kazimieras Jakstas, Robert G. Palazzo, Joseph C. O'Connell

Chapter 4. Expendable Decoys, Neal Brune

Chapter 5. Optical and Sensor Protection, Michael C. Dudzik

Chapter 6. Obscuration Countermeasures, Donald W. Hoock, Jr., Robert A. Sutherland

■ **VOLUME 8.** **Emerging Systems and Technologies,**
Stanley R. Robinson, Editor

Chapter 1. Unconventional Imaging Systems, Carl C. Aleksoff, J. Christopher Dainty, James R. Fienup, Robert Q. Fugate, Jean-Marie Mariotti, Peter Nisenson, Francois Roddier

Chapter 2. Adaptive Optics, Robert K. Tyson, Peter B. Ulrich

Chapter 3. Sensor and Data Fusion, Alan N. Steinberg

Chapter 4. Automatic Target Recognition Systems, James W. Sherman, David N. Spector, C. W. "Ron" Swonger, Lloyd G. Clark, Edmund G. Zelnio, Terry L. Jones, Martin J. Lahart

Chapter 5. Directed Energy Systems, Gary Golnik

Chapter 6. Holography, Emmett N. Leith

Chapter 7. System Design Considerations for a Visually-Coupled System, Brian H. Tsou

Copublished by

Infrared Information Analysis Center
Environmental Research Institute of Michigan
Ann Arbor, Michigan USA

and

SPIE OPTICAL ENGINEERING PRESS
Bellingham, Washington USA

Sponsored by

Defense Technical Information Center, DTIC-DF
Cameron Station, Alexandria, Virginia 22304-6145

Electro-Optical Systems Design, Analysis, and Testing

Michael C. Dudzik, *Editor*
Environmental Research Institute of Michigan

VOLUME 4

The Infrared and Electro-Optical Systems Handbook

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

Library of Congress Cataloging-in-Publication Data

The Infrared and electro-optical systems handbook / Joseph S. Accetta, David L. Shumaker, executive editors.
p. cm.
Spine title: IR/EO systems handbook.
Cover title: The Infrared & electro-optical systems handbook.
Completely rev. ed. of: Infrared handbook. 1978
Includes bibliographical references and indexes.
Contents: v. 1. Sources of radiation / George J. Zissis, editor — v. 2. Atmospheric propagation of radiation / Fred G. Smith, editor — v. 3. Electro-optical components / William D. Rogatto, editor — v. 4. Electro-optical systems design, analysis, and testing / Michael C. Dudzik, editor — v. 5. Passive electro-optical systems / Stephen B. Campana, editor — v. 6. Active electro-optical systems / Clifton S. Fox, editor — v. 7. Countermeasure systems / David Pollock, editor — v. 8. Emerging systems and technologies / Stanley R. Robinson, editor.
ISBN 0-8194-1072-1
1. Infrared technology—Handbooks, manuals, etc. 2. Electrooptical devices—Handbooks, manuals, etc. I. Accetta, J. S. II. Shumaker, David L. III. Infrared handbook. IV. Title: IR/EO systems handbook. V. Title: Infrared & electro-optical systems handbook.
TA1570.I5 1993
621.36'2—dc20 92-38055
CIP

Copublished by

Infrared Information Analysis Center
Environmental Research Institute of Michigan
P.O. Box 134001
Ann Arbor, Michigan 48113-4001

and

SPIE Optical Engineering Press
P.O. Box 10
Bellingham, Washington 98227-0010

PRINTED IN THE UNITED STATES OF AMERICA

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

This handbook was prepared by the Infrared Information Analysis Center (IRIA) in cooperation with the International Society for Optical Engineering (SPIE). The IRIA Center, Environmental Research Institute of Michigan, is a Defense Technical Information Center-sponsored activity under contract DLA-800-C-393 and administrated by the Defense Electronics Supply Center, Defense Logistics Agency.

This work relates to the aforementioned ERIM contract and is in part sponsored by the Department of Defense; however, the contents do not necessarily reflect the position or the policy of the Department of Defense or the United States government and no official endorsement should be inferred.

The use of product names does not in any way constitute an endorsement of the product by the authors, editors, Department of Defense or any of its agencies, the Environmental Research Institute of Michigan, or the International Society for Optical Engineering.

The information in this handbook is judged to be from the best available sources; however, the authors, editors, Department of Defense or any of its agencies, the Environmental Research Institute of Michigan, or the International Society for Optical Engineering do not assume any liability for the validity of the information contained herein or for any consequence of its use.

Contents

Introduction

This volume is devoted to the design, analysis, and testing methodologies of electro-optical systems. Specifically, the volume describes the image chain process used in sensor development. This volume is intended for scientists and engineers who desire a working knowledge of the principles, models, and practice of electro-optical sensor development. The volume contains information that is supplemented with relevant data, illustrative examples, and supporting references.

Although the primary thrust of this volume is devoted to imaging sensors, much of the information is relevant to other imaging and nonimaging sensor systems. The growing use of electro-optical sensors in military and commercial applications is being used to address a wide range of unique sensing problems. While the sensing problems are various, the physical principles of sensor development are similar. This volume was prepared to address those common areas of sensor development.

The volume is made up of six chapters. Chapter 1 by J. Michael Lloyd is an introduction to electro-optical imaging systems, with special emphasis on thermal imaging sensor systems. This includes a systems engineering approach to understanding source, sensor, and image scene characteristics. Central to this chapter is the application of optical transfer function theory to describe sensor response to various input functions. This chapter provides the tools to make first-order estimates of the theoretical limits of sensor resolution, frequency response and noise-limited performance of imaging sensors, SNR and resolution criteria, and scene clutter strengths relative to signal sources.

Chapter 2 by James Howe is an introduction to the fundamentals of electro-optical imaging systems performance prediction. This chapter provides a detailed description of the parameters of performance associated with detection, recognition, and identification for first- and second-generation imaging sensors. Howe's chapter continues the system engineering approach to sensor development and provides a detailed analysis on the information content of the scene image produced by the sensor system. Of note is the development of the methodology to understand the uses and limitations of first- and second-generation thermal imaging system performance models.

Chapter 3 by Daniel Vukobratovich provides theory and practical data on the application of optomechanical systems design to electro-optical sensors. This chapter treats many of the practical implementation problems associated with the design of electro-optical windows, lens housings, and structural features. This chapter is a natural progression of the sensor system design process to assist the design staff in creating actual sensor designs to meet the performance characteristics envisioned by the systems engineer. Of concern to all system

developers is the mounting of optical elements to ensure optical performance against thermal and pressure effects.

Chapter 4 by Gerald Holst is an introduction to electro-optical system testing and performance evaluation. This chapter provides the methodology and procedures for establishing many of the sensor performance measures used in imaging electro-optical sensors. In addition, Holst provides insight into the practical problems associated with optical collimator, test equipment, and electronic measurements. This chapter complements the image chain analysis by allowing the systems engineer to evaluate the results of actual hardware testing on theoretical design criteria.

Chapter 5 by Robert Nasburg provides discussions on the theory and application of tracking and control systems. With the ever-increasing demand for image processing of sensor information, Nasburg provides a systems engineering approach to sensor-derived information applications. This chapter allows the systems designer to develop an understanding of the information processing requirements placed upon the sensor output. This is especially important in those electro-optical sensors in which preprocessing or prefiltering of scene information occurs before being presented to a human operator for display.

Chapter 6 by John Conant and Malcom LeCompte is an introduction to signature modeling. This chapter provides a practical methodology to develop a first-order estimate of both emitted and reflected radiation from a complex target shape. The technical approach allows the systems developer to approximate the performance of an electro-optical sensor against a new or different target class or in a different application. The chapter further serves as a practical illustration of the physical phenomena inherent to the generation of electro-optical signatures.

This volume of course could not have been assembled without the technical contributions of my many colleagues and fellow workers in the field of electro-optics. I am gratefully indebted to the following individuals for their assistance during the preparation of this volume: H. Kennedy, D. Ratcliff, M. Michel, T. Milson, G. Zissis, J. Accetta, G. Nado, and D. Ausherman. Also, I would like to thank Ms. Nancy Hall of ERIM for her patience in the preparation and coordination of the manuscripts. Finally, I wish to thank my family, Katherine and Michael Dudzik, for allowing me the time necessary to complete this volume.

Michael C. Dudzik
Ann Arbor, Michigan

January 1993

CHAPTER 1

Fundamentals of Electro-Optical Imaging Systems Analysis

J. M. Lloyd
Loral Vought Systems
Grand Prairie, Texas

CONTENTS

1.1 INTRODUCTION

Electro-optical (EO) imaging systems convert electromagnetic radiation at optical wavelengths to electrical signals for source detection and/or analog visual display. They extend human vision in wavelength, signal sensitivity, distance, and environmental conditions. An image of a scene in the selected waveband is produced in a focal plane containing a detection medium, which performs the transduction of light to video signals.

EO imaging has seven dimensions: intensity, time, wavelength, polarization, and the three spatial dimensions. An EO imager maps the three spatial dimensions of a scene in perspective onto the two-dimensional surface of the focal plane. The dimension of range is obtainable if needed by calculations from prior knowledge (passive ranging), by pulsed illumination (range gating), or by coherent illumination and detection (laser radar).

The video produced by an imager typically is a sequence of electrical signals whose amplitudes are proportional to the intensity of the scene at every sampled point. Position information is coded by the location of the signals in time within the sequence. Temporal variations of intensity in the scene typically are sampled at a rate many orders of magnitude slower than the light frequency. Wavelength information may or may not be preserved, and polarization rarely is exploited.

This chapter is written from the viewpoint of military tactical thermal imaging. The theory presented here is applicable to other forms of EO imaging except where blackbody radiation is explicitly used in the equations. The central analyses account for the blurring of image detail by limited optical and electronic bandwidth, for the effects of sampling at discrete lattice points, and for the masking of signal sources by noise.

Those analyses are based on the hypotheses that EO systems are amenable to spatial and temporal frequency-domain analyses, that noise interferes with perception in such a way that the linear ratio of signal to noise is pertinent, and that human vision can be modeled sufficiently accurately to make useful performance predictions. For historical background, more sophisticated analyses, or detailed applications of the topics introduced here, readers are referred to Biberman,[1] Dereniak and Crowe,[2] Hudson,[3] Karim,[4] Lloyd,[5] Spiro and Schlessinger,[6] Rosell and Harvey,[7] Seyrafi,[8] and Shumaker, Wood, and Thacker.[9]

1.1.1 Definitions

Figure 1.1 depicts the basic optical collection features and parameters of an EO imaging system. Table 1.1 contains the symbols, acronyms, and abbreviations used in this chapter. An EO imager has a two-dimensional angular field of coverage, called the field of view (FOV), within which it views the scene. The FOV typically is rectangular with dimensions of A radians horizontally and B radians vertically. The FOV is distinguished from the field of regard (FOR), which is the larger two-dimensional angular coverage within which the imager may be directed by stepping or panning to survey a scene larger than the FOV.

The optical system that maps the object space onto the focal plane is characterized by the linear dimensions of the optical aperture area that collects the scene-radiated energy, and by the largest angle relative to the optical axis

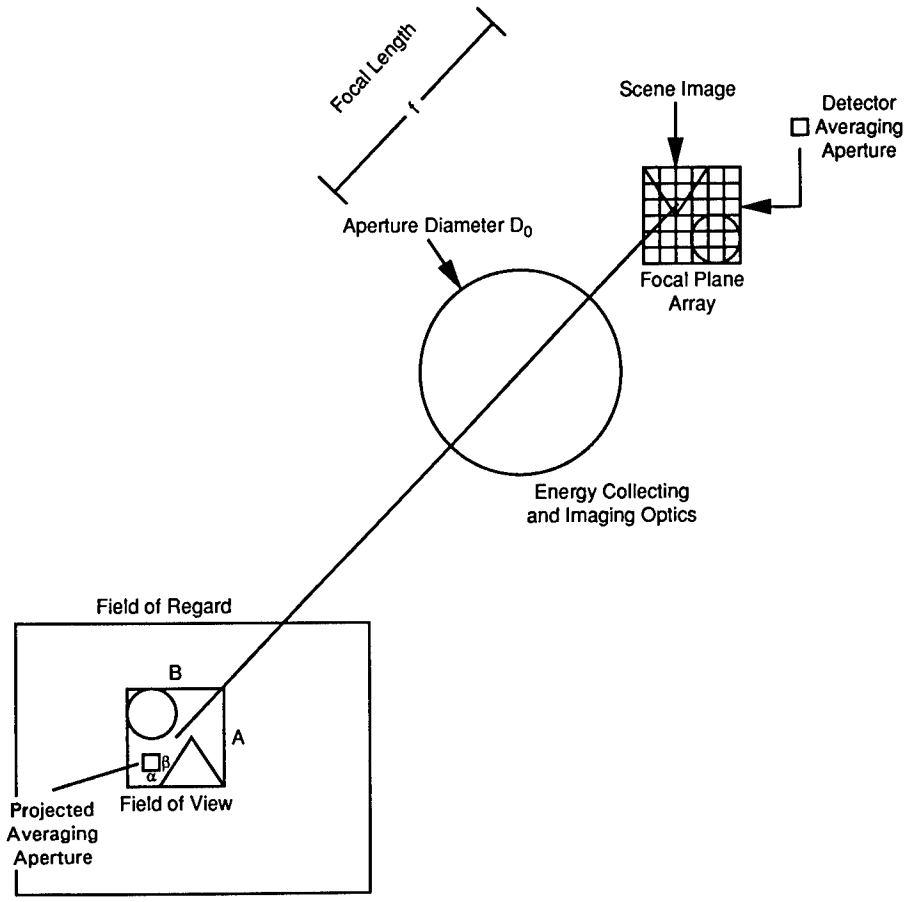


Fig. 1.1 Elementary features of an EO imaging system.

at which the rays are focused onto the detecting surface. Those dimensions and that angle determine the effective optical distance from the aperture, called the *effective focal length,* at which the energy comes to a focus.

The optical aperture area determines how much of a distant source's radiated energy is collected at a given range from the source, ultimately determining scene signal sensitivity. The effective focal length determines the dimensions in the focal plane of a particular scene source at a given range. The effective focal length, the optical aperture, and the optical transmission determine the energy flux of the collected radiation at the focal plane. The ratio of the effective focal length to the aperture diameter is called the *focal ratio,* or *f/#*. The total extent of the sensing surface projected through the optics determines the angular FOV of the device.

The optical system is further characterized by its transmission of light as a function of wavelength. A system is described as operating in the ultraviolet, the visible, or the infrared (IR). The latter is divided into the near-IR, the short-wave IR, the medium-wave IR, and the long-wave IR. The overall spectral response is the relative efficiency, as a function of wavelength, in the conversion of scene photons to electrical signal. This includes the spectral transmis-

Table 1.1 Symbols, Nomenclature, and Units

Symbol or Acronym	Definition	Units
a	detector dimension in the horizontal direction	m
A	horizontal FOV	rad
A_d	detector area	m^2
A_o	optical collecting area	m^2
b	detector dimension in the vertical direction	m
B	vertical FOV	rad
B	as subscript, indicates a quantity associated with the background	—
c	speed of light in vacuum	$m\ s^{-1}$
c	as subscript, indicates cutoff in spatial frequency due to diffraction or to detector geometry	—
c_1	first Planck radiation constant	$W\ \mu m^4\ m^{-2}$
c_2	second Planck radiation constant	$\mu m\ K$
c_3	third Planck radiation constant	—
C	mean value of a Poisson distribution	—
D	detectivity	W^{-1}
D^*	specific detectivity	$cm\sqrt{Hz}/W$
D^*_{BLIP}	background-limited D^*	$cm\sqrt{Hz}/W$
D^{**}_{BLIP}	background-limited D^* for a full hemispherical surround	$cm\sqrt{Hz}/W$
D_o	optical aperture diameter	m
e	as subscript, indicates an energy quantity	—
E	irradiance, either energy or photon, depending on the appended subscript	—
f	effective focal length	m
f_t	electrical (temporal) frequency	Hz
f_x, f_y	spatial frequency, in cycles per unit of angular measure, typically cycles/mrad	cycles/mrad
$f/\#$	focal ratio or F-number	—
f_e	electrical frequency	Hz
$F\{\ \}$	forward Fourier transform	—
$F^{-1}\{\ \}$	inverse Fourier transform	—
FOR	field of regard	rad
FOV	field of view	rad
h	Planck's constant	$W\ s^2$
I	radiant intensity, either energy or photon, depending on the appended subscript	—
k	Boltzmann's constant	—
L	radiance, either energy or photon, depending on the subscript	$J\ K^{-1}$
M	radiant exitance, either energy or photon, depending on the subscript	—
n_B	number of carriers due to background	—
n_s	number of signal carriers	—
n_t	total number of carriers	—
NEP	noise equivalent power	W

(continued)

Table 1.1 *(continued)*

Symbol or Acronym	Definition	Units
NETD	noise equivalent temperature difference	K
PSD	power spectral density, in appropriate radiation units squared per unit of appropriate bandwidth	—
PTF	phase transfer function	rad
q	as subscript, indicates a photon quantity	—
Q	radiant energy, either energy or photon, depending on the appended subscript	—
$\overline{Q_q}$	mean value of photon number	—
r	radial separation of any two points in an autocorrelation function	—
R	slant range to source	m
$\mathcal{R}$	detector responsivity	$V\ W^{-1}$
S_o	shape factor in optics-to-detector coupling	—
t	as subscript, indicates a quantity associated with a target	—
T	absolute temperature	K
T_f	frame time	s
V_s	signal voltage	V
V_n	rms noise voltage	V
Greek:		
α	horizontal detector angular subtense	rad
β	vertical detector angular subtense	rad
$\Delta\lambda$	wavelength band	m
$\Delta\lambda$	as subscript, indicates a quantity integrated over a waveband	—
Δf	noise equivalent electrical frequency bandwidth	Hz
Δn_s	change in number of signal carriers	—
ΔT	small change in temperature	K
$*$	indicates convolution of the quantities on either side of the symbol	—
η_{cs}	cold shielding efficiency	—
η_q	quantum efficiency	—
θ	angle	—
λ	wavelength in micrometers, or as subscript, indicates spectral density	—
Λ	coherence waveband	—
ξ	variable of integration; dummy variable	—
σ	the square root of variance of a random variable	—
σ_e	Stefan-Boltzmann constant in energy units	$W\ m^{-2}\ K^{-4}$
σ_q	Stefan-Boltzmann constant in photon units	$s^{-3}\ m^{-6}\ K^{-9}$
σ_Q	square root of the variance of the photons counted in a measurement	—
τ_a	atmospheric transmission	—
τ_{aa}	apparent atmospheric transmission	—
τ_{cf}	cold filter transmission	—
τ_i	integration period	s
τ_o	optical transmission	—
τ_{oa}	apparent optical transmission	—

Table 1.1 *(continued)*

Symbol or Acronym	Definition	Units
Φ	flux, either energy or photon, depending on the appended subscript	—
χ	interference length	—
ω	total detector angular subtense	sr
Ω	solid angle or total FOV	sr
Ω_{cs}	cold shield solid angle	sr
Ω_p	perfect cold shield solid angle	sr

sion of the optics and the spectral response of the detector. The aperture dimensions and the waveband combine through the phenomenon of optical diffraction to set an upper limit on the spatial frequency that can be resolved by the system. This diffraction cutoff frequency f_c is the ratio of the aperture diameter to the mean wavelength.

The process of conversion of the radiant energy into an electrical signal is called *detection*. The signal quantity that is actually measured may be a voltage or a current. This signal has a maximum value called the *saturated signal*. The detecting surface is characterized by the dimensions over which energy is spatially averaged in the focal plane, called the *averaging aperture* or *unit cell*. Concurrent with detection is the generation of noise, which obscures the signal.

The quality of the detection process is described by the signal-to-noise ratio (SNR), the linear ratio of the signal amplitude to the square root of the statistical variance of the noise. Linear SNR is used because of the well-known result from communication theory that the lowest detection false alarm rate results from using a threshold measurement to determine the presence or absence of a signal. The probability of detection is a function only of the signal threshold and the SNR. Another reason for using SNR with imaging systems is that there is ample evidence that the lower limit of visual detectability is set by a thresholding process. The value of sensor SNR when the signal is saturated is called the *dynamic range*.

The complete imaging system is characterized by the spatial fidelity with which the scene image is converted into video signals. This is the subject of quality measures loosely described as "resolution." Angular spatial resolving power is ordinarily dominated by the angular projection through the optics of the smallest independently responding (elemental) area of the detecting surface. Often the detecting surface is spatially delineated in a periodic two-dimensional pattern to form an array of sensing elements.

The linear dimension of such a detecting element is projected by the system optics into a linear angle in the scene space. That angle is the ratio of the cell dimension to the optics effective focal length. Since the detector cell cannot determine where within that angle a source originates, the angular resolution of the system is of the order of the size of the cell's angular projection.

1.1.2 Characteristics

Figure 1.2 illustrates the major features of energy transfer from the scene to the imager. Scene radiation along a line of sight (LOS) from the source to the

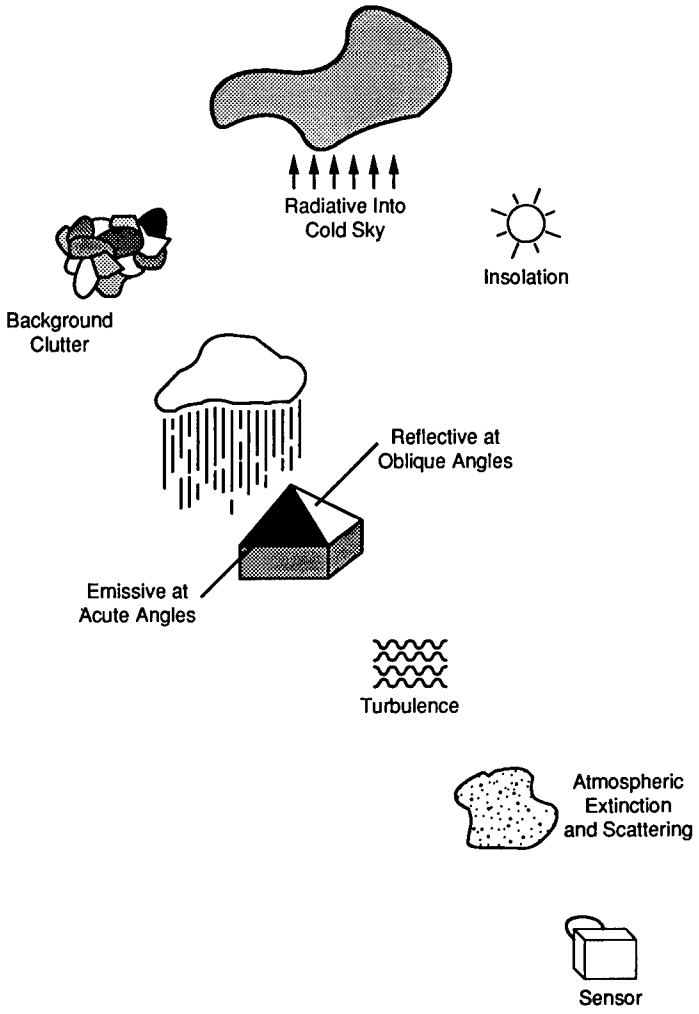


Fig. 1.2 EO imaging characteristics.

imager system arises from four mechanisms: self-emission, transmission of emissions from objects behind the source, reflection of remote emissions from objects in front of the source, and scattering and/or diffraction of all of these from outside the LOS into the LOS by the intervening atmosphere. For example, the LOS to a point in a cloud may contain emitted radiation from the cloud's internal energy, transmitted energy from sources behind it along the LOS, scattering from sources along other lines of sight, and reflections from sources external to the cloud.

All these phenomena are angularly dependent, so that the appearance of a differential element of a source's surface may depend on the viewing angle relative to the surface normal and on the angles of external sources relative to the surface normal. Thus a particular surface may take on quite different

appearances depending on its orientation to external sources and to the viewer. In the visible, a thundercloud in sunlight may appear white toward the sun and black underneath, and in the infrared an aircraft's bare aluminum skin on a clear night may appear warm on the side and cold on the top. The infrared contrast of natural terrain features is strongly dependent on the strength of insolation, and thermal washouts may occur twice per day.

While it is possible, given enough measurements, to describe the appearance of any source goniometrically, in practice the description usually is reduced to the specification of a single "apparent" quantity, which lumps the effects of all of the mediating factors together. For example, in the visible, some measure of the apparent contrast of a source relative to its background is usually specified. In the infrared, the apparent temperature difference of a source relative to the apparent temperature of its background is often specified.

The radiant characterization of a source relative to its background is called the *source signature*. Signature is composed of five elements: differences in intensity between source and background with insignificant spectral differences (contrast), differences in intensity with wavelength (color), differences in polarization, differences in the spatial correlation length and the intensity relative to surrounding scene sources (conspicuity), and time variations (motion and intensity modulation).

1.1.3 Impediments to Source Detection

Even though a source may manifest numerous differences from its surround, these differences do not necessarily propagate to the imager. The conspicuity of a source is reduced by atmospheric absorption, scattering, diffraction, and scintillation. It is also common, especially in infrared imaging and in long-range visible imaging, for the source signature to be superimposed on an uninformative and contrast-reducing background "pedestal."

This is a result, in the thermal case, of the lowest apparent temperature in the scene and, in the visible case, of atmospheric scattering. For example, in the 8- to 14-μm spectral band, the thermal radiation from a 300 K background is 65 times larger than the radiation difference due to a 1 K variation in that background. If a pedestal is not removed by signal processing, it consumes dynamic range, and reduces the displayed contrast and the detectability of the source.

In the absence of conspicuous amplitude differences, the distinctions between source and background may be the sharpness of the transition between them and the degree to which the dimensions of the target are distinguishable from those of typical background spatial variations. The source dimensions and edge sharpness do not necessarily propagate to the final displayed image with full fidelity.

Many phenomena result in blurring of details and the softening of transitions. Fundamental among these are optical diffraction, averaging of image detail by a finite-sized spatially integrating detector, and averaging of temporally coded (scanned or sampled) spatial variations due to finite electrical signal processing bandwidth. These phenomena may be described by integro-differential equations in space and/or time, but are more conveniently described in the complex frequency domain.

Another impediment to detection is the random fluctuation of noise. Noise is any spatial or temporal variation in the image of the scene that significantly interferes with detection and estimation of the source. It is often described as "unwanted signal." Noise arises from statistical variations in the spatial dimensions of the background features (clutter), temporal variations of the source relative to the background (movement and orientation), atmospheric effects (scintillation, transmission variations, and self-radiation variations), quantum statistical variations in the source and background emissions (photon noise), spatial discontinuities in signal detection (sampled-data "aliasing"), electronic noise in the detection and signal processing functions, and "optical" noise (glare, shading, ghosting).

Another limitation is that no imaging system produces imagery of a point in the scene all of the time. Some fraction of the time is always devoted to "overhead" control and processing functions, to optically directing a detector of small angular extent to cover a larger solid angle in space, or to readout discharge of detective elements. Since scene signals are always temporally integrated in some way to improve the SNR, these overhead functions reduce the time available for signal integration and therefore reduce signal sensitivity.

1.1.4 Fundamental Limits

Jamieson[10] examined the physical limits to passive spatially incoherent infrared sensing resulting from the quantum nature of light and from the uncertainty principle. He determined the number of independent measurements a sensor can make in the four-dimensional space of time, wavelength, and two angles. He assumed perfect optics, perfect detectors, the photon statistics noise limit, spatial incoherence, and temporal coherence at light frequencies.

The number of degrees of freedom is the total number of independent measurements a sensor can make. That is, it is the number of distinguishable cells into which the sensor may dissect the four-dimensional signal by spatial, temporal, or optical-temporal coherence techniques. The degrees of freedom N is given by

$$N = \frac{2A_o\Omega}{\lambda^2}\frac{2\Lambda\chi}{\lambda^2}2\Delta f T_f \ , \tag{1.1}$$

where

A_o = optical aperture area
Ω = total solid angular field of view
λ = wavelength
Λ = wavelength band used for discrimination by optical-temporal coherence techniques
χ = maximum interference length exploited by the optical-temporal coherence measurement
Δf = temporal bandwidth of the sensor
T_f = frame time required to cover Ω with the detectors used.

As an example, consider a staring imager with a unit cell solid angular dimension Ω_I operating at 100% temporal efficiency. Such an imager has a bandwidth Δf of $\Omega/(2\pi T\Omega_I)$, so that $2T\Delta f = \Omega/(\pi\Omega_I)$. Let the sensor have a

detector array with 640 columns of 480 elements each, consistent with a television format. Let the horizontal FOV be 0.04 rad and the vertical FOV be 0.03 rad so that Ω is approximately $640 \times 480\ \Omega_I$. Let the sensor operate at a frame rate of 30 Hz and work through a 20-cm-diameter optical aperture (0.0314-m^2 area) at a mean wavelength of 9 μm. Assuming that the sensor does not exploit temporal coherence, only the first and the third factors of Eq. (1.1) are used, and N has the value

$$N = \frac{2(0.0314\ \text{m}^2)(0.0012\ \text{sr})}{(9 \times 10^{-6}\ \text{m})^2} \frac{1}{\pi}(640 \times 480)$$

$$= 9.1 \times 10^{10} \text{ degrees of freedom in space and time .}$$

Jamieson also indicated an approach to determining the number of distinguishable levels within the distinguishable cells based on the fundamental limit due to statistical photon flux variations.

1.1.5 Idealized Modeling

For first-order analyses such as those presented in this chapter, simplifying assumptions are necessary to reduce the requirements for input data and to reduce the computational burden. Light from the scene is assumed to be polychromatic with sufficient spectral width to be spatially incoherent. Sources and backgrounds are assumed to be graybody radiators characterizable by equivalent blackbody temperatures. Color is assumed to be uninformative.

Polarization is assumed to be random. Signal processing is assumed to be linear for the signal amplitudes of interest. Noise is assumed to have a Gaussian amplitude distribution. These assumptions are routinely violated in the real world to some degree, but usually not to such an extent that they seriously distort the following models.

1.2 PHOTON COLLECTION AND SIGNAL-TO-NOISE RATIO

Noise in imagery masks detection of changes in the scene optical characteristics of interest. To characterize an imager's capability to sense at a distance, it is useful to have a single summary measure roughly indicative of its ability to detect signal variations. The simplest measure of noisiness is the value of the signal change, which equals the root mean square (rms) value of the broadband noise at the point of detection. Such a measure is referred to as a *sensitivity,* analogous to similar measures in communication and radar.

For example, noise equivalent power (NEP) is the incident power at a detector's surface that produces a unity ratio of rms signal to rms noise in the detector output. Noise equivalent irradiance (NEI) is the peak-to-peak irradiance differential at the optical entrance aperture of a system, which produces unity SNR at the output of a specified stage of the signal processing. NEI is useful for describing the capability of point source detection systems. The noise equivalent temperature difference (NETD) is the target temperature difference, which for a broad target produces a unity ratio of peak-to-peak signal to rms noise at a specified signal processing output.

The purpose of such measures is twofold. First, they communicate in an approximate way how noisy the instrument is relative to the signal amplitudes to be detected. Second, they appear in more sophisticated measures of device performance, such as false alarm rate or probability of recognition, which take into account the effects of subsequent signal processing.

1.2.1 Noise Characterization

Noise is characterized by the probability distribution of its amplitude and by its correlation lengths in space and in time, or by the equivalent temporal and spatial power spectral densities in the frequency domain. The theoretical limit to noise in imaging systems is the random absorption rate of photons by the detector. This is called *generation noise* because it is manifested as a random rate of signal carrier generation due to the random arrival rate of photons at the detector. The statistics of generation noise are assumed to follow those of photon noise.

For sufficiently high photon fluxes of monochromatic light temporally uncorrelated at light frequencies, the discrete photon amplitude statistics are assumed to be Poisson distributed. The variance is given by

$$\sigma_Q^2 = \overline{Q_q} \ , \tag{1.2}$$

where $\overline{Q_q}$ is the average value of Q_q, the random variable giving the number of photons absorbed by a detector within the smallest independent (correlated) time period of detection. This collection time is determined by a specific integration time or by time constants inherent in the detection process.

The discrete Poisson distribution is assumed for high fluxes to tend toward a continuous Gaussian distribution. For Gaussian statistics the probability that the random variable Q_q takes the value ξ is

$$P(\xi) = \frac{1}{\sigma_Q \sqrt{2\pi}} \exp -(\xi - \overline{Q_q})^2 / 2\sigma_Q^2 \ . \tag{1.3}$$

For the case of blackbody radiation (neither monochromatic nor temporally uncorrelated), the variance is not given by Eq. (1.2), but by the Bose-Einstein equation:

$$\sigma_Q^2 = \overline{Q_q} \left[\frac{\exp(hc/\lambda kT)}{\exp(hc/\lambda kT) - 1} \right] \ . \tag{1.4}$$

The term in brackets is ordinarily close to unity. The SNR is defined as the ratio of the mean number of photons counted to the square root of the variance of the count. Then the central result of photon detection theory is that the SNR in most cases is given by

$$\text{SNR} = \frac{\overline{Q_q}}{\sigma_q} = \sqrt{\overline{Q_q}} \ . \tag{1.5}$$

With all noise, care must be taken in calculating statistics from ensembles and applying them to a single member of the population, or vice versa. If the ergodic hypothesis does not hold, the expected value for one member may not equal the average value for the full ensemble. This is usually the case when describing systems with many detectors. Remember that a statistical description of one detector's behavior does not necessarily apply to the entire array.

The power spectral density (PSD) in the temporal frequency domain is defined as the Fourier transform of the autocorrelation function of the noise, as discussed in Chapter 2 of Dereniak and Crowe.[2] Equivalently, it can be calculated from the square of the modulus of the Fourier transform of a sufficiently long time sequence of adequately sampled noise. The temporal PSD of photon noise is flat ("white") except as modified by time constants in the detection mechanism. The power spectrum in the spatial frequency domain of a noise process that has a spatial variation is the multidimensional Fourier transform of the autocorrelation function of the noise.

The types of random noise encountered in EO imaging are the aforementioned generation noise, followed in relative importance by noises referred to as generation-recombination, excess ("$1/f$" or "flicker"), Johnson, tunneling, thermal, and pickup. The reader is referred to Dereniak and Crowe[2] or Karim[4] for detailed treatments of detection noise.

Generation noise is present in all detectors, but in some device structures there are also statistical fluctuations in the rate at which signal carriers recombine. This recombination noise usually has the same power spectrum as generation noise, and the two combine as the root-sum-square. The result is called *generation-recombination* noise, and its power spectrum is usually a factor of $\sqrt{2}$ higher than generation noise alone. The existence of recombination noise depends on whether recombination takes place before or after detection. In a photodiode device, only generation noise is present. In a photoconductive device, both generation and recombination are present.

Excess or $1/f$ noise is due largely to variations in contact resistances, believed to be caused by chaotic variations in surface state potentials. Its power spectrum typically is proportional to the inverse of the electrical frequency f_t raised to a power of approximately 1. Thermal noise, also called Johnson noise, comes from random variation in the current in a resistive element due to interactions between carriers and phonons. Tunneling is a quantum effect that occurs in reverse-biased junctions. "Temperature" (as distinct from thermal) noise in photon detectors is due to gain and offset variations in detectors that are extremely sensitive to small variations in cryocooler coldtip temperature, such as extrinsic silicon devices.

Pickup noises are deterministic and are due to microphonics, electrical ground voltage variations, and electromagnetic interference. They tend to be spectral rather than white, and generally are not easily characterized because they depend on the spectral content of the external forcing function. For example, microphonic noise may appear in a cryostatically cooled detector due to the pulsating impact of the cooling liquid on the inner wall of the dewar. A similar effect may occur in a mechanically refrigerated device due to piston motion or gas pressure variations.

In addition to the above-mentioned noises, there are fixed pattern noises in arrays of detectors that cause the ergodic hypothesis to be violated. These are

"fixed" in the sense that their characteristics vary slowly with time, but they are distributed, not necessarily randomly, in space. Fixed pattern noises may be either signal independent or signal dependent. Examples of signal-independent noise are gain and offset variations from detector to detector in an array when the irradiance has small variations. The same array might exhibit signal-dependent noise due to nonlinearity when irradiated with large signal differences.

1.2.2 Radiant Energy Transfer

Radiometric units are discussed in detail in Dereniak and Crowe,[2] by Meyer-Arendt,[11] and by Zissis.[12] Table 1.2 defines the radiometric quantities used herein in their alternative energy and photon forms. A subscript is used to distinguish between the two, e for energy and q for quanta. The dominant type of detector in use now is the quantum detector, whose signal is proportional to the incident photon flux integrated over a fixed time interval. This makes it desirable to use the photon forms. However, the historical detector performance measures are in the energy forms, and detector performance measurements are usually still reported that way. Therefore, it is necessary to be adept at using both approaches.

The quantities being transported radiatively are the radiant energy Q_e in joules and the number of photons Q_q. These are functions of wavelength, time, and the radiation geometry, including surface area and solid angle. Consequently, either type of Q results from integration of a spectral density or its partial derivative over some specified radiation geometry, spectral bandpass, and time period. At a specific wavelength the two Q's are related by the photon energy hc/λ. The energy transfer can be a combination of emitted, reflected, or transmitted energies. The following paragraphs first summarize radiometric quantities for extended surfaces, then treat point sources.

The partial derivatives with respect to time of radiant energy and of photon number are, respectively, the radiant flux Φ_e in watts and the photon flux Φ_q in photons per second. These fluxes are the time rate of transport of energy or of photons integrated over a specified radiation transfer geometry in a specified spectral bandpass. Fluxes are functions of the angles relative to the normals to the differential emitting and receiving surface areas being described.

For an emitting surface, the partial derivatives of emitted flux with respect to surface area are the radiant exitance M_e in watts per square meter and the photon exitance M_q in photons per second per square meter. For an absorbing surface, the partial derivatives of incident flux with respect to surface area

Table 1.2 Definition of Radiometric Quantities

Quantity	Defintion	Energy Symbol	Photon Symbol
Radiant energy Q	Total radiated energy	Q_e (J)	Q_q (photons)
Radiant flux Φ	$\partial Q/\partial t$	Φ_e (W)	Φ_q (photons s^{-1})
Radiant exitance M	$\partial\Phi/\partial A$	M_e (W m^{-2})	M_q (photons s^{-1} m^{-2})
Irradiance E	$\partial\Phi/\partial A$	E_e (W m^{-2})	E_q (photons s^{-1} m^{-2})
Radiance L	$\dfrac{\partial^2\Phi}{(\partial A\ \cos\theta)\partial\Omega}$	L_e (W m^{-2} sr^{-1})	L_q (photons s^{-1} m^{-2} sr^{-1})
Radiant intensity I	$\partial\Phi/\partial\Omega$	I_e (W sr^{-1})	I_q (photons s^{-1} sr^{-1})

are the energy irradiance E_e and the photon irradiance E_q in the same units as the corresponding exitances.

The partial derivatives of the exitance with respect to the solid angle through which the source radiates, at a specified angle θ to the surface normal, are the energy radiance L_e in watts per square meter per steradian and the photon radiance L_q in photons per second per square meter per steradian.

Radiation from a point source is described by the partial derivative of the fluxes with respect to the solid angle through which the fluxes are radiated. These quantities are the energy radiant intensity I_e in watts per steradian and the photon radiant intensity I_q in photons per second per steradian. A point source is not necessarily a mathematical point. It could also be an object, which from the perspective of the viewing device has an angular subtense much smaller than the spatial resolution of the device.

Each of the above quantities is derived by integration over a specified wavelength band of the associated spectral concentration, or wavelength density, of the quantity. Spectral concentration is denoted by an additional subscript λ, for example, as in $M_{q\lambda}$, which is called the *spectral photon exitance* and has units of photons per square meter per second per micrometer.

For first-order analyses, we usually assume that Lambert's cosine law holds, meaning that extended sources are diffuse radiators with their radiance being proportional to the cosine of the angle at which they are viewed, so that integration of L over a hemisphere yields

$$L = M/\pi \ . \tag{1.6}$$

For blackbody radiation, the fundamental radiometric quantities from which all others are derived are the spectral radiant exitances

$$M_{e\lambda} = \frac{2\pi hc^2}{\lambda^5[\exp(hc/\lambda kT) - 1]} = \frac{c_1}{\lambda^5[\exp(c_2/\lambda T) - 1]} \tag{1.7}$$

and

$$M_{q\lambda} = \frac{\lambda}{hc} M_{e\lambda} = \frac{2\pi c}{\lambda^4[\exp(hc/\lambda kT) - 1]} = \frac{c_3}{\lambda^4[\exp(c_2/\lambda T) - 1]} \ , \tag{1.8}$$

where the blackbody radiation constants are as given in Table 1.3.

Integration of the spectral radiant exitances over all wavelengths gives the Stefan-Boltzmann law for the total radiant exitances into a hemisphere:

$$M_e = \frac{\pi^5 k^4 T^4}{45c^2h^3} = \sigma_e T^4 \tag{1.9}$$

and

$$M_q = \frac{2\pi k^3 T^3}{c^2 h^3} = \sigma_q T^3 \tag{1.10}$$

with the constants as given in Table 1.3.

Table 1.3 Blackbody Radiation Constants

h = Planck's constant = 6.626176×10^{-34} (W s^2)
c = vacuum speed of light = 2.997925×10^{8} (m s^{-1})
k = Boltzmann's constant = 1.380662×10^{-23} (J K^{-1})
c_1 = first radiation constant = 3.741832×10^{8} (W μm^4 m^{-2})
c_2 = second radiation constant = 1.438786×10^{4} (μm K)
c_3 = third radiation constant = 1.883652×10^{27} (μm^3 s^{-1} m^{-2})
σ_e = energy Stefan-Boltzmann constant = 5.6703×10^{-8} (W m^{-2} K^{-4})
σ_q = photon Stefan-Boltzmann constant = 5.5204×10^{23} (s^{-3} m^{-6} K^{-9})

1.2.3 Extended Source Focal Plane Flux Equation

Figure 1.3 depicts the geometry assumed for the calculation of irradiance on the image plane from a remote extended resolved source. The extended surface is assumed to be planar, to be normal to the optical axis, and to follow Lambert's law. The radiating surface, the receiving aperture, and the receiving detector are assumed to lie in parallel planes. Appropriate cosine factors may be applied if this is not the case due to target orientation, sensor panning, sensor scanning, or off-axis detector location. In the worst case this results in an irradiance dependence on the cosine to the fourth power, as described by Dereniak and Crowe[2] and Karim.[4]

In all of the following equations, use of consistent units has been assumed. Bear in mind that the use of convenient units will require the addition to these equations of appropriate unit conversion constants. The receiving aperture has an area A_o, is at a range R from the source, and forms an image at an effective distance f behind it. For sufficiently large R, small-angle approximations may be made and the aperture receives source radiation within a solid angle in steradians of A_o/R^2 relative to the source.

The source signature is described by its radiant exitance M, in either the energy or the photon form. Lambert's law is assumed to hold so that the source radiance L_λ equals M_λ/π. Unless explicitly indicated otherwise, the source dimensions are assumed to be much larger than the dimensions subtended by the projected detector angular subtense. Then the irradiance on the detector is not influenced by optical blurring or detector location in the focal plane relative to the source image.

The detector is assumed to be rectangular with linear dimensions a and b in the focal plane. These project through the optics of focal length f with angular dimensions α and β in radians, given respectively by a/f and b/f. The projected detector solid angle in radians is ω, which equals the product $\alpha\beta$. The detector collects radiance from a source area of $\alpha\beta R^2$ corresponding to the area projection of the detector solid angular subtense at the range R.

The intervening atmosphere transmits the source radiance by the factor $\tau_a(\lambda)$ and is assumed to have an insignificant signature itself. If this assumption is untenable, the more general approach of Findlay and Cutten[13] must be used. In most of the subsequent radiation equations, the symbols are given from left to right in the order in which they appear in the imaging process to aid in comprehension.

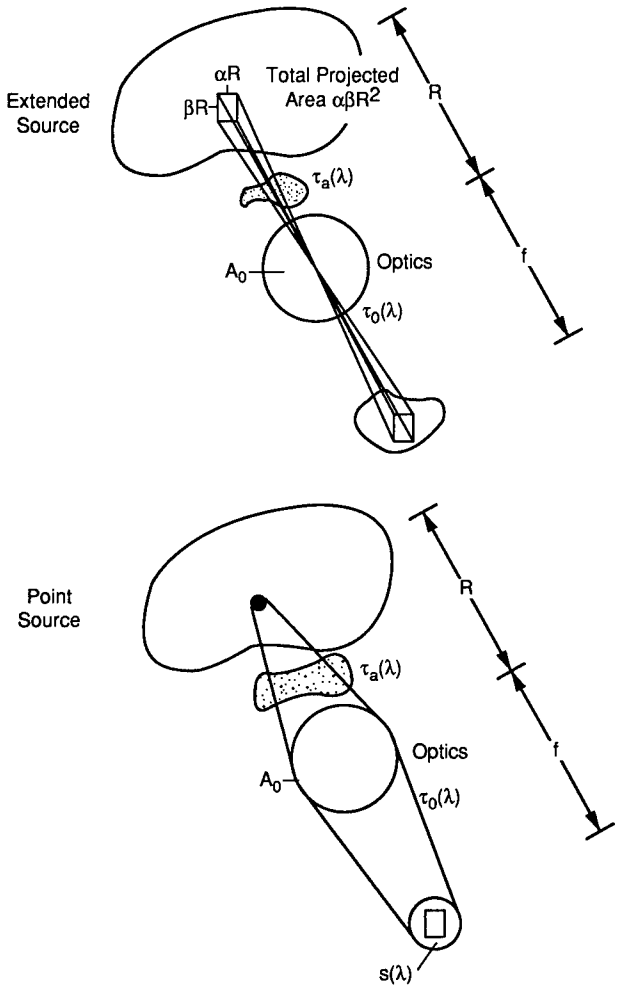


Fig. 1.3 Optical energy collection.

The spectral flux $\Phi_{\lambda o}$ on the optical collecting aperture in either energy or photon units is then

$$\Phi_{\lambda o} = \alpha\beta R^2 L_\lambda \tau_a(\lambda)\frac{A_o}{R^2} = \alpha\beta L_\lambda \tau_a(\lambda) A_o \ . \tag{1.11}$$

Using the small-angle approximation and allowing for the transmission $\tau_o(\lambda)$ of the optics, the spectral flux of the radiation reaching the detector element in either photon or energy units is

$$\Phi_{\lambda d} = \alpha\beta L_\lambda \tau_a(\lambda) A_o \tau_o(\lambda) \ . \tag{1.12}$$

The small-angle approximation is made to simplify the sensitivity equations for the fundamental photon noise limit, and is sufficiently accurate for most

purposes for $f/\#$'s greater than 1.5. See Chapter 2 of Karim[4] for the exact expression.

If the flux is desired in watts, $L_{e\lambda}$ will be used. If the flux is desired in photons per second, $L_{q\lambda}$ will be used. If the irradiance on the detector focal plane is desired, the above expression may be divided by the detector area. For a circular optical aperture you will often see the alternative form

$$\alpha\beta A_o = \frac{ab}{f^2}\frac{\pi D_o^2}{4} = \frac{\pi A_d}{4(f/\#)^2} , \tag{1.13}$$

resulting in this expression for the spectral flux on the detector:

$$\Phi_{\lambda d} = \frac{\pi A_d}{4(f/\#)^2} L_\lambda \tau_a(\lambda)\tau_o(\lambda) . \tag{1.14}$$

The exact expression avoiding the approximation made in Eq. (1.12) has the factor $[1 + 4(f/\#)^2]$ in place of the factor $[4(f/\#)^2]$. Note that the only dependence on range here is implicit in $\tau_a(\lambda)$. Up to this point there is no difference between the power and the photon formulations for an extended source. They diverge in the derivation of the detector output SNR in Sec. 1.2.5.1.

1.2.4 Point Source Irradiance Equations

Referring to Fig. 1.3, the point source is assumed to be radiating uniformly in all directions. Then regardless of its orientation to the LOS, it may be characterized relative to its background by a constant difference in spectral radiant intensity ΔI_λ in watts per steradian per micrometer, or in photons per second per steradian per micrometer.

The change in spectral flux received by the optical aperture relative to the background flux is

$$\Delta\Phi_{\lambda o} = \Delta I_\lambda \tau_a(\lambda)\frac{A_o}{R^2} . \tag{1.15}$$

Making the small-angle approximation as before, the change in spectral flux on the detector is

$$\Delta\Phi_{\lambda d} = \Delta I_\lambda \tau_a(\lambda)\frac{A_o}{R^2}\tau_o(\lambda)S_o(\lambda) \tag{1.16}$$

where the optical "shape" factor $S_o(\lambda)$ has a value between zero and one. The shape factor accounts for two effects. First, the optics blur causes some of the point source flux to fall off the detector. Second, the center of the point source image may not fall on the center of the detector due to the registration of the array structure.

The dependence of S_o on wavelength allows for optical chromatic aberrations. See Barhydt[14] for a more detailed discussion of the components of the shape factor, referred to by him as "efficiencies." The need for a shape factor

opens the detector flux description up to a statistical treatment with a minimum value, a mean value, a variance, and a maximum value. These may be derived by considering the possible range of misalignments of the detector relative to the source image.

1.2.5 Detector Output Signal-to-Noise Ratio

The equation for the SNR of the detector output depends on whether the source is unresolved or extended, and on whether an energy form or a photon form is desired. The energy derivation proceeds assuming that the detector is a power-to-voltage transducer. The detector could as easily be assumed to be a power-to-current transducer, in which case references to voltages would be replaced with currents. The photon derivation proceeds assuming that the detector is an integrating photon counter.

1.2.5.1 Extended Source SNR. Using Eq. (1.12) for the spectral flux on a detector from an extended source, the spectral fluxes from a target and its background are approximately

$$\Phi_{\lambda dT} = \alpha\beta L_{\lambda T}\tau_a(\lambda)A_o\tau_o(\lambda) \tag{1.17}$$

and

$$\Phi_{\lambda dB} = \alpha\beta L_{\lambda B}\tau_a(\lambda)A_o\tau_o(\lambda) \ , \tag{1.18}$$

where the subscripts T and B refer to target and background, respectively. The derivations of the energy and photon extended source SNR equations diverge at this point as follows.

Energy Form. In the energy formulation, the detector converts power to voltage with a spectral transduction factor called the spectral responsivity $\mathcal{R}(\lambda)$ defined by

$$\mathcal{R}(\lambda) = \frac{V_s(\lambda)}{E_{\lambda d}A_d} \quad [\text{V/W}] \ , \tag{1.19}$$

where $E_{\lambda d}$ is the rms value of the fundamental temporal frequency of the radiant test input, and $V_s(\lambda)$ is the rms value of the fundamental temporal frequency of the resulting detector output signal. The specific temporal frequency of the measurement and the temperature of the source often are made explicit inside the parentheses of $\mathcal{R}(\lambda)$. Introducing the subscript e on the radiance, for either target or background the spectral signal voltage is

$$V(\lambda) = \alpha\beta L_{e\lambda}\tau_a(\lambda)A_o\tau_o(\lambda)\mathcal{R}(\lambda) \ . \tag{1.20}$$

Then the signal voltage difference in the transition from background to target is

$$\Delta V(\lambda) = \alpha\beta(L_{e\lambda T} - L_{e\lambda B})\tau_a(\lambda)A_o\tau_o(\lambda)\mathcal{R}(\lambda) \ . \tag{1.21}$$

It is assumed that the noise voltage is sufficiently independent of the difference between target and background that we may write the spectral signal-to-noise ratio as

$$\mathrm{SNR}(\lambda) = \frac{\Delta V(\lambda)}{V^n} = \frac{\alpha\beta A_o}{V^n}(L_{e\lambda T} - L_{e\lambda B})\tau_a(\lambda)\tau_o(\lambda)\mathcal{R}(\lambda) \ . \tag{1.22}$$

When the target-to-background difference is small, Eq. (1.22) is valid. Otherwise an expression for V_n incorporating $L_{e\lambda T}$ must be used.

The SNR is obtained by spectral integration:

$$\mathrm{SNR} = \frac{\alpha\beta A_o}{V^n}\int_0^\infty (L_{e\lambda T} - L_{e\lambda B})\tau_a(\lambda)\tau_o(\lambda)\mathcal{R}(\lambda)\, d\lambda \ . \tag{1.23}$$

The ratio $V_n/\mathcal{R}(\lambda)$ is defined as the spectral noise equivalent power, NEP(λ). This quantity is the change in energy flux at a specified wavelength incident on the detector, which produces a unity SNR in the detector output. Equation (1.23) can be rewritten using NEP(λ) as

$$\mathrm{SNR} = \alpha\beta A_o\int_0^\infty (L_{e\lambda T} - L_{e\lambda B})\tau_a(\lambda)\tau_o(\lambda)\frac{1}{\mathrm{NEP}(\lambda)}\, d\lambda \ . \tag{1.24}$$

NEP(λ) is no longer reported as a detector performance measure. It was replaced by its reciprocal, $\mathcal{R}(\lambda)/V_n$, called the *spectral detectivity* $D(\lambda)$, in order to report a larger number when the detector SNR is higher. Detectivity, in turn, was replaced by a normalized quantity called specific detectivity, defined by

$$D^*(\lambda) = D(\lambda)(A_d\Delta f)^{1/2} = \frac{(A_d\Delta f)^{1/2}}{\mathrm{NEP}(\lambda)} = \frac{(A_d\Delta f)^{1/2}\mathcal{R}(\lambda)}{V^n} \ , \tag{1.25}$$

where Δf is the noise equivalent bandwidth and D^* is reported in units of $\mathrm{cm}\sqrt{\mathrm{Hz}}/\mathrm{W}$, so care must be taken in equations where the other parameters are in MKS units.

Specific detectivity bears a similar relationship to detectivity as that of specific heat to heat capacity. Specific detectivity is a detector material property, whereas detectivity describes the result of using that material property to fabricate a particular detector. The factor $(A_d\Delta f)^{1/2}$ is the normalization, analogous to dividing the heat absorbed by a sample by its mass and temperature rise to obtain specific heat. It accounts for the observation that in most detectors the SNR is inversely proportional to the square root of the product of detector area and electrical bandwidth. Putting all this together, a commonly seen form for the SNR is

$$\mathrm{SNR} = \frac{\alpha\beta A_o}{(A_d\Delta f)^{1/2}}\int_0^\infty (L_{e\lambda T} - L_{e\lambda B})\tau_a(\lambda)\tau_o(\lambda)D^*(\lambda)\, d\lambda \ . \tag{1.26}$$

At this point blackbody radiation is introduced for the first time in these derivations by assuming that the spectral radiance is a function of temperature.

For cases where detection of small temperature differences against a uniform background at temperature T_B is of interest, the approximation is made that

$$L_{e\lambda T} - L_{e\lambda B} = \Delta T \frac{\partial L_{e\lambda}}{\partial T}(T_B) \ , \tag{1.27}$$

where the T_B in parentheses indicates evaluation of the partial derivative at that temperature. The SNR equation then becomes

$$\text{SNR} = \frac{\alpha\beta A_o}{(A_d \Delta f)^{1/2}} \Delta T \int_0^\infty \frac{\partial L_{e\lambda}}{\partial T}(T_B)\tau_a(\lambda)\tau_o(\lambda)D^*(\lambda)\, d\lambda \ . \tag{1.28}$$

Longshore, Raimondi, and Lumpkin[15] made the definition

$$M^* = \pi \int_0^\infty \frac{\partial L_{e\lambda}}{\partial T}(T_B)\tau_a(\lambda)\tau_o(\lambda)D^*(\lambda)\, d\lambda \tag{1.29}$$

and proposed its use as a figure of merit for spectral band selection.

The SNR equation is sometimes further developed by incorporating the theoretically ideal form of spectral detectivity for a bandgap semiconductor detector,

$$D^*(\lambda) = \frac{\lambda}{\lambda_p} D^*(\lambda_p) \ , \qquad \lambda \leq \lambda_p \ , \tag{1.30}$$

where λ_p is the theoretically sharp "cutoff" wavelength corresponding to the bandgap. The formulation of D^* in energy units, with the value of D^* increasing with wavelength, suggests the erroneous notion that longer wavelengths are more efficacious from a SNR viewpoint than are shorter wavelengths. In fact, the response of a photon detector at any wavelength is proportional to the photon flux at that wavelength, not to the energy flux.

Photon Form. The photon counting SNR is derived from Eqs. (1.17) and (1.18) by finding the change Δn_s in the number of signal carriers due to the radiance difference between target and background. That change is then divided by the number of so-called "noise carriers," the square root of the variance in the number of total carriers from all sources, including noise sources. The variance is assumed to equal the mean value $\overline{n_T}$ of the total number of carriers, as in Eq. (1.2).

Paralleling Eq. (1.21), the time rate of generation of the spectral density of the number of signal carriers corresponding to the difference between signal and background is approximately

$$\frac{d\Delta n_{\lambda s}}{dt} = \alpha\beta A_o (L_{q\lambda T} - L_{q\lambda B})\tau_a(\lambda)\tau_o(\lambda)\eta_q(\lambda) \ , \tag{1.31}$$

where $\eta_q(\lambda)$ is the quantum efficiency as a function of wavelength, $0 \leq \eta_q(\lambda) \leq 1$. The quantum efficiency is the efficiency in the conversion of photons to signal carriers, assuming the device is not an avalanche structure. Note the use of the subscript q on the photon spectral radiances.

Assuming that the detector is "on" and collecting photons for a fixed integration time τ_i before it is "read out," the change in the number of signal carriers in an integration time, after integration over all wavelengths, is

$$\Delta n_s = \alpha\beta A_o \tau_i \int_0^\infty (L_{q\lambda T} - L_q\lambda_B)\tau_a(\lambda)\tau_o(\lambda)\eta_q(\lambda)\, d\lambda \ . \tag{1.32}$$

Then the SNR is

$$\text{SNR} = \frac{\Delta n_s}{\sqrt{\overline{n_T}}} \ . \tag{1.33}$$

The total number of carriers arises from background, signal, and signal-independent noise sources ("dark current"). The components include the noise sources listed in Sec. 1.2.1. It is dependent on the device structure, and a general treatment is beyond the scope of this chapter. However, the fundamental limit due to photon noise is derivable as follows. If the difference between the target and background photon radiances is small, and if the generation of noise carriers is wholly due to background flux, then $n_T = n_B$, the number of carriers generated by background flux. If the detector is perfectly cold-shielded so that the only background flux it receives arrives from the scene through the optics, then following the approach of Borg, Davis, and Thompson[16] the variance of the number of carriers is

$$\sigma_n^2 = \overline{n_T} = \alpha\beta A_o \tau_i \int_0^\infty L_{q\lambda}(T_B)\tau_a(\lambda)\tau_o(\lambda)\eta_q(\lambda)\, d\lambda \ . \tag{1.34}$$

Making the assumption that the detector background is uniform at a temperature T_B, the SNR is

$$\text{SNR} = \frac{\Delta n_s}{\sqrt{\overline{n_B}}} = \frac{\alpha\beta A_o \tau_i \int_0^\infty (L_{q\lambda T} - L_{q\lambda B})\tau_a(\lambda)\tau_o(\lambda)\eta_q(\lambda)\, d\lambda}{\left[\alpha\beta A_o \tau_i \int_0^\infty L_{q\lambda}(T_B)\eta_q(\lambda)\, d\lambda\right]^{1/2}} \ . \tag{1.35}$$

This is a highly idealized case. In general, it is necessary to model the detector fluxes originating from the atmosphere, the optical elements, and the structures around the optics and the detector.

1.2.5.2 Point Source SNR. Recall from Eq. (1.16) that the change in flux arriving at a detector due to the appearance of a point source against a uniform background, for either energy or photons, is approximately

$$\Delta\Phi_{\lambda d} = \Delta I_\lambda \tau_a(\lambda) \frac{A_o}{R^2} \tau_o(\lambda) S_o(\lambda) \ .$$

Energy Form. Paralleling the derivation of the energy form from Sec. 1.2.5.1, the point source energy SNR is derived by writing the equation for the spectral change in signal voltage $\Delta V_s(\lambda)$, dividing it by the noise voltage V_n, and integrating over all wavelengths. The result is

$$\text{SNR} = \frac{A_o}{R^2} \int_0^\infty \Delta I_{e\lambda} \tau_a(\lambda) \tau_o(\lambda) S_o(\lambda) \frac{\mathcal{R}(\lambda)}{V_n} \, d\lambda \ . \tag{1.36}$$

As before, $\text{NEP}(\lambda) = V_n/\mathcal{R}(\lambda)$, so

$$\text{SNR} = \frac{A_o}{R^2} \int_0^\infty \Delta I_{e\lambda} \tau_a(\lambda) \tau_o(\lambda) S_o(\lambda) / \text{NEP}(\lambda) \, d\lambda \ . \tag{1.37}$$

Photon Form. Paralleling the derivation of the photon form in Sec. 1.2.5.1, the point source photon SNR is derived by writing the equation for the spectral change in the rate of generation of signal carriers $d\Delta n_s/dt$, dividing it by the square root of the total number of carriers from all sources, and integrating over all wavelengths. The result is

$$\text{SNR} = \frac{A_o \tau_i}{R^2 \sqrt{n_T}} \int_0^\infty \Delta I_{q\lambda} \tau_a(\lambda) \tau_o(\lambda) S_o(\lambda) \eta_q(\lambda) \, d\lambda \ . \tag{1.38}$$

The limiting case is given by the background-limited infrared photodetection (BLIP) condition so that the total number of carriers is given by Eq. (1.34), and the theoretical limit to SNR for a small signal relative to the background is

$$\text{SNR}_{\text{BLIP}} = \frac{\dfrac{A_o \tau_i}{R^2} \displaystyle\int_0^\infty \Delta I_{q\lambda} \tau_a(\lambda) \tau_o(\lambda) S_o(\lambda) \eta_q(\lambda) \, d\lambda}{\left[A_o \alpha \beta \tau_i \displaystyle\int_0^\infty L_{q\lambda}(T_B) \tau_a(\lambda) \tau_o(\lambda) \eta_q(\lambda) \, d\lambda \right]^{1/2}} \ . \tag{1.39}$$

1.2.6 Sensitivity

Sensitivity as used here is defined as the value of the change in a distinguishing source radiant characteristic that equals the rms value of the noise at some specified point in the imaging process. Therefore, sensitivities are called *noise equivalent* parameters. In thermal imaging, a commonly used noise equivalent parameter is the noise equivalent temperature difference (NETD) in the amplified detector output. In threat warning, a common parameter is the noise equivalent irradiance (NEI) on the collecting aperture.

1.2.6.1 NETD. NETD is a laboratory measure defined[5] with the assumptions that the measurement sources are blackbodies and that the atmospheric transmission is unity. It is the target ΔT against a uniform background at T_B

that produces a unity ratio of peak signal to rms noise when measured at some specified point in the electronic signal processing. The target is sized sufficiently large relative to the system resolution such that the system frequency response does not significantly reduce the signal amplitude.

NETD may be specified as a minimum value to be achieved in all detector channels or as a channel average. To be meaningful, the background temperature, the spectral band, the exact point of measurement in the video chain, the method of noise estimation, and the extent of any external video filtering must all be specified. NETD is measured, as described by Driggers et al.,[17] by the ratio of the noise voltage to the slope of the curve of signal voltage ΔV versus ΔT for small ΔT:

$$\text{NETD} = \frac{V_n}{\Delta V/\Delta T} \,. \tag{1.40}$$

This is so defined because the signal transduction curve may have an offset from the origin due to thermal offsets in the test equipment, causing a single point signal measurement to be in error. They also point out that reporting the NETD of multiple channels as a single composite number by taking the rms value of all the channel NETDs is more sensitive to the effect of a few very noisy channels than is reporting the average channel NETD.

Even though NETD is a simplistic performance parameter, easily misinterpreted or misrepresented, it is still useful because it is an absolute measure of low-frequency SNR. As a summary measure its order of magnitude is indicative of system performance, and the equations for more complicated performance measures can often be cast in forms that include NETD as a convenient factor.

Energy Form. The energy form of the NETD equation is derived from Eq. (1.28) by setting the $\tau_a(\lambda)$ term to unity, setting SNR equal to unity, solving for ΔT, and identifying the result as NETD:

$$\text{NETD} = \Delta T(\text{SNR} = 1) = \frac{(A_d \Delta f)^{1/2}}{\alpha\beta \int_0^\infty \frac{\partial L_{e\lambda}}{\partial T}(T_B)\tau_o(\lambda)D^*(\lambda)\, d\lambda} \,. \tag{1.41}$$

For the case of a circular aperture and using Eq. (1.13), you will often see Eq. (1.41) rewritten as

$$\text{NETD} = \frac{4(f/\#)^2\sqrt{\Delta f}}{\pi\sqrt{A_d}\int_0^\infty \frac{\partial L_{e\lambda}}{\partial T}(T_B)\tau_o(\lambda)D^*(\lambda)\, d\lambda} \,. \tag{1.42}$$

Recall that this is an approximation for large $f/\#$'s.

Equations (1.41) and (1.42) are applicable whenever the spectral form of D^* is known. Most often that is not the case, the D^* being reported by the manufacturer as either the $D^*(\lambda_p)$ or the $D^*(500)$. The latter is an effective quantity integrated over a specified spectral bandpass by measurement with a blackbody

at 500 K as the chopped source and a blackbody at 300 K as the chopping background. Note that this is not the same as the "blackbody" D^* defined by Dereniak and Crowe.[2]

A further complication is that a $D^*(300)$ often is cited because with its use the NETD equation may be written without a spectral integral. That form is derived by noting that Eq. (1.41) contains the following integral in its denominator:

$$\int_0^\infty \frac{\partial L_{e\lambda}}{\partial T}(T_B)\tau_o(\lambda)D^*(\lambda)\,d\lambda \ . \tag{1.43}$$

The first step in this conversion is to assume the background temperature is 300 K. Then the integral is replaced by measurable spectrally integrated quantities. An "effective" optical transmission τ_o is sifted out of the integral:

$$\tau_o = \frac{\displaystyle\int_0^\infty \frac{\partial L_{e\lambda}}{\partial T}(300\text{ K})\tau_o(\lambda)D^*(\lambda)\,d\lambda}{\displaystyle\int_0^\infty \frac{\partial L_{e\lambda}}{\partial T}(300\text{ K})D^*(\lambda)\,d\lambda} \ . \tag{1.44}$$

Then the original integral is rewritten as:

$$\int_0^\infty \frac{\partial L_{e\lambda}}{\partial T}(300\text{ K})\tau_o(\lambda)D^*(\lambda)\,d\lambda = \tau_o\int_0^\infty \frac{\partial L_{e\lambda}}{\partial T}(300\text{ K})D^*(\lambda)\,d\lambda \ . \tag{1.45}$$

Then the integral on the right side of Eq. (1.45) is used to sift out the $D^*(300)$ defined by:

$$D^*(300) = \frac{\displaystyle\int_0^\infty \frac{\partial L_{e\lambda}}{\partial T}(300\text{ K})D^*(\lambda)\,d\lambda}{\displaystyle\int_0^\infty \frac{\partial L_{e\lambda}}{\partial T}(300\text{ K})\,d\lambda} \ . \tag{1.46}$$

Assuming Lambertian behavior and rearranging Eq. (1.46) we get

$$\int_0^\infty \frac{\partial L_{e\lambda}}{\partial T}(300\text{ K})D^*(\lambda)\,d\lambda = D^*(300\text{ K})\frac{1}{\pi}\int_0^\infty \frac{\partial M_{e\lambda}}{\partial T}(300\text{ K})\,d\lambda \ . \tag{1.47}$$

Recognizing that the integral on the right side of Eq. (1.47) is the thermal derivative of the Stefan-Boltzmann law evaluated at 300 K,

$$\int_0^\infty \frac{\partial M_{e\lambda}}{\partial T}(300\text{ K})\,d\lambda = 4\sigma_e(300\text{ K})^3 \ , \tag{1.48}$$

we rewrite Eq. (1.47) as

$$\int_0^\infty \frac{\partial L_{e\lambda}}{\partial T}(300\text{ K})D^*(\lambda)\,d\lambda = \frac{1}{\pi}4\sigma_e(300\text{ K})^3 D^*(300)\ . \tag{1.49}$$

Then replacing integral (1.43) by the product $4\sigma_e(300\text{ K})^3\tau_o D^*(300)/\pi$, the NETD equation becomes:

$$\text{NETD} = \frac{\pi(A_d\Delta f)^{1/2}}{4\sigma_e(300\text{ K})^3\alpha\beta\tau_o D^*(300)}\ . \tag{1.50}$$

Again Eq. (1.13) may be employed to produce the simplest possible approximate NETD equation,

$$\text{NETD} = \frac{(f/\#)^2\sqrt{\Delta f}}{\sqrt{A_d}\sigma_e(300\text{ K})^3\tau_o D^*(300)}\ . \tag{1.51}$$

As a numerical example, consider a noninterlaced scanning FLIR with a 4:3 aspect ratio and a linear contiguous 480-element detector array. Let this FLIR use sampled integrating detectors at a 30-Hz frame rate with a scan efficiency of 75%, and let the number of samples per detector dwell time be 1.5, yielding 960 samples per scan line. Let the detector integration time equal the sample interval, so that the noise bandwidth $\Delta f = 1/(2\tau_i)$ = (30 Hz) (960 samples/line)/[2(0.75)] = 19.2 kHz.

Let the $f/\#$ be 2, the square root of the detector area be 0.0025 cm, the optical transmission be 0.75, and the $D^*(300)$ be 2×10^{10} cm$\sqrt{\text{Hz}}$/W. Then the NETD is calculated to be

$$\text{NETD} = \frac{2^2(19.2\text{ kHz})^{1/2}}{(0.0025\text{ cm})(5.67 \times 10^{-12}\text{ W cm}^{-2}\text{ K}^{-4})(300\text{ K})^3 \times (0.75)(2 \times 10^{10}\text{ cm}\sqrt{\text{Hz}}/\text{W})}$$

$$= 0.097\text{ K}\ .$$

With the NETD formulated in this way, the effective optical transmission τ_o can be measured by the insertion loss method with small signals and the detector of interest, which is exactly the definition of τ_o, and the $D^*(300)$ can be measured with small signals in a test set having the same spectral shape as $\tau_o(\lambda)$. In practice, the detector is usually measured by the difference between a 300 K source and a 500 K source, and the resulting $D^*(500\text{ K})$ is converted to the desired $D^*(300)$ by

$$D^*(300) = D^*(500)\frac{(500\text{ K})^4 - (300\text{ K})^4}{4(300\text{ K})^3} \times \frac{\int_0^\infty D^*(\lambda)\frac{\partial L_{e\lambda}}{\partial T}(300\text{ K})\,d\lambda}{\int_0^\infty D^*(\lambda)[L_{e\lambda}(500\text{ K}) - L_{e\lambda}(300\text{ K})]\,d\lambda}\ . \tag{1.52}$$

The effective quantities of Eqs. (1.44) and (1.46) were sifted out of their home integral in the order given so that when they are measured their product will have the same value as the integral. Therefore, they are not necessarily the same as the "apparent" or "average" quantities defined in Sec. 1.2.7, which are used to collect experimental data at the system level.

The fundamental limit to NETD is derived by substituting the D^* expression for BLIP performance. The BLIP condition is attained when the detector noise is dominated by the random absorption rate of background photon flux, for example, by generation noise in a photodiode or other charge separation structure, or by generation-recombination noise in a photoconductor.

In the photodiode case, the $D^*(\lambda)$ for the BLIP condition is given by

$$D^*_{\mathrm{BLIP}}(\lambda) = \frac{\lambda}{hc\sqrt{2}} \frac{\eta_q(\lambda)}{(E_{qB})^{1/2}}$$

$$= \frac{\lambda}{hc\sqrt{2}} \frac{\eta_q(\lambda)}{\left[\dfrac{\Omega_{cs}}{\pi}\displaystyle\int_0^\infty \eta_q(\lambda)\tau_{cf}(\lambda)E_{q\lambda B}\, d\lambda\right]^{1/2}}, \tag{1.53}$$

where E_{qB} is the spectrally integrated photon irradiance on the detector from a full hemispherical surround, $\eta_q(\lambda)$ is the spectral quantum efficiency, Ω_{cs} is the effective solid angle to which reception of background irradiance is limited by a cold shield (if any), and $\tau_{cf}(\lambda)$ is the transmission of a background-reducing cold filter, if used. For a photoconductor, in most cases, the factor $\sqrt{2}$ in the denominator of Eq. (1.53) is replaced by the factor 2 due to the presence of recombination noise having the same PSD as the generation noise.

Insight into the design implications of BLIP operation is achieved as follows. First define D^{**}_{BLIP} as the D^*_{BLIP} for an unshielded detector, that is, one having a cold shield solid angle of π. Then, since the background photon flux is proportional to the solid angle, and D^* is inversely proportional to the inverse square root of the solid angle,

$$D^*_{\mathrm{BLIP}} = (\pi/\Omega_{cs})^{1/2} D^{**}_{\mathrm{BLIP}} \,. \tag{1.54}$$

Then let the perfect cold shield solid angle conforming to the cone of flux converging on the detector be Ω_p. Next, define a cold shielding efficiency η_{cs} by the ratio of Ω_p to the actual cold shield solid angle Ω_{cs} so that:

$$\Omega_{cs} = \frac{\Omega_p}{\eta_{cs}} \,. \tag{1.55}$$

Then Eq. (1.41) may be rewritten for the BLIP condition per Laakmann[18] or Lloyd[5] as

$$\mathrm{NETD}_{\mathrm{BLIP}} = \frac{(A_d \Delta f)^{1/2}}{\alpha\beta\left(\dfrac{\pi\eta_{cs}}{\Omega_p}\right)^{1/2} \displaystyle\int_0^\infty \frac{\partial L_{e\lambda}}{\partial T}(T_B)\tau_o(\lambda)D^{**}_{\mathrm{BLIP}}(\lambda)\, d\lambda} \,. \tag{1.56}$$

For the common case where the cone of irradiance converging toward the detector is circular, and the optical $f/\#$ is slower than $f/2$, a satisfactory approximation to the perfect cold shield angle Ω_p is

$$\Omega_p = \frac{\pi}{(2f/\#)^2} = \frac{\pi D_o^2}{4f^2} \ . \tag{1.57}$$

Substitution of Eq. (1.57) into Eq. (1.56) followed by some manipulation yields

$$\mathrm{NETD_{BLIP}} = \frac{2f\sqrt{\Delta f}}{\pi\sqrt{A_d}D_o\sqrt{\eta_{cs}}\displaystyle\int_0^\infty \frac{\partial L_{e\lambda}}{\partial T}(T_B)\tau_o(\lambda)D^{**}_{\mathrm{BLIP}}(\lambda)\,d\lambda} \ , \tag{1.58}$$

showing the key result that BLIP NETD depends on D_o to the inverse first power. Barhydt[19] provides a formalism for describing nearly BLIP detectors and in a subsequent paper[20] gives a derivation similar to the above but for nearly BLIP conditions.

Photon Form. The photon form of the NETD equation is derived from Eqs. (1.32) and (1.33) by dropping the $\tau_a(\lambda)$ term, making the small-signal approximation as in Eq. (1.27), setting SNR equal to unity, solving for ΔT, and identifying the result as NETD:

$$\mathrm{NETD} = \frac{\sqrt{\overline{n_T}}}{\alpha\beta A_o\tau_i\displaystyle\int_0^\infty \frac{\partial L_{q\lambda}}{\partial T}(T_B)\tau_o(\lambda)\eta_q(\lambda)\,d\lambda} \ . \tag{1.59}$$

Using the result of Eq. (1.35) the BLIP limit may be rewritten as

$$\mathrm{NETD_{BLIP}} = \frac{\left[\displaystyle\int_0^\infty L_{q\lambda}(T_B)\eta_q(\lambda)\,d\lambda\right]^{1/2}}{(\alpha\beta A_o\tau_i)^{1/2}\displaystyle\int_0^\infty \frac{\partial L_{q\lambda}}{\partial T}(T_B)\tau_o(\lambda)\eta_q(\lambda)\,d\lambda} \ . \tag{1.60}$$

1.2.6.2 NEI. The noise equivalent irradiance (NEI) is the value of the irradiance in the plane of the optical entrance aperture that produces a unity SNR in the detector output. This is also called noise equivalent flux density (NEFD). It is used primarily as a figure of merit for point source detection systems. It is derived as follows. Let $E_{e\lambda o}$ be the irradiance from a point source on an aperture of area A_o. Accounting as in Sec. 1.2.4 for an optical transmission $\tau_o(\lambda)$ and a detector-to-image coupling shape factor $S_o(\lambda)$, the spectral SNR is

$$\begin{aligned}\mathrm{SNR}(\lambda) &= \frac{E_{e\lambda o}A_o\tau_o(\lambda)S_o(\lambda)}{\mathrm{NEP}(\lambda)} \\ &= \frac{E_{e\lambda o}A_o\tau_o(\lambda)S_o(\lambda)D^*(\lambda)}{(A_d\Delta f)^{1/2}} \ .\end{aligned} \tag{1.61}$$

Integrating over all wavelengths,

$$\text{SNR} = \frac{A_o}{(A_d \Delta f)^{1/2}} \int_0^\infty E_{e\lambda o} \tau_o(\lambda) S_o(\lambda) D^*(\lambda)\, d\lambda \ . \tag{1.62}$$

Defining the average in-band aperture irradiance as

$$E_{eo} = \frac{\displaystyle\int_0^\infty E_{e\lambda o} \tau_o(\lambda) S_o(\lambda) D^*(\lambda)\, d\lambda}{\displaystyle\int_0^\infty \tau_o(\lambda) S_o(\lambda) D^*(\lambda)\, d\lambda} \ , \tag{1.63}$$

Eq. (1.62) may be rewritten as

$$\text{SNR} = \frac{A_o E_{eo}}{(A_d \Delta f)^{1/2}} \int_0^\infty \tau_o(\lambda) S_o(\lambda) D^*(\lambda)\, d\lambda \ . \tag{1.64}$$

Setting the SNR equal to one, solving for E_{eo} and identifying it as NEI,

$$\text{NEI} = E_{eo}(\text{SNR} = 1) = \frac{(A_d \Delta f)^{1/2}}{A_o \displaystyle\int_0^\infty \tau_o(\lambda) S_o(\lambda) D^*(\lambda)\, d\lambda} \ . \tag{1.65}$$

Barhydt[14] shows that the extension of NEI to include the electrical signal processing requires the addition of a correction factor accounting for the peak-to-peak amplitude response of the electronics to the specific shape of the detector output waveform for the source. He called this the *electrical efficiency* or *pulse gain factor*. Barhydt[21,22] extended the simple formulation given above to the case of near-BLIP conditions.

It is common practice to describe a point source by its average in-band radiant intensity $\Delta I_{\Delta\lambda}$, relative to the background, in watts per steradian as measured by integration over a specified spectral band for specific atmospheric conditions:

$$\Delta I_{\Delta\lambda} = \frac{\displaystyle\int_{\Delta\lambda} \Delta I_\lambda \tau_a(\lambda, R)\, d\lambda}{\displaystyle\int_{\Delta\lambda} \tau_a(\lambda, R)\, d\lambda} \ , \tag{1.66}$$

where $\tau_a(\lambda, R)$ is the atmospheric transmission for a specific range R. Deriving the $\Delta I_{\Delta\lambda}$ for other ranges and atmospheric conditions obviously is subject to error given only an average $\Delta I_{\Delta\lambda}$ for another set of conditions as a starting point. However, this problem is unavoidable if a spectral shape is not specified. A photon form of NEI could be defined paralleling the derivation given above, but it would not be useful because point source data customarily are reported in energy units.

1.2.7 Apparent Quantities

Section 1.2.6.1 defined an "effective" or "blackbody" optical transmission useful for simplifying the form of the NETD equation. This is different from the usual definition of an "apparent" or "average" or "in-band" optical transmission useful in reporting the results of experiments. For example, it is useful to make generalizations such as the following:

- "The atmospheric transmission varied between 40 and 50%."
- "The optical transmission degraded by 1% per degree C."
- "The apparent delta-T was 10 degrees C."

Such simplifications are made by calculating these apparent quantities by the insertion loss method. The variation in effective quantities is what we infer from measurements made over a spectral band without benefit of spectral discrimination inside that band. Recalling Eq. (1.21) and integrating over all wavelengths, the change in signal determined by two measurements of sources at different temperatures is

$$\Delta V = \int_0^\infty \Delta V(\lambda)\, d\lambda = \alpha\beta A_o \int_0^\infty [L_{\lambda T}(T_T) - L_{\lambda B}(T_B)]\tau_a(\lambda)\tau_o(\lambda)\mathcal{R}(\lambda)\, d\lambda\,. \quad (1.67)$$

Then the apparent change in target radiance is defined as

$$\Delta L_a = \frac{\int_0^\infty [L_{\lambda T}(T_T) - L_{\lambda B}(T_B)]\tau_a(\lambda)\tau_o(\lambda)\mathcal{R}(\lambda)\, d\lambda}{\int_0^\infty \tau_a(\lambda)\tau_o(\lambda)\mathcal{R}(\lambda)\, d\lambda}\,, \quad (1.68)$$

the apparent atmospheric transmission as

$$\tau_{aa} = \frac{\int_0^\infty [L_{\lambda T}(T_T) - L_{\lambda B}(T_B)]\tau_a(\lambda)\tau_o(\lambda)\mathcal{R}(\lambda)\, d\lambda}{\int_0^\infty [L_{\lambda T}(T_T) - L_{\lambda B}(T_B)]\tau_o(\lambda)\mathcal{R}(\lambda)\, d\lambda}\,, \quad (1.69)$$

the apparent optical transmission as

$$\tau_{oa} = \frac{\int_0^\infty [L_{\lambda T}(T_T) - L_{\lambda B}(T_B)]\tau_a(\lambda)\tau_o(\lambda)\mathcal{R}(\lambda)\, d\lambda}{\int_0^\infty [L_{\lambda T}(T_T) - L_{\lambda B}(T_B)]\tau_a(\lambda)\mathcal{R}(\lambda)\, d\lambda}\,, \quad (1.70)$$

and an apparent responsivity as

$$\mathcal{R}_a = \frac{\int_0^\infty [L_{\lambda T}(T_T) - L_{\lambda B}(T_B)]\tau_a(\lambda)\tau_o(\lambda)\mathcal{R}(\lambda)\, d\lambda}{\int_0^\infty [L_{\lambda T}(T_T) - L_{\lambda B}(T_B)]\tau_a(\lambda)\tau_o(\lambda)\, d\lambda} . \tag{1.71}$$

In general, however,

$$V \neq \alpha\beta A_o \Delta L_a \tau_{aa} \tau_{oa} \mathcal{R}_a \tag{1.72}$$

because the apparent quantities were not successively sifted out of Eq. (1.68) like the effective quantities of Eqs. (1.44) and (1.46).

For $T_T - T_B$ on the order of a few degrees kelvin, an apparent ΔT can be calculated by defining the apparent change in signal with a small change in temperature as

$$\frac{\partial V}{\partial T} = \alpha\beta A_o \int_0^\infty \frac{\partial L_\lambda(T_B)}{\partial T} \tau_a(\lambda)\tau_o(\lambda)\mathcal{R}(\lambda)\, d\lambda . \tag{1.73}$$

Dividing Eq. (1.68) by Eq. (1.73) we get the apparent ΔT:

$$\Delta T_a = \frac{\Delta V}{\partial V/\partial T} = \frac{\int_0^\infty [L_{\lambda T}(T_T) - L_{\lambda B}(T_B)]\tau_a(\lambda)\tau_o(\lambda)\mathcal{R}(\lambda)\, d\lambda}{\int_0^\infty \frac{\partial L_\lambda(T_B)}{\partial T} \tau_a(\lambda)\tau_o(\lambda)\mathcal{R}(\lambda)\, d\lambda} . \tag{1.74}$$

Note that, in general, this is not necessarily the same as the actual blackbody temperature difference.

1.2.8 Spectral Band Selection

The determination of which spectral band maximizes the SNR depends on the spectral natures of the atmospheric transmission, of the optical transmission, and of the detector quantum efficiency. Dependencies of noise on the spectral band are also a consideration, as in the case of spectrally cold-filtered BLIP performance.

Longshore, Raimondi, and Lumpkin[15] presented examples of spectral band trade-offs for the case where the threshold of target detection is set by small temperature differences. They expressed the spectrally dependent quantities in the NETD equation by the radiation function M^* given by Eq. (1.29) in Sec. 1.2.5. Their examples used the LOWTRAN model for atmospheric transmission and assumed unity optical transmission, blackbody targets, and constant quantum efficiency in the spectral band.

Findlay and Cutten[13] presented a more general method, appropriate for any conditions, but required for ranges and/or atmospheres such that large ΔT's determine the limits of detectability. Although their equations are too complex

to reproduce and explain here, it is important to note that their approach is required for cases in which:

- the combination of background and atmosphere may not be a blackbody, as in the case of high elevation angle lines of sight to the sky
- measured $D^*(T_B)$ values are not available for the apparent background temperature of interest
- the background radiance is spectrally different from that of the target
- the SNR threshold crossover in the comparison of two candidate spectral bands occurs at high ΔT's
- sources of detector background flux from the target background, the atmosphere, and the optics must be considered separately.

Findlay and Cutten argue that the M^* approach to spectral band selection is appropriate only when a near-ambient blackbody target is viewed against an ambient blackbody background through a ground-level horizontal path.

Kleinhans[23] proved that the ideal spectral filter has either unity or zero transmission at every wavelength. An intuitive justification is as follows. It is likely that the combination of source signature and atmospheric transmission has a maximum and is relatively flat over a certain narrow waveband, and that optical and detector materials can be found to match that band. This narrow band is taken as the starting point for an optimization.

The exact limits of the optimum band can be investigated by eliminating a very narrow waveband from the starting passband. If elimination of this narrower band increases the integrated SNR, it should be left out. Otherwise, it should be kept. Likewise, expansion of the limits of the passband can be determined by adding a narrow waveband. The optimum passband is then found by iteration.

As an example, one of the perennial problems of terrestrial thermal imaging is the choice between the 3- to 5-μm and the 8- to 12-μm atmospheric windows. That choice should be based on a tailoring of the problem at hand to the generalized analysis procedure of Findlay and Cutten.[13] In general, the 8- to 12-μm range is preferred unless some shorter wavelength characteristics such as exhaust plumes or sun glints are of particular interest.

The cases analyzed by Findlay and Cutten indicate that the 8- to 12-μm window is superior to the 3- to 5-μm window at all ranges of tactical interest, unless the targets are very hot and ranges of tens of kilometers are required. The availability of detector arrays with selectable or simultaneous dual-color response offers the best of both bands with a modest increase in design complexity for circumstances where there is no clear advantage for either band.

1.3 CLUTTER NOISE

The experience of looking right at something and not seeing it is a common one: Your ears can guide you to look at a bird calling from a tree, but you may not be able to distinguish it within the clutter of branches and leaves. The experiential definition of clutter is that it consists of variations in the background having about the same size, radiant exitance, and texture of the source you are searching for. Clutter consists of scene elements similar enough in size and contrast to the sources sought that each one has to be considered in

detail as a potential target. In heavy clutter you need enough sensor resolution to delineate clearly straight lines on targets. Typically this means more than twice as much resolution than is needed without clutter. Clutter can be so strong that no amount of improvement in the SNR will reduce the search time to find the target, although resolution improvement might help.

1.3.1 Description

Mathematically, clutter is described by the variance of the scene radiant exitance after removal of low-frequency content such as the mean value and any slowly varying trends, and by the correlation lengths in the two angular dimensions in the orientation in which the scene is viewed.

Itakura, Tsutsumi, and Kakagi[24] observed that the statistics of scenes dominated by thermal emission tend to exhibit a random distribution of radiance which is Gaussian:

$$P(L) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(L - \bar{L})^2/2\sigma^2] \ , \tag{1.75}$$

where L is the radiance of an element of the scene, $\bar{L}$ is the average scene radiance, and σ^2 is the variance of the scene radiance.

They also observed that strongly reflective scenes with a relatively few highly impulsive, high-amplitude components tend to exhibit Poisson statistics:

$$P(L) = \frac{1}{C} \exp(-L/C) \ , \tag{1.76}$$

where C is the mean and square root of the variance.

Scene amplitude statistics can also be a combination of the Gaussian and the Poisson:

$$P(L) = \frac{1}{\sqrt{\pi C}} \exp(\sigma^2/2C^2) \exp[-(L - \bar{L})/C] \times \operatorname{erfc}\left(\frac{\sigma}{\sqrt{2C}} - \frac{L - \bar{L}}{\sqrt{2C\sigma}}\right) \ , \tag{1.77}$$

where the complementary error function is given by

$$\operatorname{erfc}(x) = \int_x^\infty \exp(-t^2)\, dt \ . \tag{1.78}$$

The spatial autocorrelation function tends toward the simple form

$$P(r) = \alpha \exp(-\alpha r) \ , \tag{1.79}$$

where r is the separation of any two points and α is the reciprocal of the average pulse width, that is, the reciprocal of the background correlation length.

The PSD combining both amplitude and spatial characteristics as derived by Ben-Yosef, Rahàt, and Feigen[25] for zero mean amplitude is given by

$$\mathrm{PSD}(f_x,f_y) = \frac{2\pi\alpha\sigma^2}{\alpha^2 + (2\pi f_x)^2 + (2\pi f_y)^2} , \tag{1.80}$$

where f_x and f_y are spatial frequencies in cycles per linear dimension.

For the thermal infrared, Ben-Yosef, Rahat, and Feigen[25] observed that where scene temperature variations are small, say about $\pm 10°C$ about an average temperature $\langle T \rangle$, the radiation functions are approximately linear in ΔT so that

$$P(T) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[- \frac{(T - \langle T \rangle)^2}{2\sigma^2} \right] , \tag{1.81}$$

where σ^2 is the variance of T.

Ator and White[26] concluded from comparisons of spatially registered infrared and visible scene data that anisotropic power spectral densities are common due to the presence of geological features and cloud structures. Thus a scene may not always be describable by a one-dimensional PSD.

Ben Yosef et al.[27–36] made critical contributions to the understanding of clutter and clutter statistics in the thermal infrared. They found that the measured statistics of IR scenes depend on the spatial resolution of the thermal imager used to make the observations because the imager resolution may be averaging over many smaller independent scene elements. They also found that although the scene statistics as measured by a high-resolution imager may be non-Gaussian, averaging by a lower resolution imager will tend, consistent with the central limit theorem, to produce Gaussian statistics.

Extreme care must be taken in the calculation and interpretation of PSDs to avoid deceptive results arising from the inclusion of low-frequency scene components. Scenes commonly contain slowly varying gradients where the amplitude of the gradient greatly exceeds the amplitudes of the superimposed informative variations. Large abrupt transitions are also common, as in the boundaries between ground and sky, between earth and water, or between sky and clouds.

In such cases the Fourier transform of the autocorrelation function will be dominated by the contribution of such features, and use of such a PSD will overestimate the masking effects of clutter. Therefore, it is necessary in most cases to remove large uniform trends and abrupt transitions from the scene data prior to calculating the PSD. Similarly, before using published PSD results, the method of calculation must be checked.

1.3.2 Diurnal Variation

Collectively, the papers of Ben-Yosef et al.[24,27–36] made four key observations about PSDs. First, the dominant factors are variations in the heat capacity, thermal conductivity, and solar irradiation absorption of the terrain. Second, in the presence of solar heating, variations in scene emissivity are not important because the emissivity of natural ground materials typically is greater

than 0.9. Third, as insolation increases, the standard deviation increases, and the correlation length decreases. Fourth, terrain features that are unobservable in steady-state conditions may become apparent during warming or cooling transients.

They wrote a heat balance equation accounting for heat capacity, thermal conductivity, thermal radiation, absorption, and insolation. Experiments validated this equation for the case of dry, still summer desert terrain. It reasonably accurately predicts observed washouts and contrast reversals, and leads to approximately Gaussian amplitude statistics. The equation predicts the skewing of the distribution observed under extreme conditions. The spatial autocorrelation is predictable from the insolation, with the correlation length decreasing as insolation increases.

1.3.3 Effects

Scene clutter has three major effects on target detection. First, when the density of clutter is low and the size and intensity of the clutter mimic the appearance of the target, there is a drastic slowing of search speed because each candidate target must be considered individually. A classic infrared example is sagebrush in the desert. Second, when the density of clutter is high and the PSD mimics that of sensor noise, masking the target, there is a reduction in the conspicuity of targets with consequent slowing of search speed.

Third, in automatic moving target detection using frame-to-frame differencing, the motion of the sensor between frames produces the appearance of motion in a large percentage of the scene elements. This can result from platform motion, scene motion, or platform jitter, and generates false tracks in the motion detection processing. Weathersby and Schmieder[37] defined a signal-to-clutter ratio and showed how detection requirements for spatial resolution may vary over a range of ten-to-one depending on the clutter. Pohlman[38] showed how to combine the effects on performance of noises due to background clutter, scintillation, photon statistics, and the detector.

1.4 SPATIAL FREQUENCY RESPONSE

Imaging has a fundamental limit to the size of the smallest angular detail that can be distinguished, called the *resolution*. There are three main reasons for this. First, all of the light collected from a mathematical point in the scene is not imaged at a single point in the image, due to the blurring effects of diffraction, geometrical aberrations, and scattering. Second, atmospheric scintillation causes disturbances in the LOS to a point in the scene. Third, the light sensing element is not a point, but is extended, so that spatial integration causes light from any point in the scene to be affected by the responses to adjacent scene points.

The resolution of a system may be described and analyzed using either of two approaches. One approach is to write the integrodifferential equations that describe the effects of each resolution-limiting phenomenon moving successively from input to output. The result is a superposition integral[39]:

$$I(x,y) = \iint_{-\infty}^{\infty} \mathbb{O}(\xi,\eta) S[\delta(x - \xi)\delta(y - \eta)]\, d\xi\, d\eta \ , \tag{1.82}$$

where $I(x,y)$ describes the image amplitude variation, $\mathbb{O}(\xi,\eta)$ describes the object or source amplitude variation in the dummy variables ξ and η, $\delta(x - \xi)\delta(y - \eta)$ is an ideal point source at the variable point (ξ,η), and the function S is the system response to that impulse, called the *impulse response.*

The second approach is applicable if the superposition process behaves sufficiently like a convolution to permit the use of transform theory. In that case a series of cascaded independent convolutionary imaging processes transforms to a series of multiplied frequency responses. Frequency-domain analysis ordinarily is simpler mathematically and yields more insight into the imaging process.

1.4.1 Linear Filter Theory

Linear filter theory is the application of frequency transforms to convolutionary spatiotemporal processes. It is useful for determining the response to specific signals (especially periodic ones), for predicting resolution, and for designing and evaluating image enhancement processes. When a system is linear a sine-wave input produces a sine-wave output. This is the reason Fourier transform analysis is used to decompose the inputs and outputs of a linear system into sine waves. Goodman[39] provides an excellent introduction to this subject. Its application to thermal imaging systems is given in Lloyd.[5]

Three conditions must hold for linear filter theory to be applicable. First, the processes must be linear. Second, they must be stationary over a sufficiently large span of time and/or space, called an *isoplanatic patch,* to be meaningfully applicable. Third, the imaging processes must be single-valued mappings of input points to output points, which implies that they are deterministic and effectively noise free.

The superposition integral indicates that the image of a source point is the sum of the responses of the system to that point and to all the surrounding points. For the condition of spatial invariance, the superposition integral reduces to a convolution integral:

$$I(x,y) = \iint_{-\infty}^{\infty} \mathbb{O}(\xi,\eta) r(x - \xi, y - \eta)\, d\xi\, d\eta \triangleq \mathbb{O}(x,y) * r(x,y) \ , \tag{1.83}$$

where $r(x - \xi, y - \eta)$ is the reverted point source response (impulse response) of the system. See Refs. 39 and 5 for the derivation of this equation.

Two types of transforms may be used depending on the nature of the process to be described. If the process occurs in an analog electrical circuit using a real-time data stream, the realizability requirement dictates that there can be no output before there is an input. Then the single-sided Laplace transform must be used, in which only the data appearing before the point in time under consideration influence the response at that time. If the process occurs in an optical device or in a software computation working on stored data, the Fourier transform is appropriate.

1.4.2 System Optical Transfer Function

The optical transfer function (OTF), or frequency amplitude and phase response, of a complete imaging system is the Fourier transform of its impulse response $r(x,y)$. This clearly is not properly defined if the impulse response is significantly affected by nonlinearities or noise, or if there are residual sampling effects in the image. This means that strictly speaking, Fourier transform analysis only approximates the behavior of most electronic imaging devices.

The one-dimensional Fourier transform is

$$F\{\mathbb{O}(x)\} = \int_{-\infty}^{\infty} \mathbb{O}(x) \exp(-2\pi i x f_x)\, dx \triangleq \tilde{\mathbb{O}}(f_x) \tag{1.84}$$

and the inverse transform is

$$F^{-1}\{\tilde{\mathbb{O}}(f_x)\} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \tilde{\mathbb{O}}(f_x) \exp(2\pi i x f_x)\, df_x = \mathbb{O}(x) \ . \tag{1.85}$$

By the convolution theorem,[39] in one dimension the transform of the image is

$$F\{\mathbb{O}(x) * r(x)\} = \tilde{\mathbb{O}}(f_x)\tilde{r}(f_x) \tag{1.86}$$

and the transform of the impulse response is defined as

$$\tilde{r}(f_x) = \mathrm{OTF}(f_x) = \mathrm{MTF}(f_x) \exp[i\mathrm{PTF}(f_x)] \ . \tag{1.87}$$

The modulus of the OTF is called the modulation transfer function (MTF) and describes the amplitude response to sine-wave inputs. The argument of the OTF is called the phase transfer function (PTF) and describes the phase response. Phase shifts are evidenced by asymmetry of the impulse response or, in the case of linear phase shift, by spatial translation of the center of the impulse response relative to the point source input. If all the processes of an imaging system are separable, that is, uncoupled and independent, then the system transfer function MTF_s is the product of the n component transfer functions MTF_i:

$$\mathrm{MTF}_s = \prod_{i=1}^{n} \mathrm{MTF}_i \ . \tag{1.88}$$

Following the notation of Schade,[40] the MTF is denoted by $\tilde{r}$ in lengthy equations.

There is no single ideal shape for the MTF of a system whose output is to be viewed by a human. Kusaka[41] showed for high-definition television imagery that the MTF that produces the most aesthetically satisfying picture depends on the picture content.

In addition to the primary sources of MTF loss described in subsequent paragraphs, many minor deviations from ideality exist that collectively can seriously reduce the system MTF. Some examples are optical material scattering, optical assembly errors such as lens tilt, the inability to focus precisely, scan speed variations in fixed-time-base systems, scan position measurement

errors in variable-time-based systems, scan-to-scan misregistration, internal structural vibration, and lossy video transmission lines.

1.4.2.1 Equivalent Bandwidth. Several investigators have found that the integral of the squared MTF correlates well with subjective measures of image quality. This quantity is usually referred to as the equivalent line number N_e, after Schade,[40] and is defined in one dimension by

$$N_e \triangleq \int_0^\infty [\tilde{r}(f_x)]^2 \, df_x \ . \tag{1.89}$$

It is conventional to normalize the modulus to a maximum value of 1 at very low, but not necessarily zero, video frequency. Line number N_e is related by a constant to visual acutance, a common measure of photographic resolution. It is intuitively justifiable as a summary measure because it is analogous to the noise-equivalent bandwidth, the integral of the product of the squared frequency response with the noise power spectrum. An equivalent resolution r is usefully defined by

$$\bar{r} = (2N_e)^{-1} \ . \tag{1.90}$$

This definition is chosen to make the $\bar{r}$ for a rectangular impulse response of width α equal to α. Some examples of equivalent bandwidth are:

Circular diffraction-limited aperture: (by numerical integration)	$N_e = 0.072/f_c$ $= 0.272\lambda/D$,	(1.91)
Square diffraction-limited aperture:	$N_e = \lambda/(3D_o)$,	(1.92)
Gaussian impulse response:	$N_e = (4\sigma\sqrt{\pi})^{-1}$,	(1.93)
Rectangular detector impulse response:	$N_e = (2\alpha)^{-1}$.	(1.94)

See subsequent sections for definitions of the preceding impulse responses.

1.4.2.2 Central Limit Theorem and Equivalent Gaussian. The central limit theorem of statistics as applied to optics states that the product of a series of monotonically decreasing component MTFs will tend toward a Gaussian shape. Since there are usually at least six transfer functions in a system (diffraction, optical aberrations, detector geometry, detector time constant or integration time, signal processing, and display), and usually many more, the system MTF often is reasonably well approximated by a circularly symmetrical Gaussian impulse response

$$r(x,y) = \frac{1}{2\pi\sigma^2} \exp\left(- \frac{x^2 + y^2}{2\sigma^2} \right) \tag{1.95}$$

whose MTF in either direction is

$$\tilde{r}(f) = \exp(-2\pi^2\sigma^2 f^2) \ . \tag{1.96}$$

The definition of the Gaussian function equivalent to a given near-Gaussian MTF could be made in several ways, such as defining equivalence at a critical spatial frequency or by making a least-squares fit. The recommended way is to determine the σ that produces the same equivalent bandwidth:

$$\sigma = (4N_e\sqrt{\pi})^{-1} = 0.142/N_e \ . \tag{1.97}$$

1.4.2.3 Central Ordinate Theorem. The central ordinate theorem of Fourier transform theory states that the area of a function $f(x)$ equals the central ordinate of its transform $F(f_x)$:

$$F(0) = \int_{-\infty}^{\infty} f(x)\, dx \ , \tag{1.98}$$

and that the area of a function's transform equals the central ordinate of the function

$$f(0) = \int_{-\infty}^{\infty} F(f_x)\, df_x \ . \tag{1.99}$$

This is useful for calculating the peak value of a waveform from knowledge of its transform.

1.4.2.4 Electronic Boost. Electronically improving the overall system MTF by high-frequency emphasis is called *video boosting* or *aperture correction*. The theory is well grounded in experiment, and is based on two considerations. First, matched filter theory proves that the detectability of a signal is maximized by first whitening the noise power spectrum prior to application of the matched filter. Second, it is well known from television experiments that the parameter of noise that is significant is not the rms noise, but the amplitude of the noise PSD.

It is demonstrable that a less-than-white PSD can be whitened with no deleterious effect. Therefore, in theory, the signal response can safely be boosted by at least as much boost as will whiten the noise; in practice, it is common to boost just enough so that the noise does not "ring." Kusaka[41] showed that a flat to slightly peaked overall MTF is aesthetically pleasing.

1.4.3. Optical OTFs

There are two limiting cases for which closed-form expressions are available. The first is the case of diffraction-limited optics with no geometrical aberrations. Then the OTF is the scaled autocorrelation of the shape of the entrance pupil. The second is the case where the geometrical aberrations are so strong that their effect is describable by ray-trace statistics, as opposed to wave-front variation calculations. Then if the aberrations are not too severe, they can be collectively reasonably approximated by a Gaussian MTF using Eq. (1.96). In between is the general case where wave-front calculations including diffraction must be used, wherein there are few closed solutions.

For a clear circular diffraction-limited aperture monochromatically illuminated,

$$\tilde{r}(f) = \frac{2}{\pi}\left\{\arccos\left(\frac{f}{f_c}\right) - \left(\frac{f}{f_c}\right)\left[1 - \left(\frac{f}{f_c}\right)^2\right]^{1/2}\right\}, \qquad \text{for } \frac{f}{f_c} \leq 1 \tag{1.100}$$

where $f_c = D_o/\lambda$.

The MTF of an incoherent polychromatic optic is the normalized weighted integral of the OTFs at each component wavelength, given by Levi[42] as

$$\text{MTF} = \frac{\int_{\lambda_1}^{\lambda_2} \text{MTF}(f,\lambda) E_\lambda \tau_o(\lambda) \mathcal{R}(\lambda)\, d\lambda}{\int_{\lambda_1}^{\lambda_2} E_\lambda \tau_o(\lambda) \mathcal{R}(\lambda)\, d\lambda} \tag{1.101}$$

where, as before, E_λ is the irradiance of the source, $\tau_o(\lambda)$ is the optical transmission, and $\mathcal{R}(\lambda)$ is the detector responsivity.

Gafni et al.[43] indicate that in the infrared the polychromatic equation often may be adequately approximated by the monochromatic equation, using the midband wavelength, for spatial frequencies well below the cutoff.

1.4.4 Detector OTFs

The most common detector spatial geometry is rectangular. If the detector dissects the scene by some form of motion such as scanning or dithering, its impulse response in one direction is:

$$r(x) = \text{rect}\left(\frac{x}{\alpha}\right) = \begin{cases} 1, & \text{for } |x/\alpha| \leq \dfrac{1}{2} \\ 0, & \text{for } |x/\alpha| > \dfrac{1}{2} \end{cases} \tag{1.102}$$

and its MTF is

$$\tilde{r}(f_x) = \frac{\sin(\pi\alpha f_x)}{\pi\alpha f_x} \triangleq \text{sinc}(\alpha f_x)\,. \tag{1.103}$$

If the detector dissects the scene by staring, the detector shape influences the frequency content of the image, but the MTF is undefined because the stationarity existence condition is violated.

1.4.5 Signal Processing OTFs

The conversion in a scanning sensor from spatial frequency f_x to electrical frequency f_t is given by:

$$f_t = f_s \dot{\theta}\,, \tag{1.104}$$

where $\dot{\theta}$ is the scan velocity in angular units per second. If a scanning detector of angular dimension α is temporally integrated for a time t_i, that integration

has an equivalent spatial dimension of α_i:

$$\alpha_i = \alpha t_i/\tau_d \ , \tag{1.105}$$

where τ_d is the scan dwell time corresponding to the angular dimension α. The resulting MTF is

$$\tilde{r}(f) = \text{sinc}(\alpha_i f) \ . \tag{1.106}$$

If this or any other electrical MTF is significant in a bidirectionally scanning design, the MTF loss from the associated phase transfer function must be accounted for.

A single-RC low-pass filter has

$$\text{OTF} = (1 + jf/f_c)^{-1} \ , \tag{1.107}$$

$$\text{MTF} = [1 + (f/f_c)^2]^{-1/2} \ , \tag{1.108}$$

$$\text{PTF} = \arctan(-f/f_c) \ , \qquad f_c = (2\pi\text{RC})^{-1} \ . \tag{1.109}$$

A single-RC high-pass filter has

$$\text{OTF} = jf/(1 + jf/f_c) \ , \tag{1.110}$$

$$\text{MTF} = (f/f_c)[1 + (f/f_c)^2]^{-1/2} \ , \tag{1.111}$$

$$\text{PTF} = \arctan(f_c/f) \ . \tag{1.112}$$

According to Sequin and Thompsett,[44] a charge transfer device such as might be used in a focal plane readout, a time-delay-and-integrated (TDI) circuit, or an analog memory has an MTF and a PTF due to a nonunity charge transfer efficiency of

$$\text{MTF} = \exp\{-m(1 - \varepsilon)[1 - \cos(2\pi f/f_s)]\} \tag{1.113}$$

$$\text{PTF} = -m\left[\frac{2\pi f}{f_s} - \sin\left(\frac{2\pi f}{f_s}\right)\right] \tag{1.114}$$

where m is the number of gates in the transfer from the detector to the output amplifier, ε is the charge transfer efficiency of each gate, and f_s is the sampling frequency of the structure.

1.4.6 Square-Wave Response

Truncated square-wave targets ("bar charts") are often used instead of extended sine-wave targets because of the difficulty of making the latter, especially in the infrared. Bar charts also are used because the theory of target

recognition relates the probability of recognition to the resolvability of bar charts having the same dimension and signal amplitude as the target. System responses to a series of bar charts are also used to make an estimate of the MTF when impulse response measuring equipment is not available. A four-bar target of bar width W may be described in the direction of the modulation using the notation of Sec. 1.5.1:

$$\mathcal{O}(x) = \text{rect}\left(\frac{x}{W}\right) * [\delta(x - W) + \delta(x + W) + \delta(x - 3W) + \delta(x + 3W)] \; . \tag{1.115}$$

Fourier transforming and normalizing, its amplitude spectrum is:

$$\tilde{\mathcal{O}}(f_x) = \text{sinc}(Wf_x)\left[\frac{1}{2}\cos(2\pi f_x W) + \frac{1}{2}\cos(6\pi f_x W)\right] \; . \tag{1.116}$$

The four-bar response of a Gaussian MTF does not differ appreciably from its infinite square-wave response, so Coltman's relations[45] between square-wave response $\tilde{r}(f)$ and sine-wave response $\tilde{r}(f)$ for a fundamental frequency f and its harmonics nf usually can be used with confidence for a four-bar chart:

$$\tilde{r}(f) = \frac{4}{\pi}\left[\tilde{r}(f) - \frac{\tilde{r}(3f)}{3} + \frac{\tilde{r}(5f)}{5} - \frac{\tilde{r}(7f)}{7} + \frac{\tilde{r}(9f)}{9} + \ldots + \right] \tag{1.117}$$

plus other regular terms, and

$$\tilde{r}(f) = \frac{\pi}{4}\left[\tilde{r}(f) + \frac{\tilde{r}(3f)}{3} - \frac{\tilde{r}(5f)}{5} + \frac{\tilde{r}(7f)}{7} + \frac{\tilde{r}(11f)}{11} - \frac{\tilde{r}(13f)}{13} - \frac{\tilde{r}(15f)}{15} + \ldots + \right] \tag{1.118}$$

plus other irregular terms per Coltman.

1.4.7 Theoretical Limit to Target Position Estimation

Fried[46,47] derived the rms uncertainty with which the position of a single point source target may be measured by a scanning sensor with one-dimensional output limited by additive white Gaussian noise. This noise-limited position measurement precision depends on the nature of the source, the transfer function of the measuring system, and the noise of the system. The result is similar to that of radar theory where the angular measurement precision is a resolution-related constant divided by the voltage SNR.

Fried defined a "resolution scale" $\delta\theta$ given by

$$\delta\theta = \frac{1}{2\pi}\left[\frac{\int_0^\infty \mathrm{MTF}^2(f)\,df}{\int_0^\infty f^2\mathrm{MTF}^2(f)\,df}\right]^{1/2} \tag{1.119}$$

such that the angular measurement precision θ is

$$\theta = \frac{\delta\theta}{\mathrm{SNR}} \ . \tag{1.120}$$

Further, the measurement of the separation of two targets is the resolution scale divided by the "composite SNR" given by

$$\mathrm{SNR}_c = (\mathrm{SNR}_1^{-2} + \mathrm{SNR}_2^{-2})^{-1/2} \ . \tag{1.121}$$

The result is that position determination is achievable to a much finer accuracy than the conventionally defined resolution of the system, for example, much less than λ/D_o, given sufficient SNR. This result does not apply to the cross-scan direction or to staring systems in which the only knowledge about the target location is that it falls within a detector angular subtense β. In that case, the rms angular uncertainty σ_θ in one dimension for one line or pixel is

$$\sigma_\theta = \frac{\beta}{\sqrt{12}} \ . \tag{1.122}$$

As an example of the application of Eq. (1.119), if the MTF is a Gaussian as defined previously in Eq. (1.96),

$$\delta\theta = \sigma\sqrt{2\pi} \ . \tag{1.123}$$

For two targets having equal SNR and the $\delta\theta$ given by Eq. (1.123), Eq. (1.121) produces the result

$$\delta\theta = 2\sigma \ . \tag{1.124}$$

For a circular diffraction-limited aperture of diameter D_o, monochromatically illuminated by light of wavelength λ, the numerically integrated value of Eq. (1.119) is

$$\delta\theta = \frac{3.704}{2\pi f_c} = 0.59/f_c \ . \tag{1.125}$$

If the system is diffraction-limited by a rectangular aperture of width D_o, and the illumination is monochromatic with wavelength λ, Eq. (1.119) produces

$$\delta\theta = 0.5\lambda/D_o \ . \tag{1.126}$$

The best possible case is that of a unity MTF out to a cutoff of f_c:

$$\delta\theta = \frac{\sqrt{3}}{2\pi f_c} = 0.276/f_c \ . \tag{1.127}$$

1.5 SAMPLING AND ALIASING

Sampling occurs whenever a system deviates from convolutionary image formation because its impulse response is not spatially or temporally invariant. This occurs in space when image dissection is performed at discrete loci, such as along scan line centers or at fixed array points. It occurs in time when the image is sampled at discrete intervals, such as by framing or periodic image integration. Inadequately closely spaced sampling causes disruption of periodic patterns, a blocky appearance in static imagery, and shimmering in moving images. Many image interpretation experiments have shown that sampling effects are as detrimental to visual recognition as are noise and blurring.

A common example of a spatial sampling effect is the disruption of the pattern of a sportscaster's plaid coat by the fixed red-green-blue pattern of a color TV camera. An example from "most wanted" crime videos is the unrecognizability of the face of a witness whose anonymity is protected by constructing his or her facial image with a few coarse blocks of video. In these cases some of the higher spatial frequency power in the original scene spectrum is shifted to both higher and lower frequency bands called *aliases* or *aliased frequencies*. A common temporal frequency example is the case of a rotating wheel sampled in time by a TV camera's framing action. The wheel appears to move faster, slower, or even backward, depending on its real rotation rate relative to the frame rate.

1.5.1 Elemental Structures and Replication

Bracewell[48] and Goodman[39] introduced a shorthand notation that produces compact equational representations of complicated sampled imaging processes. Any sampling process based on a fixed one-dimensional or a fixed orthogonal two-dimensional sampling lattice is easily analyzed using their technique. It requires only that the sampling structure be composed of elemental averaging apertures replicated in the regular pattern of a sampling lattice. This notation avoids a much more complicated representation by a series of equations with boundary conditions.

The fundamental element in this notation is the Dirac delta function defined for one direction by

$$\delta(x) = \begin{cases} 1, & \text{for } x = 0 \\ & \text{for all } y \ . \\ 0, & \text{for } x \neq 0 \end{cases} \tag{1.128}$$

In this case, $\delta(x)$ is a knife-edge extending infinitely along the y axis.

A Dirac delta located other than at $x = 0$ is indicated by $\delta(x - \xi)$ and exists at $x = \xi$ for all y. Periodic imaging structures such as scanned or staring detectors are constructed by replicating the unit cell of the array on lattice structures, which are themselves arrays of delta functions. These lattice structures are called *comb functions*.

A one-dimensional array of Dirac deltas separated by the dimension d in the y direction is given by

$$\operatorname{comb}\left(\frac{y}{d}\right) = \sum_{n=-\infty}^{\infty} \delta\left(\frac{y}{d} - n\right) = \begin{cases} 1, & \text{for } y = nd \\ 0, & \text{otherwise} \end{cases} \quad \text{for all } x \ . \tag{1.129}$$

This can be visualized as an infinite series of knife-edges parallel to the x axis.

An infinite array of points along the y axis with separation d is given by

$$\operatorname{comb}\left(\frac{y}{d}\right) \cdot \delta(x) \ . \tag{1.130}$$

A two-dimensional array of mathematical points spaced by c in the x direction and by d in the y direction is represented by

$$\operatorname{comb}\left(\frac{x}{c}, \frac{y}{d}\right) = \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \delta\left(\frac{x}{c} - m\right)\delta\left(\frac{y}{d} - n\right) \ . \tag{1.131}$$

The very useful rectangular function is defined by

$$\operatorname{rect}\left(\frac{x}{a}\right) = \begin{cases} 1, & \text{for } |x/a| \leq \dfrac{1}{2} \\ 0, & \text{for } |x/a| > \dfrac{1}{2} \end{cases} \quad \text{for all } y \ . \tag{1.132}$$

This is used to represent the unit cell of a detector array by

$$\operatorname{rect}\left(\frac{x}{a}, \frac{y}{b}\right) = \begin{cases} 1, & \text{for } |x/a| \leq \dfrac{1}{2} \text{ and } |y/b| \leq \dfrac{1}{2} \\ 0, & \text{otherwise} \end{cases} \ , \tag{1.133}$$

where a and b are the detector dimensions in the x and y directions, respectively.

This notation is also useful for deriving the spectrum of periodic structures such as a bar chart (see Sec. 1.4.6), of a compound detector structure such as a CCD element centrally obscured by an opaque contact, of the impulse response of a detector array exhibiting crosstalk, and of the impulse response of a TDI array with misregistration due to scan velocity error.

1.5.2 Sampling Mathematics

Any sampled imaging process is characterized by three stages. The first is convolution of the object distribution $\mathcal{O}(x,y)$ with the before-sampling impulse response $r_{bs}(x,y)$. The second is multiplication of that intermediate image by

the sampling lattice $L(x,y)$ of delta functions. Proof that those two processes occur in that order is given by Lloyd.[5] The third stage is convolution of the sampled image with the after-sampling image processing and reproduction impulse response $r_{as}(x,y)$. The spatial domain sampled image signal equation is:

$$I(x,y) = \{[\mathbb{O}(x,y)*r_{bs}(x,y)] \times L(x,y)\}*r_{as}(x,y) \ . \tag{1.134}$$

The equivalent steps in the frequency domain are multiplication of the object spectrum $\tilde{\mathbb{O}}(f_x,f_y)$ by the before sampling (prefiltering) MTF $\tilde{r}_{bs}(f_x,f_y)$, convolution of this product with the transform of the sampling lattice $\tilde{L}(f_x,f_y)$, and multiplication of the resulting aliased spectrum by the after-sampling reproducing (postfiltering) MTF $\tilde{r}_{as}(f_x,f_y)$. The frequency domain image equation is:

$$\tilde{I}(f_x,f_y) = \{[\tilde{\mathbb{O}}(f_x,f_y) \times \tilde{r}_{bs}(f_x,f_y)]*\tilde{L}(f_x,f_y)\} \times \tilde{r}_{as}(f_x,f_y) \ . \tag{1.135}$$

1.5.3 Signal Aliasing and Noise Foldover

The sequence of processes in Eq. (1.135) shows that the postfilter can pass aliased signal power into the output if the first harmonic of the sampling frequency is less than twice the highest frequency present in the prefiltered object spectrum. This occurs because the aliased sideband spectra will overlap the fundamental spectrum and masquerade as components of the source. The requirement of a sampling frequency twice the frequency of interest is called the *Nyquist criterion,* and is elaborated by the Whittaker-Shannon sampling theorem.

Similar equations can be written for the noise component for the direction in which noise is sampled. In that case the signal spectrum is replaced by the noise PSD and the analysis is done in power space rather than signal space, with the corresponding change from linear addition to addition in quadrature. The noise PSD is likely to be broader than the signal spectrum because more transfer functions limit the width of the signal spectrum. Thus noise "foldover" is a more serious effect in multiplexed systems than is signal aliasing. Usually such an analysis is necessary only for channel-multiplexed imagers.

Aliasing of signal masks perception of scene detail much the same as does random noise. Resnikoff[49] offered a proof based on the uncertainty principle, which demonstrated the equivalence between aliasing and noise. Resnikoff's proposed alternative to disruptive aliasing is to sample randomly such that scene content which would be inadequately sampled by a regular structure is converted to less disruptive uncorrelated noise, rather than to aliases.

Another alternative is to sample in a fixed lattice, but to translate the lattice by sub-sample-spacing dimensions between frames. If the frame rate is such that several frames are created during the eye integration time appropriate to the viewing conditions, then moving the sampling lattice in increments smaller than the sample separation will result in the perception that the imagery is sampled at a higher spatial frequency, or even is unsampled. Dann, Carpenter, and Seamer[50] show the benefits of such regular periodic displacement, which they called *microscanning,* with 2×2 and 3×3 patterns in a 64×64 imaging array.

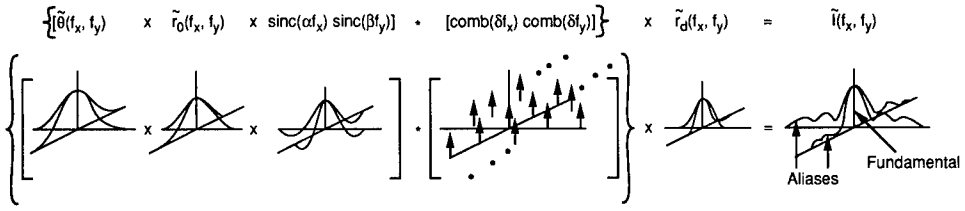


Fig. 1.4 Aliasing in a staring imager.

In situations where the sampling rate cannot be changed, considerable attention has been given to the question of whether one should deliberately increase prefiltering to reduce aliasing or simply let the aliasing occur. The current consensus based on simulated imagery is that one should not prefilter more than the process naturally dictates. For example, see the experiment reported by van Meeteren and Mangoubi.[51]

1.5.4 Staring Array Example

The image signal equation for a staring array with unit cell dimension α and β on lattice centers separated by δ and γ is:

$$I(x,y) = \left\{\left[\mathbb{O}(x,y)*r_o(x,y)*\text{rect}\left(\frac{x}{\alpha},\frac{y}{\beta}\right)\right] \times \left[\text{comb}\left(\frac{x}{\delta},\frac{y}{\gamma}\right)\right]\right\}*r_d(x,y) \quad , \quad (1.136)$$

where $r_o(x,y)$ is the optics impulse response and $r_d(x,y)$ is the display (reconstruction) impulse response. The signal spectrum is

$$\tilde{I}(f_x,f_y) = \{[\tilde{\mathbb{O}}(f_x,f_y) \times \tilde{r}_o(f_x,f_y) \times \text{sinc}(\alpha f_x)\,\text{sinc}(\beta f_y)] * [\text{comb}(\delta f_x)\,\text{comb}(\gamma f_y)]\} \times \tilde{r}_d(f_x,f_y) \quad . \quad (1.137)$$

The components of Eq. (1.137) are shown in Fig. 1.4. Note that there is cross aliasing in the x and y frequencies. Because the sampling frequencies $1/\delta$ and $1/\gamma$ necessarily are less than half the maximum frequencies $1/\alpha$ and $1/\beta$ of interest in the presampled spectrum, i.e., $\delta \geq \alpha$ and $\gamma \geq \beta$, strong aliasing is unavoidable. Bradley and Dennis[52] give examples of imagery from such staring arrays and show corresponding aliasing calculations. The microscanning approach of Dann, Carpenter, and Seamer[50] is the only solution to the problem of such aliasing short of optical prefiltering so strong that its effect is worse than that of the aliasing it eliminates.

1.5.5 Interpolative Display Reconstruction

Although the aliased signal information in an inadequately sampled image cannot be unambiguously recovered by any postsampling processing, the disturbing "blockiness" or "edginess" of the aliased power can be reduced by image reconstruction with an interpolator. This is what is done in compact disk players, which claim "oversampling." This is not really an increase in input

sampling but an interpolation to increase the number of output samples prior to reproduction.

A simple illustrative example is the use of a one-dimensional linear interpolator to produce one new "sample" between each of the original samples. This is actually a sinc filter with a first null at the sampling frequency, followed by convolution with the original sampling lattice shifted by one-half sample separation. The net effect is to postfilter the sampled spectrum and shift the sidebands out in frequency. From a mathematical point of view information has been lost, but the image is much more "viewable" because the edginess produced by the sidebands has been reduced by pushing the sidebands out prior to reconstructive postfiltering. In practice, a two-dimensional sampling process with unavoidable aliasing should be reconstructed with a two-dimensional interpolator.

1.6 VISUAL DETECTION OF NOISY SIGNALS

Two common experiences exemplify the theory that follows. The first is the experience of driving in heavy fog where you are straining to detect small features like lights. Then suddenly you become aware of an unlighted vehicle at the threshold of your perceptibility, close and looming large in your field of vision, but devoid of details. The ability to draw that large object out of an otherwise washed-out low-contrast scene is an example of visual spatial integration. The second experience is the appearance of a single frame of video frozen on a VCR, which looks noticeably noisier than the continuous video. That difference is an example of the phenomenon of visual temporal integration.

A model of vision must account for these effects to predict the visual response to the SNR of processed and reconstructed video. Coltman and Anderson,[53] Schade,[54] and Rosell and Willson[55] observed that vision seems to operate as if the threshold of perception of signal sources in noisy video is a function of a hypothetical perceived SNR. This quantity is proportional to the square root of the ratio of the displayed target area A_T to the displayed noise correlation area A_c:

$$\mathrm{SNR}_p \propto \left(\frac{A_T}{A_c}\right)^{1/2} \mathrm{SNR}_i \ . \tag{1.138}$$

Vision is presumed to produce this SNR by an approximation to matched filtering with adaptive selection of tunable compound spread functions. The net effect functions similar to spatial integration and explains the phenomenon that large sources are more detectable than smaller sources with the same point SNR, within the limit of about 12 mrad of diameter discovered by Bagrash, Kerr, and Thomas.[56]

DeVries,[57] Rose,[58] Coltman and Anderson,[53] Schade,[54] and Rosell and Willson[55] observed that the visual detection process also behaves as if it were integrating in time with an integration period ranging from 0.05 s at high display luminance to 0.25 s at low luminance. This is presumed, similar to spatial integration, to produce a perceived SNR proportional to the square root of the product of the integration time and the video frame rate:

$$\mathrm{SNR}_p \propto \left(\dot{F} t_e\right)^{1/2} \mathrm{SNR}_i \ , \tag{1.139}$$

where t_e is the eye integration time. Combining both observations, we model vision as if it detects noisy sources by thresholding the signal after a signal processing enhancement, producing an effective SNR of

$$\mathrm{SNR}_p = \mathrm{SNR}_i (\dot{F} t_e A_T / A_c)^{1/2} \ . \tag{1.140}$$

The resulting probability of detection as it relates to vision in electronic displays is given by Lloyd.[5]

An example of the application of the idea of SNR_p is the minimum resolvable temperature difference (MRTD) concept of thermal imaging. MRTD is the minimum temperature difference between the bars and the background of a four-bar chart that permits a 70% probability of visual recognition of the correct orientation of the bars. The pattern is presumed to be recognized when the SNR_p produced by matched filtering and temporal integration is high enough to support detection of the individual bars.

MRTD is plotted as a function of the bar chart fundamental spatial frequency. The concept was introduced by Sendall and Lloyd,[59] and by Laakmann,[18] as an amplification of the recognition by Genoud[60] that the television and photographic analyses of Schade[40] are applicable to thermal imaging. The theory of MRTD can be formulated in a variety of ways. The model of Ratches[61] and his associates is the most widely used because it is the basis for the U.S. Army's computer codes for synthesis and evaluation of thermal imagers. Alternative derivations are given in Refs. 5 and 7. Gao, Kharim, and Zheng[62] generalized the Ratches model to permit extension of the MRTD concept to any system, even one that uses machine recognition.

The model is widely used because of government advocacy, but some details are questionable. For example, Vortman and Bar-Lev[63] found that a synchronous integrator model limited to an angular extent of 0.5 deg better fits measured data than the matched filter used in Ratches. That is, the synchronous integrator model shows asymptotic behavior of the MRTD at low spatial frequency, whereas the accepted model does not. Examples of the trade-offs possible within the model are given in the papers of Vortman and Bar-Lev.[64,65] Kennedy[66] also questioned the low-frequency part of the model, and suggested improvements to account for aliasing and fixed pattern noise.

1.7 APPLICATION TO SYSTEM DESIGN

This chapter provides the tools to make first-order estimates of the following:

- the theoretical limits to resolution and SNR
- the frequency response and noise-limited performance of imaging sensors
- how SNR and resolution combine to determine position location accuracy
- the strength of scene clutter relative to signal sources.

The system specifier or designer must recognize what class of design problem confronts him or her so that the fundamental problem of that class of design

can be identified and solved. Then the designer can select from the tools summarized here to analyze the problem and allocate performance requirements. For example, the design of a system for penetration of severe atmospheric attenuation must emphasize sensitivity. One designed to work in high clutter must have higher resolution than is customary. An intelligence gathering system should have an extremely wide dynamic range, cleverly compressed for data extraction.

Every design problem is characterized by unavoidable limits or restrictions, such as the available spectrum or the permissible optical aperture. A good design grows around those limits in such a way that the resulting concept is the obvious solution. This is achieved by systematic trade-offs radiating from the limitations, and involving FOV, resolution, signal sensitivity, target peculiarities, technology availability, cost, and risk.

The designer must avoid "blind spots" in the technologies used, which might allow opponents to choose to operate in conditions that conceal their presence. This means that elements such as rain, clutter, signature inversions, and diurnal changes must be considered. The designer's own blind spots are an issue as well. Examples abound of what the social critic Lewis Mumford called "technological exhibitionism," wherein a pet concept is implemented at tremendous expense in a regime where it is clearly suboptimal.

Every design problem is sufficiently unique such that analyses in the literature or standard tools are rarely wholly adequate. A custom analysis based on a tailored theory is almost always required because of peculiarities in targets, backgrounds, and environments. The original sources listed here as references should therefore be consulted to provide the insight and raw material for model development.

Acknowledgments

My sincere thanks go to Dave Shumaker of IRIA for proposing that I author this chapter. I am grateful to Mike Dudzik of ERIM for his kind invitation to contribute to his volume, and for his generous patience, reviews, and editorial suggestions. I am indebted to Handbook editor Joe Accetta of IRIA for providing outstanding leadership of this project, much encouragement, and a thorough review.

This opportunity would not have come my way had not Luc Biberman of the Institute for Defense Analyses so generously aided my career many years ago. Nor would I have grasped the subject matter without the tutelage of Bob Sendall of Hughes Aircraft Company, who continues to set an ever-rising standard of achievement in the art of thermal imaging.

More recently, I am deeply indebted to my friend Paul Cox of the U.S. Army's Night Vision and Electro-Optics Directorate for his review and for many stimulating conversations. For their reviews and assistance with permission to publish, I owe thanks to Jack Maziorka and Bill Cannon of Loral Vought Systems. Readers with suggestions for improving the presentation of this material are encouraged to contact me.

References

1. L. M. Biberman, Ed., *Perception of Displayed Information,* Plenum Press, New York (1973).
2. E. L. Dereniak and D. G. Crowe, *Optical Radiation Detectors,* John Wiley & Sons, New York (1984).
3. R. D. Hudson, Jr., *Infrared System Engineering,* Wiley-Interscience, New York (1969).
4. M. A. Karim, *Electro-Optical Devices and Systems,* PWS-Kent Publications, Boston (1990).
5. J. M. Lloyd, *Thermal Imaging Systems,* Plenum Press, New York (1975).
6. I. J. Spiro and M. Schlessinger, *Infrared Technology Fundamentals,* Marcel Dekker, New York (1989).
7. F. Rosell and G. Harvey, Eds., "The fundamentals of thermal imaging systems," Report 8311, Naval Research Laboratory, Washington, DC (1979).
8. K. Seyrafi, *Electro-Optical Systems Analysis,* 3rd ed., Electro-Optical Research Company, Los Angeles (1985).
9. D. L. Shumaker, J. T. Wood, and C. R. Thacker, *FLIR Performance Handbook,* DCS Corporation, Alexandria, VA (1988).
10. J. A. Jamieson, "Passive infrared sensors: limitations on performance," *Applied Optics* **15**(4), 891–909 (1976).
11. J. R. Meyer-Arendt, "Radiometry and photometry: units and conversion factors," *Applied Optics* **7**(10), 2081–2084 (1968).
12. G. J. Zissis, "Infrared technology fundamentals," *Optical Engineering* **15**(6), 484–497 (1976).
13. G. A. Findlay and D. R. Cutten, "Comparison of performance of 3–5 and 8–12 micrometer infrared systems," *Applied Optics* **28**(23), 5029–5037 (1989).
14. H. Barhydt, "Figures of merit for infrared sensors," in *Modern Utilization of Infrared Technology V, Proceedings of the SPIE* **197,** 64–79 (1979).
15. R. Longshore, P. Raimondi, and M. Lumpkin, "Selection of detector peak wavelength for optimum infrared system performance," *Infrared Physics* **16,** 639–647 (1976).
16. E. J. Borg, J. S. Davis, and B. D. Thompson, "Modeling approaches to thermal imaging systems," in *Thermal Imaging, Proceedings of the SPIE* **636,** 2–16 (1986).
17. R. G. Driggers, C. E. Halford, G. D. Boreman, and M. R. Wellfare, "Comparison of two frame noise calculations for infrared linescanners," *Optical Engineering* **29**(7), 781–785 (1990).
18. P. Laakmann, "Multiple detector thermal imaging systems," in *Infrared Technology, Proceedings of the SPIE* **62,** 243–248 (1975).
19. H. Barhydt, "Performance of nearly BLIP detectors in infrared sensors," *Optical Engineering* **15**(6), 498–509 (1976).
20. H. Barhydt, "Effect of F/number and other parameters on FLIR in nearly BLIP systems," *Optical Engineering* **17**(2), SR28–32 (1977).
21. H. Barhydt, "Effect of F/number and other parameters on performance of nearly BLIP search and surveillance systems (Part 1)," *Optical Engineering* **17**(5), SR94–96 (1978).
22. H. Barhydt, "Effect of F/number and other parameters on performance of nearly BLIP search and surveillance systems (Part 2)," *Optical Engineering* **18**(1), SR6–7 (1979).
23. W. A. Kleinhans, "Optimum spectral filtering for background-limited infrared systems," *Journal of the Optical Society of America* **55,** 104 (1965).
24. Y. Itakura, S. Tsutsumi, and T. Kakagi, "Statistical properties of the background noise for the atmospheric windows in the intermediate infrared region," *Infrared Physics* **14,** 17–29 (1974).
25. N. Ben-Yosef, B. Rahat, and G. Feigen, "Simulation of IR images of natural backgrounds," *Applied Optics* **22**(1), 190–193 (1983).
26. J. T. Ator and K. P. White III, "Investigations of isotropy in earth background clutter," in *Modern Utilization of Infrared Technology IV, Proceedings of the SPIE* **156,** 157–163 (1978).
27. N. Ben-Yosef and K. Wilner, "Temporal behavior of thermal images," *Applied Optics* **24**(2), 284–286 (1985).
28. N. Ben-Yosef, K. Wilner, S. Simhony, and G. Feigen, "Measurement and analysis of 2-D infrared natural background," *Applied Optics* **24**(14), 2109–2113 (1985).

29. N. Ben-Yosef, K. Wilner, I. Fuchs, S. Simhony, and M. Abitbol, "Natural terrain infrared radiance statistics: daily variation," *Applied Optics* **24**(23), 4167–4171 (1985).

30. N. Ben-Yosef, K. Wilner, S. Simhony, and M. Abitbol, "Correlation length of natural terrain infrared images; daily variation," *Applied Optics* **25**(6), 866–869 (1986).

31. N. Ben-Yosef, K. Wilner, and M. Abitbol, "Radiance statistics vs ground resolution in infrared images of natural terrain," *Applied Optics* **25**(16), 2648–2649 (1986).

32. N. Ben-Yosef, S. Lashansky, K. Wilner, and M. Abitbol, "Temporal prediction of infrared images of ground terrain," *Applied Optics* **26**(11), 2128–2130 (1987).

33. N. Ben-Yosef, K. Wilner, S. Lashansky, and M. Abitbol, "Prediction of temporal changes of natural terrain infrared images," in *Passive Infrared Systems and Technology, Proceedings of the SPIE* **807,** 58–60 (1987).

34. N. Ben-Yosef, K. Wilner, and M. Abitbol, "Natural terrain in the infrared: measurements and modeling," in *Infrared Technology XIII, Proceedings of the SPIE* **819,** 66–71 (1987).

35. N. Ben-Yosef, K. Wilner, and M. Abitbol, "Measurement and modelling of natural desert terrain in the infrared," *Optical Engineering* **27**(11), 928–932 (1988).

36. K. Wilner, R. Dombrowsky, and N. Ben-Yosef, "Correlation between LWIR and SWIR image of ground terrain," in *Sixth Meeting in Israel on Optical Engineering, Proceedings of the SPIE* **1038,** 127–136 (1988).

37. M. R. Weathersby and D. E. Schmieder, "An experiment quantifying the effect of clutter on target detection," in *Infrared Technology X, Proceedings of the SPIE* **510,** 26–33 (1984).

38. R. T. Pohlman, "Staring mosaic sensor noise sources," in *Modern Utilization of Infrared Technology II, Proceedings of the SPIE* **95,** 138–147 (1976).

39. J. W. Goodman, *Introduction to Fourier Optics,* McGraw-Hill, New York (1968).

40. O. H. Schade, Sr., "Electro-optical characteristics of television systems," in four parts, *RCA Review* **9,** 5–37 (1948); 245–286 (1948); 490–530 (1948); and 653–686 (1948).

41. H. Kusaka, "Consideration of vision and picture quality—Psychological effects induced by picture sharpness," in *Human Vision, Visual Processing, and Digital Display, Proceedings of the SPIE* **1077,** 50–55 (1989).

42. L. Levi, *Imaging Systems: Analysis and Evaluation,* Vol. 2, John Wiley & Sons, New York (1980).

43. G. Gafni, M. Azoulay, C. Shiloh, Y. Noter, A. Saya, H. Galron, and M. Roth, "Large diameter germanium single crystals for infrared optics," *Optical Engineering* **28**(9), 1003–1008 (1989).

44. M. P. Sequin and M. F. Thompsett, *Charge Transfer Devices,* Academic Press, New York (1975).

45. J. W. Coltman, "The specification of imaging properties in response to a sine wave input," *Journal of the Optical Society of America* **44,** 468–471 (1954).

46. D. L. Fried, "Resolution, signal-to-noise ratio, and measurement precision," *Journal of the Optical Society of America* **69**(3), 399–406 (1979).

47. D. L. Fried, "Resolution, signal-to-noise ratio, and measurement precision: addendum," *Journal of the Optical Society of America* **70**(6), 748–749 (1980).

48. R. N. Bracewell, *The Fourier Transform and Its Applications,* McGraw-Hill, New York (1965).

49. H. L. Resnikoff, "The duality between noise and aliasing and human understanding," in *Image Understanding and the Man-Machine Interface, Proceedings of the SPIE* **758,** 31–38 (1987).

50. R. J. Dann, S. R. Carpenter, and C. Seamer, "Sampling effects in CdHgTe focal plane arrays—practical results," in *Infrared Technology XII, Proceedings of the SPIE,* **685,** 123–128 (1986).

51. A. van Meeteren and S. Mangoubi, "Recognition of thermal images: effects of scan-line density and signal-to-noise ratio," in *Assessment of Imaging Systems: Visible and Infrared, Proceedings of the SPIE* **274,** 230–238 (1981).

52. D. J. Bradley and P. N. J. Dennis, "Sampling effects in CdHgTe focal plane arrays," in *Infrared Technology and Applications, Proceedings of the SPIE* **590,** 53–60 (1985).

53. J. N. Coltman and A. E. Anderson, "Noise limitations to resolving power in electronic imaging," *Proceedings of the IRE* **48,** 858–865 (1960).

54. O. H. Schade, Sr., "An evaluation of photographic image quality and resolving power," *Journal of the Society of Motion Picture and Television Engineers* **73**(2), 81–120 (1964).

55. F. A. Rosell and R. H. Willson, "Recent psychophysical experiments and the display signal-to-noise ratio concept," Chap. 5 in *Perception of Displayed Information,* L. M. Biberman, Ed.,

Plenum Press, New York (1973).

56. F. M. Bagrash, L. G. Kerr, and J. P. Thomas, "Patterns of spatial integration in the detection of compound visual stimuli," *Vision Research* **11,** 625–634 (1971).
57. H. L. DeVries, "The quantum character of light and its bearing upon the threshold of vision, the differential sensitivity, and visual acuity of the eye," *Physica* **10,** 553–564 (1943).
58. A. Rose, "The sensitivity performance of the human eye on an absolute scale," *Journal of the Optical Society of America* **38,** 196–208 (1948).
59. R. L. Sendall and J. M. Lloyd, "Improved specifications for infrared imaging systems," *Proceedings of the Infrared Information Symposium,* (unclassified paper), **14**(2), 109–129 (1970).
60. R. H. Genoud, Hughes Aircraft Company, El Segundo, CA, personal communication (1968).
61. J. A. Ratches, "Static performance model for thermal imaging systems," *Optical Engineering* **15**(6), 525–530 (1976).
62. M. L. Gao, M. A. Karim, and S. H. Zheng, "Device nonspecific minimum resolvable temperature difference for infrared imaging systems characterization," *Optical Engineering* **29**(8), 905–910 (1990).
63. J. G. Vortman and A. Bar-Lev, "Improved minimum resolvable temperature difference model for infrared imaging systems," *Optical Engineering* **26**(6), 492–498 (1987).
64. J. G. Vortman and A. Bar-Lev, "Optimal electronics response for parallel thermal imaging systems," *Optical Engineering* **23**(4), 431–435 (1984).
65. J. G. Vortman and A. Bar-Lev, "Dependence of thermal imaging system performance on optics and detector area," *Optical Engineering* **25**(1), 123–131 (1986).
66. H. V. Kennedy, "Modeling second-generation thermal imaging systems," *Optical Engineering* **30**(11), 1771–1778 (1991).

CHAPTER 2

Electro-Optical Imaging System Performance Prediction

James D. Howe
U.S. Army Night Vision and Electronic Sensors Directorate
Fort Belvoir, Virginia

CONTENTS

2.1 INTRODUCTION

Accurate predictions of electro-optical (EO) imager performance are important for several reasons. The predictions serve as a guide for system development by providing a means for comparing systems in their design phase, and they can be used to decide if systems with a specific design will meet established requirements. The predictions are also used in war game simulations that directly influence engagement tactics. In this way, performance predictions play a large role in the establishment of future weapons system requirements.

A complete military mission can encompass reconnaissance, surveillance, navigation, driving or pilotage, as well as attacking the target and subsequent damage assessment. Ultimately, any prediction of weapon system performance should account for the whole mission. While efforts are made to model complete missions in large war game models, the portion of the mission to be focused on in this chapter is simply the location of targets on an EO system display and their further discrimination after location.

The prediction of military target acquisition using real-time EO imaging systems, either thermal, image intensifier (I^2), or television (TV), will be described. Target acquisition is taken here to encompass the tasks of target location and discrimination to the desired level, from detection, through classification, recognition, and identification. Most of the following discussion will focus on predicting the performance of forward-looking infrared systems (FLIRs). Although many of the component descriptions and system level figures of merit to be discussed are specialized to the infrared, the modeling methods described are general and can be applied to most real-time EO imagers with only minor modification.

The target acquisition process has often been broken conceptually into two distinct parts, search and static performance. In search tasks (often referred to as dynamic or time-dependent tasks) the position of the target is not known and time to locate the target is of fundamental importance. Static performance tasks (or time-independent tasks) are those where the position of the target is approximately known and time to perform the task is not as important. It is usually assumed that the observer has as much time as desired to perform the tasks. Examples of static performance tasks are detection, recognition, and identification. Models of both types of processes are discussed. The static and search performance prediction methodologies that will be described here in most depth originated at the U.S. Army Night Vision Laboratory. The basic methodology developed at this laboratory serves as the basis for most performance models in the United States and several other countries.

Target acquisition is complex. Many models of the process have been developed, and often they are specialized to only a few military scenarios. For most models only partial validation exists due to the difficulties in carrying out realistic field tests. EO system technology is rapidly progressing and model modifications and updates are necessary in order for the predictions to reflect accurately the performance of modern systems. Consequently, the current prediction methods are bound to evolve and improve. The reader interested in further detail or historical perspective on modeling EO system performance is referred to Rosell and Harvey,[1] Lloyd,[2] Ratches et al.,[3] Biberman,[4] Overington,[5] Greening,[6] Schumaker, Wood, and Thacker,[7] and Jones et al.[8]

The units of quantities listed in Table 2.1 are those commonly found in the literature rather than SI units.

Table 2.1 Symbols, Nomenclature, and Units

Symbol	Nomenclature	Units
A_d	Detector area	cm^2
A_i	Isothermal area of target or background	m^2
A_T	Total target area	m^2
b	Optics Gaussian blur spot parameter	$mrad^2$
c	Speed of light	$cm\ s^{-1}$
D	Optics entrance pupil diameter	cm
D^*_λ	Spectral D-star	$cm\ Hz^{1/2}\ W^{-1}$
D^{**}_λ	Spectral D-double star	$cm\ Hz^{1/2}\ W^{-1}$
E	Exponent in TTPF equation	—
f_x	Spatial frequency	$cycles\ mrad^{-1}$
f_0	Measuring frequency of $S(f)$, customarily $\frac{1}{2}\tau_D$	$cycles\ s^{-1}$
f^*	3-dB point of detector response	$cycles\ s^{-1}$
f_c	Cutoff frequency of digital filter	$cycles\ s^{-1}$
f_N	Nyquist frequency	frequency units
f_{max}	Frequency of maximum electronic boost	$cycles\ s^{-1}$
f_{on}	Cut-on frequency of preamplifier	$cycles\ s^{-1}$
f_{off}	Cutoff frequency of postamplifier	$cycles\ s^{-1}$
f_s	Sample frequency	frequency units
f_t	Temporal frequency	$cycles\ s^{-1}$
F_R	Frame rate of sensor	$frames\ s^{-1}$
h	Planck's constant	$W\ s^{-2}$
H_{targ}	Target height	m
k	Boltzmann's constant	$J\ K^{-1}$
K	Boost amplitude at f_{max}	—
L	Radiance	$W\ cm^{-2}\ sr^{-1}$
L_{bar}	Angular subtense of bar length	mrad
L_{targ}	Target length	m
M	System magnification	—
n	Number of device FOVs in a system FOR	—
N	Number of resolvable cycles across a target dimension	—
N_{50}	Number of cycles required for 50% probability of performing a static discrimination task	—
$(N_{50})_D$	Number of cycles required for 50% probability of performing search detection	—
n_p	Number of detectors in parallel	—
N_s	Number of detectors in series	—
P_D	Probability of detection	—
P_0	Single glimpse probability of search detection	—
P_∞	Probability of finding target given sufficient time	—
Q_B	Background photon density	$photons\ s^{-1}\ cm^{-2}$
SNR_T	Threshold signal-to-noise ratio required for detection of a bar	—

Table 2.1 *(continued)*

Symbol	Nomenclature	Units
$S(f)$	Normalized noise power spectrum	—
t	Time	s
T_{avg}	Average temperature	K
T_{back}	Background temperature	K
t_E	Integration time of the eye	s
t_f	Mean fixation time of the eye	s
T_i	Temperature of isothermal area of target or background	K
T_{targ}	Mean temperature of whole target	K
v	Scan velocity	mrad s^{-1}
w	Angular subtense of Gaussian spot (width at 1/*e* point)	mrad
X	Physical spacing of image samples	m
Greek:		
α	System horizontal field of view (FOV)	mrad
β	System vertical FOV	mrad
γ	Atmospheric extinction coefficient	km^{-1}
Γ	Exponent in eye modulation transfer function	mrad
Δf_n	Electronic noise bandwidth	cycles s^{-1}
$\Delta\lambda$	Spectral band	μm
ΔT	Temperature difference	°C
Δx	Detector instantaneous field of view (IFOV) in azimuth	mrad
Δy	Detector IFOV in elevation	mrad
Δx_{sample}	Angular subtense of a single sample along the x direction	mrad
Δy_{sample}	Angular subtense of a single sample along the y direction	mrad
ε	Charge transfer efficiency	—
η_{cs}	Cold-shield efficiency	—
η_{ovsc}	Overscan ratio	—
η_q	Quantum efficiency	—
η_{scan}	Scan efficiency	—
λ	Wavelength of light	μm
λ_{co}	Cutoff wavelength of detector	μm
λ_M	Mean wavelength of EO system spectral passband	μm
π	3.14159...	—
$\sigma_{(\ )}$	Standard deviation of subscripted variable	(subscripted variable units)
$\sigma^2_{(\ )}$	Variance of subscripted variable	(σ units)^2
τ	Mean acquisition time	s
τ_A	Atmospheric transmission	—
τ_D	Dwell time for scanning detector	s
τ_{FOV}	Mean acquisition time for FOV search	s
τ_{FOR}	Mean acquisition time for field of regard (FOR) search	s
τ_o	Optics transmission	—

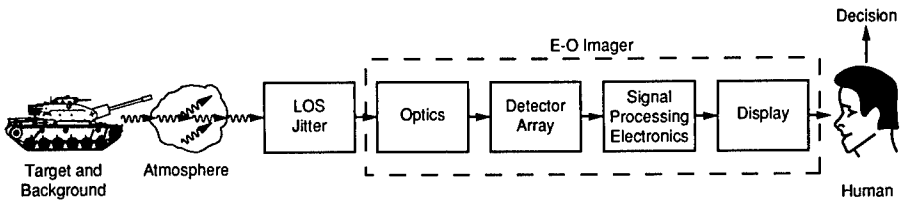


Fig. 2.1 The elements of target acquisition.

2.2 THE ELEMENTS OF TARGET ACQUISITION

2.2.1 General Description

A simplified diagram illustrating the important components of target acquisition is presented in Fig. 2.1. The radiance from the scene is a complex function of target and background material properties, heat sources, and atmospheric conditions. The emitted or reflected radiation traveling toward the imager passes through the atmosphere where some of the radiance is absorbed or scattered. Background radiation can also be scattered into the same path toward the imager. Turbulence along the path can cause a difference in the amount of refraction of adjacent rays causing distortions in the image formed by the imager.

Optical elements gather the radiation, forming an image of the scene at the image plane of the system. A single detector or array of detectors transduces the incident radiation into electrical signals, which are processed and formatted for display. The image can be sampled in various ways: A single detector or an array of detectors can rapidly scan across an image, or a two-dimensional array can simply stare at the scene. A human views the image and tries to make a decision as to the existence, location, and identity of a target. A model of target acquisition requires descriptions of all parts of the imaging chain as well as the decision process.

2.2.2 Complexities of Target Acquisition

Table 2.2 lists several factors and subfactors pertinent to target acquisition. Although the list is by no means complete, its length underscores the complexity of target acquisition. No single model of the target acquisition process could possibly account for all of the factors listed; because of this, many models are specialized to a particular handful of scenarios and system types. Frequently they incorporate simplifying assumptions, and they are usually validated for only a small portion of the possible missions or situations. Using a model to predict performance for scenarios where the model is not validated or specialized can lead to very inaccurate predictions.

2.2.3 Target Acquisition Vocabulary

> The three most commonly used and confused terms employed to describe the visual problems of targeting are detection, recognition, and identification.[9]

As Bliss[9] implies in the preceding comment, no universally accepted set of definitions of target acquisition terminology exists in the literature. Although

Table 2.2 Factors and Subfactors Influencing Target Acquisition Performance

Major factor	Subfactors	
Mission-related factors	Reconnaissance Surveillance Navigation Vectored attack	Attack targets of opportunity Attack prebriefed targets Damage assessment
Target and background	Type Size Shape Contrast Temperature Reflectance Emittance	Motion Shadow Masking Camouflage Clutter Cues Distinctiveness
Environment	Visibility Cloud cover Sun angle Solar loading Diurnal variation Seasonal variation	Illumination level Attenuation Transmittance Scattering Path radiance Turbulence
Platform	Crew size and interaction Vibration environment Target exposure time	Altitude Range Speed
Sensor parameters	Sensor type Spectral band Field of view Resolution Dynamic range Sensitivity	Sample spacing Interlace Frame rate LOS panning capability Field of regard Search field size
System display	Screen angular subtense Aspect ratio Color SNR Resolution Phosphor persistence Gray-scale rendition	Gamma Interlace Contrast ratio Veiling glare Surround lighting Flicker Frame rate
Observer variables	Training Motivation Experience Prebriefing Stress Task load	Visual acuity Search pattern Fatigue Age IQ

dictionaries of military terms have been compiled,[10] their definitions are often inappropriate for the discussion of target acquisition prediction. The literature of target acquisition is often misleading or inconsistent due to the lack of a standard vocabulary. A glossary of search and target acquisition nomenclature has been developed[11] in an effort to remedy this situation. The definitions presented in the glossary are the most appropriate for description of target acquisition tasks, and they will be used throughout this chapter.

Following the glossary,[11] definitions of some key terms that apply to the target acquisition process will be presented. A distinction that is accentuated in the glossary is the difference between static and search performance tasks. Some important time-dependent acquisition definitions from Ref. 11 follow.

- *Target acquisition:* All those processes required to locate a target image whose position is uncertain and to discriminate it to the desired level (detection, classification, recognition, identification). The target acquisition process includes the search process at the end of which the target is located and the discrimination process at the end of which the target is acquired.
- *Search:* The process of visually sampling the search field using either a device display or the unaided eye in an effort to locate or acquire targets.
- *Display search:* The search process limited to one display field. This is also called FOV search.
- *Localize:* To establish the position of the target image through the search process.

Important static target acquisition task definitions, also from Ref. 11, follow and examples are given:

- *Discrimination:* A process in which an object is assigned to a subset of a larger set of objects based on the amount of detail perceived by the observer.

For example, discrimination at the recognition level for an army application would correspond to the assignment of an object to a subset of objects that contains either tanks, trucks, jeeps, or armored personnel carriers.

- *Detection:* A temporally unconstrained perception of an object image or a target image at a specific display field location.
- *Pure detection:* Detection where the two alternative choices are (1) that something is present or (2) that nothing is present.

An example of pure detection would be the perception of a helicopter against a bland sky or a hot tank against an uncluttered background. Pure detection plays a role in cluttered environments as well since it describes the observer's ability to state whether an object is present or not.

- *Discrimination detection:* A detection where the two alternative choices are (1) the object is a target or (2) the object is something else, such as a natural scene element.

The scenario to which one refers when performing discrimination detection is one where some shape or form needs to be discerned in order for military objects to be distinguished from clutter objects in the background. Simply seeing a blob in a cluttered image does not usually indicate the presence of a target with a high degree of confidence. For performance of discrimination detection, the blob in question must be compared with other blobs, and features of the blob must set it apart from the others indicating the need for a resolving capability for this type of detection.

Another type of detection commonly discussed in the literature is military detection. It is the determination that an object in an image is of military interest. If the simple determination that an object is present is enough to indicate the object is of military interest, for example, seeing a point object in the sky, then pure detection is equivalent to military detection. If, on the other hand, the observer has to discriminate between a particular object and other objects in the FOV in order to determine which is of military interest, then discrimination detection is equivalent to military detection.

- *Orientation, classification, recognition:* Intermediate levels of discrimination used variously in the services.

Classification for army scenarios has sometimes been chosen to mean the discrimination between wheeled and tracked vehicles. Discrimination tasks will vary in difficulty depending on how the target subsets are defined, i.e., how many targets are in each subset, and how similar the target appears to other targets within the same subset, as well as to targets in different subsets.

- *Identification:* The most detailed level of discrimination generally used to describe the acquisition of military targets.

For example, a target is perceived to be a particular type of tank, an M1 or a T72.

2.3 THE IMAGING CHAIN

2.3.1 Target and Background

The spatial and spectral distributions of radiance from a scene are complicated functions involving the reflective, absorptive, and emissive properties of the target and background. For imagers operating in the visible spectrum, a scene element's emissive properties usually do not play a role, and the radiance distribution is determined through knowledge of properties of illuminating sources such as the sun or moon and the reflective properties of scene objects. In the far infrared the contribution of reflected radiation is negligible, and only the emissive properties of targets and backgrounds are important. In the near and mid infrared both the reflective and emissive properties of the targets and backgrounds can be significant.

An EO imager responds to spatial distributions of radiance or "signatures" from targets and backgrounds. This signature is a function that specifies the radiance from every point in the scene projected onto a plane or image perpendicular to the system line of sight (LOS). While the variation from point to point in a signature can be due to many things besides temperature variation, it is convenient to simplify the situation and assume that all signal variations are due to differences in the temperatures of blackbody sources. Signatures can then be specified using radiation temperature, which is the temperature of a blackbody that gives the same total radiance in the particular waveband of interest as the scene element.

It is common practice to specify a temperature difference between target and background for use in a performance prediction model. Since a ΔT of, for

example, 5°C will correspond to different amounts of power depending on the actual target and background temperatures, it is important, for consistency, to reference the ΔT to a particular temperature, T_{ref}, chosen in practice[7] to be 300 K. If this is done properly, each degree of temperature difference will correspond to the same power difference for a particular imager operating in a specific spectral band.

Referenced temperature differences can be calculated from the power difference, or from the difference between radiometric temperatures, using

$$\Delta T = \frac{\Delta L}{\left(\frac{\partial L}{\partial T}\right)_{T=T_{ref}}} , \tag{2.1}$$

where

L = $\int_{\Delta\lambda} L_\lambda \, d\lambda$, the blackbody radiance ,

$\Delta\lambda$ = the spectral band of the imager ,

$\left(\frac{\partial L}{\partial T}\right)_{T=T_{ref}}$ = the derivative with respect to temperature of the total blackbody radiance evaluated at the reference temperature of 300 K.

In the infrared, several factors can influence the target signature. The atmospheric conditions, past history (running engine or not), solar loading, camouflage, and the local background will all influence the target signature, resulting in a thermal signature that is highly variable.

Two different types of signatures are normally discussed in the context of imaging: An *inherent* signature is a signature that has not been affected by the atmosphere, and an *apparent* signature refers to a signature at the imager after having passed through the atmosphere.

Modeling the Target and Background. The description of the target and background in most target acquisition models is extremely simple. Two parameters have normally been used to describe the target: its size, expressed in terms of a *critical dimension,* and its contrast with respect to background. The focus here will be the infrared in which the contrast takes the form of a temperature difference.

In the particular prediction methodology discussed here in most detail, and in most other methodologies, the target shape is taken to be a rectangle, not an object with a complex shape. Often the critical dimension is taken to be the minimum dimension of the projected target, which usually turns out to be the target height H_{targ} for ground-based tactical vehicles. The larger dimension of the rectangle L_{targ} is chosen so that the rectangle has an area, $A_T = H_{targ} \times L_{targ}$, equal to the total projected area of the actual target. Other models use $A_T^{1/2}$ as the critical dimension.

In many models the target is assumed to have a uniform temperature, an approximation that will apply well to some real targets without internal struc-

ture and not as well to targets with a lot of visible internal detail. A target's internal structure may not be visible at the long ranges considered when discussing pure detection, but for shorter ranges and higher order discrimination tasks, the influence of internal "cues" on discrimination performance can be significant.

In the infrared, the inherent target signature is divided into i subareas A_i of equal temperature T_i, and the target's area weighted average temperature is calculated using

$$\overline{T}_{\text{targ}} = \frac{\sum_i A_i T_i}{\sum_i A_i} \quad . \tag{2.2}$$

The area weighted average temperature of a region of background directly adjacent to the target is calculated similarly. For targets where the temperature differences between each target subarea and the average background are all positive, it is appropriate to apply the formula for ΔT given in Ref. 3:

$$\Delta\overline{T} = \overline{T}_{\text{targ}} - \overline{T}_{\text{back}} \quad . \tag{2.3}$$

This procedure runs into trouble sometimes for targets with prominent hot spots. Many tactical targets have been found to have some areas that are hotter and some that are colder than the average background temperature. Application of Eqs. (2.2) and (2.3) leads to the possibility of targets having a total inherent temperature difference of zero even though portions of the target have high and low temperatures. Application of the performance prediction methodology to be discussed could then result in very low or even zero probability of detection although some parts of the target were very bright.

Characterization of the detail within a target is normally not used in acquisition prediction, although it may, sometimes, be that detail which is used by an observer to perform discrimination. Several variations to the ΔT definition described earlier have been proposed; however, none has been shown to improve the accuracy of performance prediction. The investigation of target descriptions that can accurately account for the presence of hot spots on the target continues to be an active area of research.

2.3.2 Atmosphere

Radiation emitted by a scene is attenuated as it passes through the atmosphere to the imager. The attenuation is due to scattering and absorption by gas molecules, aerosols, dust, fog, rain, snow, or man-made obscurants. Not all the radiation reaching an imager comes directly from the scene since radiation can be scattered from a source outside the scene into the FOV of the imager. In the infrared the atmospheric path along the imager LOS can also act as a source of radiance that reaches the imager.

The random irregularities in the index of refraction of the atmosphere, called *turbulence,* that are caused by local variations in the temperature, pressure, or humidity can cause distortions in an image, motion of the whole image or pieces of the image, a blurring of the image, or temporal fluctuations called

scintillations. Turbulence often exhibits a strong diurnal cycling, and is distinctly influenced by the presence of sources of strong upwelling along a particular atmospheric path. For an EO imaging device, the relation between the integration time of each pixel in the image and the temporal extent of the local variations will determine whether the turbulence manifests itself mainly as blurring of a single image, or as an image distortion that changes from one image to the next in an image sequence.

Modeling Atmospheric Effects. The absorption, scattering, and path radiance of the atmosphere can be accounted for by using a computer program[12] such as LOWTRAN 7. The atmospheric transmission through a particular pathlength is calculated for the spectral band of the imager and is applied to the target ΔT. The resulting apparent ΔT is then used in the prediction methodology.

Although turbulence in the atmosphere can result in blurring and distortion of the final image, it has not usually been accounted for in most tactical prediction methods. If turbulence is manifested mainly as a distortion of the image, its influence could not be accounted for simply by using another modulation transfer function (MTF).

2.3.3 FLIR Systems

Figure 2.2 illustrates the system architecture for a typical "first-generation" tactical FLIR system. The infrared scene is scanned across a parallel array of detectors inside a cryogenic dewar. Signals from each detector are fed in parallel to visible light-emitting diodes (LEDs), the light from which is scanned across the face of a vidicon that produces the signal for display. More recent

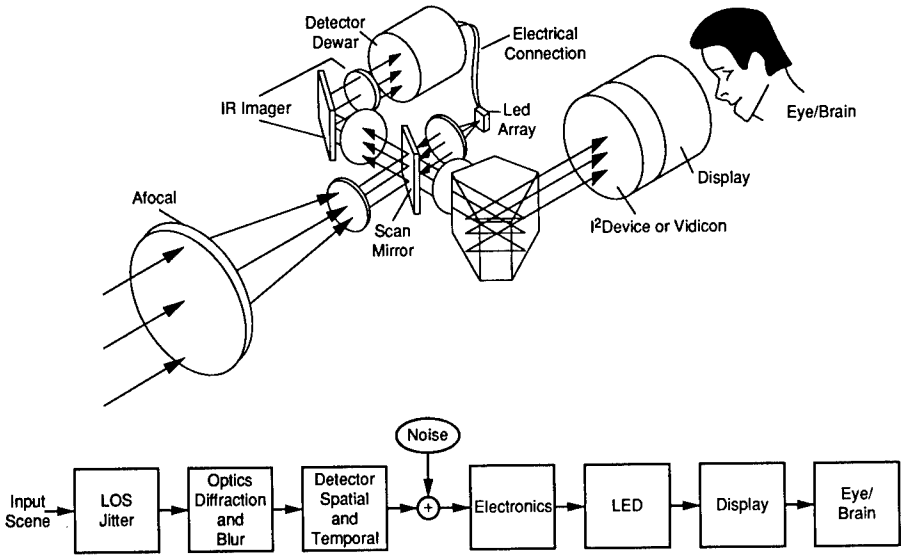


Fig. 2.2 Typical "first-generation" scanning FLIR system architecture and associated signal processing block diagram.

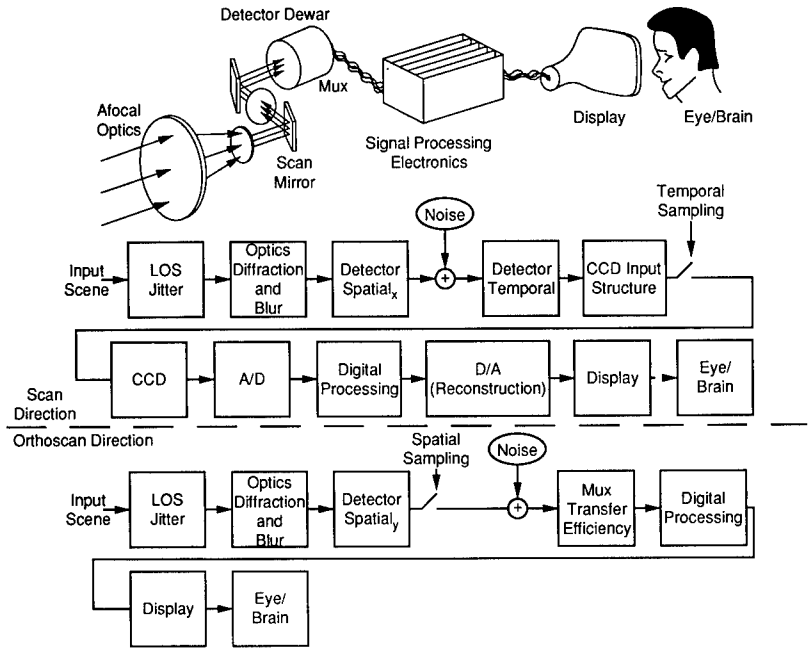


Fig. 2.3 Typical "second-generation" FLIR system architecture and associated signal processing block diagrams, along and perpendicular to the scan direction.

imagers utilize an electronic multiplexing scheme to circumvent the optical multiplexing performed by the LEDs and vidicon.

The bottom half of Fig. 2.2 shows a block diagram indicating the important processing steps of the imager for the purpose of performance modeling. As has been the case until recently, only system transfer characteristics along the direction of scan are considered.

A modern "second-generation" FLIR system architecture is depicted in Fig. 2.3. The continuous image formed by the optics is made discrete, or is sampled, by the array of detectors. The regular array spatially samples the continuous image perpendicular to the scan direction, producing a parallel set of spatially discrete analog signals. This same spatial sampling perpendicular to scan exists for the first-generation systems depicted in Fig. 2.2.

For the scanning system in Fig. 2.3, the process of integration of a signal for a fixed time during scanning also takes a continuous signal and samples it, resulting in a series of discrete pulses, this time along each scan line. The signal, which is now a discrete analog signal in two dimensions, is read off the focal plane by a charge transfer mechanism and is passed to an analog-to-digital (A/D) converter that assigns a binary value to (or quantizes) each analog sample. Digital processing such as histogram modification or filtering can be applied before the digital signal is converted back to an analog signal and finally displayed. Block diagrams of processing components both along the scan direction and perpendicular to it are shown paralleling recent trends in FLIR system performance modeling that attempt to account for system transfer characteristics in both directions.

Two dimensional (2-D) staring arrays are being proposed for more applications since the size of the arrays is increasing while the quality of the imagery they produce is improving. The lack of a scanner leads to simpler, smaller, more reliable, and more robust systems.

Modeling FLIR Systems. System component transfer functions, which for linear systems are the ratios of output frequency spectra to input spectra, play a central role in the system level figures of merit described in the next section. Therefore, an appreciation of typical component transfer characteristics is necessary for a complete understanding of system performance prediction. (See Chapter 1 in this Volume, and Chapters 2 and 3 in Volume 5 of this handbook.)

For the purposes of performance modeling of EO systems, several simplifying assumptions are usually made about an imaging system. Naturally the accuracy of the resulting performance predictions relies on how large a departure from reality is involved in these assumptions. For systems where the departure is not large, the decrease in complexity of the problem afforded by the assumptions is usually significant.

EO system components are assumed to be linear, or at least to be operating in a linear region. The systems are also assumed to be shift invariant, or isoplanatic. [The response of a system to a point source, the system point spread function (PSF) or impulse response, does not change appreciably from point to point in an image.]

The Fourier transform $\mathcal{F}$ of the PSF is the optical transfer function (OTF), a function that describes the ability of a system component to transfer an object's spatial distribution of light to an image. The modulus of the OTF is the modulation transfer function (MTF) and the argument is the phase transfer function (PTF):

$$\mathcal{F}(\text{PSF}) = \text{OTF}(f_x) = \text{MTF}(f_x) \exp[i \cdot \text{PTF}(f_x)] \ . \tag{2.4}$$

The MTF expresses the amplitude response of a system or system component to a sine-wave input. For many typical system components that act as low-pass filters, the MTF describes the attenuation of contrast with increasing frequency. The PTF describes the phase response due to a shift or an asymmetry in the PSF. If each component of a system is linear and shift invariant, then the amplitude response of the total system can be modeled using a single MTF that is the product of the component MTFs.

In frequency space an input spectrum is multiplied by the transfer function of a system yielding an output spectrum. Working in frequency space and multiplying transfer functions is considerably simpler than working in direct space where the signal is convolved by the impulse response of the system resulting in the output signal. The use of linear systems theory and Fourier analysis has shaped not only the analysis of EO systems, it has also influenced system measurement methodology and system performance summary measures.

Until recently, performance prediction for FLIR systems has involved the calculation of the system transfer function in a single dimension only, chosen to be the scan direction. Perhaps this has been the case because of the development of a performance prediction methodology based on the calculation of the number of resolvable cycles across a single dimension of a target. The

prevalence of the unidimensional approach may have resulted partially from the difficulty of dealing with the sampling of an image in scanning systems perpendicular to the scan direction. Recent efforts have been directed toward the utilization of component descriptions in two orthogonal dimensions in a prediction methodology.[13–15]

For the model discussed here in the most detail, we have assumed that there is only one noise term, or rather that one noise source, usually the detector, dominates all others. Although this may be the case in many systems, keep in mind that many components of the system can act as noise sources, and different system components can dominate the system noise power spectrum in different frequency bands.

Many component characterizations involve analytical functions of spatial or temporal frequency. For scanning systems the two are related by

$$f_t = vf_x \ , \tag{2.5}$$

where v is the scan velocity (mrad s^{-1}), f_t is temporal frequency (cycles s^{-1}), and f_x is spatial frequency (cycles $mrad^{-1}$) along the scan direction.

The following descriptions of system components and signal processing are specific to thermal imaging systems. Transfer functions of typical first-generation FLIR systems are described first. System level performance measures are then presented, after which static performance prediction methods for typical first-generation systems are outlined. Some modern system components that are not treated by the first-generation static performance methodology are then described as motivation for a discussion of recent efforts to update the static performance modeling procedures.

Line of Sight Jitter. For the purposes of system performance modeling, the influence of several causes of image position uncertainty are lumped together into one term. Vibration of the sensor with respect to the scene is combined with scanner position uncertainty and other causes of image jitter. The observer's temporal integration of several frames of imagery randomly displaced with respect to one another has a blurring effect most often accounted for by a Gaussian of the form

$$H_{\mathrm{LOS}}(f_x) = \exp[-2(\pi\sigma_{\mathrm{LOS}}f_x)^2] \ , \tag{2.6}$$

where σ^2_{LOS} is the variance of the total vibration from all sources.

Optics. For a diffraction-limited system with a clear circular aperture, the MTF of the optics is given by

$$H_{\mathrm{diff}}(f_x) = \frac{2}{\pi}[\cos^{-1}X - X(1 - X^2)^{1/2}] \ , \tag{2.7}$$

where

X = f_x/f_c
f_c = D/λ_M, the cutoff spatial frequency
λ_M = mean optical wavelength of the system
D = clear aperture diameter.

For apertures that are not circular or are partially occluded, the form of the MTF can be significantly different. For optical systems that are not diffraction limited the detailed form of the optical MTF will depend on the specific aberrations present in the system. See Goodman[16] for more details. The MTF due to the blur is sometimes assumed to be a Gaussian of the form

$$H_{\text{blur}}(f_x) = \exp(-bf_x^2) \ , \tag{2.8}$$

where b is related to the width of the Gaussian spot by

$$b = \frac{\pi^2 w^2}{4} \tag{2.9}$$

and w is equal to the blur spot diameter at the $1/e$ point.

Computer implementations of many performance models usually allow the input of system component MTFs of arbitrary shape. This allows inclusion of a measured optical system MTF or an MTF predicted using a ray tracing program often leading to greater accuracy than the Gaussian blur assumption.

Detector. It is usually assumed that the PSF of the detector is a rect function in direct space. Then

$$H_{\text{det}}(f_x) = \text{sinc}(\pi f_x \Delta x) \ , \tag{2.10}$$

where Δx is the instantaneous field of view (IFOV), or angular subtense, of the detector in the direction being considered. Depending on detector fabrication techniques and material properties such as carrier diffusion, the detector spatial response can have more rounded edges and sloping sides than a rect function. The temporal response of detectors to incident radiation is not instantaneous and, for scanning systems, this results in a blurring in the scan direction given by

$$H_{\text{temp}}(f_t) = \left[1 + \left(\frac{f_t}{f_t^*}\right)^2\right]^{-1/2} \ , \tag{2.11}$$

where f_t^* is the 3-dB point of the detector temporal response.

A measure of the detector sensitivity that is commonly used in FLIR system performance calculations is detector specific detectivity, D^{**}. For an ideal background-limited detector with no cold shielding exposed to 2π steradians of background radiation,

$$D^{**}(\lambda) = \eta_q(\lambda)^{1/2} \frac{\lambda}{2hc} Q_B^{-1/2} \ , \tag{2.12}$$

where

$D^{**}(\lambda)$ = detectivity at λ (cm $\text{Hz}^{1/2}$ W^{-1})

$Q_B = \int_0^{\lambda_{co}} Q_B(\lambda,T_B)\, d\lambda$

λ_{co} = cutoff wavelength

$Q_B(\lambda,T_B) = \frac{2\pi c}{\lambda^4}\left(\exp\frac{ch}{\lambda k T_B} - 1\right)^{-1}$ = spectral photon flux density from a background of temperature T_B (photons s^{-1} cm^{-2})

k = Boltzmann's constant = 1.3805×10^{-23} J K^{-1}

h = Planck's constant = 6.6256×10^{-34} W s^{-2}

c = speed of light = 2.9979×10^{10} cm s^{-1}.

The detectivity of a cold-shielded detector, D^*, assuming a small effective cold-shield $f/\#$, f_{cs}, is

$$D^* = 2(f/\#)\eta_{cs}D^{**} ,$$

where

$$\eta_{cs} = f_{cs}/(f/\#) .$$

Electronics. The signal from the detector is usually ac coupled in first-generation FLIR systems in order to get rid of the large background pedestal. This ac coupling is modeled as a simple RC high-pass filter where

$$H_{\text{preamp}}(f_t) = \frac{f_t/f_{\text{on}}}{[1 + (f_t/f_{\text{on}})^2]^{1/2}} . \tag{2.13}$$

The cut-on frequency f_{on} is the 3-dB point of the preamplifier response.

The postamplifier is modeled as an RC low-pass filter and is given by

$$H_{\text{postamp}}(f_t) = \frac{1}{[1 + (f_t/f_{\text{off}})^2]^{1/2}} , \tag{2.14}$$

where f_{off} is the 3-dB point of the postamplifier response.

Electronic boost or aperture correction is sometimes used to flatten the system response out to a frequency where the noise level becomes appreciable in relation to the signal. Boost transfer characteristics are given by

$$H_{\text{boost}}(f_t) = 1 + \left(\frac{K-1}{2}\right)\left[1 - \cos\left(\frac{\pi f_t}{f_{\text{max}}}\right)\right] , \tag{2.15}$$

where f_{max} is the frequency at which the boost is maximum and K is the boost amplitude at f_{max}. The boost operation has normally been applied only along the scan direction.

Display. The display acts to reconstruct the signal for the observer. If we assume the display to be a CRT with a Gaussian spot shape, the display MTF is given by

$$H_{\text{display}}(f_x) = \exp\left[-\left(\frac{\pi w f_x}{2}\right)^2\right], \tag{2.16}$$

where w is the width of the Gaussian spot at the $1/e$ point. Flat-panel displays are becoming attractive alternatives to CRT displays due to their light weight, low power consumption, and small volume. Most flat-panel display technologies employ a regular grid of rectangular (or square) pixels. Thus, instead of a Gaussian spot as for the CRT, the display component PSF is more nearly a rect function, resulting in a sinc function MTF.

2.3.4 The Observer

The human visual system, consisting of the eye and parts of the brain, is the final component in a target acquisition system. The human plays the central role in the target acquisition process, and a knowledge of the functions and capabilities of this final component in the system is necessary in order to appreciate completely the role of the observer in performance prediction methodologies. A short discussion will give an understanding of the complexity of the human visual system and the contrasting simplicity with which the eye is usually treated in performance models. See Refs. 17 through 20 for more complete descriptions of the anatomy and functions of the visual system.

Sensitivity.[21,22] The human eye is able to respond to illumination levels that vary over 10 orders of magnitude due to a combination of different mechanisms. Two different receptor types, rods and cones, located in different portions of the retina (see Fig. 2.4) respond to different intensities, contributing to the

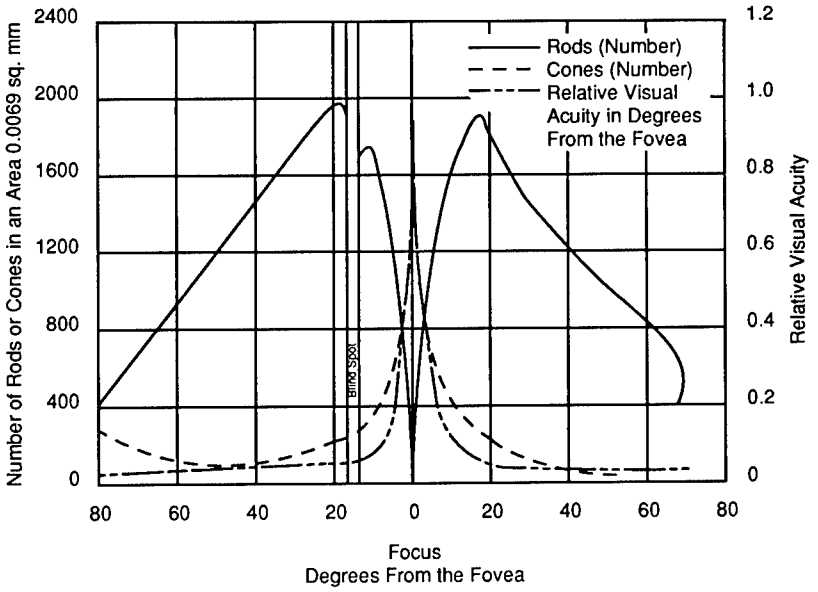


Fig. 2.4 The number of rods and cones (left axis) and relative visual acuity (right axis) versus angle from the fovea (from Refs. 19 and 120).

large dynamic range of the eye. The cones, which are located mostly within a small angle subtending approximately 1.5 deg about the visual axis called the fovea, provide color and detail information. The fovea is the area used to interrogate an EO system display, fixate on a target, and provide detail information. The cones function primarily at high, or photopic, illumination levels. The rods respond at relatively low, or scotopic, illumination levels where the cones are not sensitive. The response of both types of receptor to input intensity is roughly proportional to the logarithm of the input illumination, another feature that contributes to the eye's dynamic range.

A relatively small portion of the ability of the visual system to adapt to different intensities and to remain sensitive over a wide range of illumination levels is accomplished simply by opening and closing the pupil. One main utility of the pupil is its rapid response. A much larger part of the adaptation is handled by the variation of the receptor response to different levels of illumination. The rods in the periphery are more sensitive, and they become saturated with increased background levels. The cones in the fovea do not appear to saturate and they rapidly adapt to changes in background level.

Acuity. The ability of an observer to discriminate fine detail in an image, the acuity, is influenced by several factors.[23] Acuity is usually defined as the reciprocal of the threshold visual angle in minutes of arc as measured using a standard test pattern. Several different test patterns are commonly used, i.e., bar patterns, sine-wave gratings, and Landolt Cs. Threshold is defined as the point at which an observer or ensemble of observers is just able to perform a task such as detection. It is usually chosen as a particular point (50% or 75%) on a psychometric function (a relation between stimulus level and proportion of correct responses). As shown in Fig. 2.4, the acuity of the eye closely follows the density of cones on the retina, being greatest at the fovea and falling off rapidly in the periphery. For this reason, the luminance level strongly affects acuity, being greatest at high light levels where the cones operate. Acuity decreases as pupil size increases, and the outer portions of the lens of the eye, which contain more aberrations, are used to form the image.

The acuity of the human visual system is commonly expressed using sine-wave contrast threshold measurements for varying spatial frequencies. Threshold modulation averaged over a large group of test subjects is plotted in Fig. 2.5 as a function of spatial frequency.[24] Also indicated are limits that include 90% of the test population results. It is interesting to note that the visual system modulation sensitivity is best near 3 to 5 cycles deg^{-1} and requires more modulation at lower as well as at higher spatial frequencies. The visual system is not able to spatially integrate signals completely from the large (low-frequency) targets as it is from intermediate and high spatial frequency targets.

Spatial Summation.[25] The eye-brain is able to sum an input stimulus over space and time. This capability profoundly influences the observer's target acquisition capabilities, and it is reflected in the system performance measures discussed in the following section. Spatial summation has been investigated by measuring the threshold detectability of simple stimuli of varying spatial and temporal extent.

Figure 2.6 indicates how threshold intensity varies with target area for foveal targets presented for 12.8 ms. The curve exhibits a region where thresh-

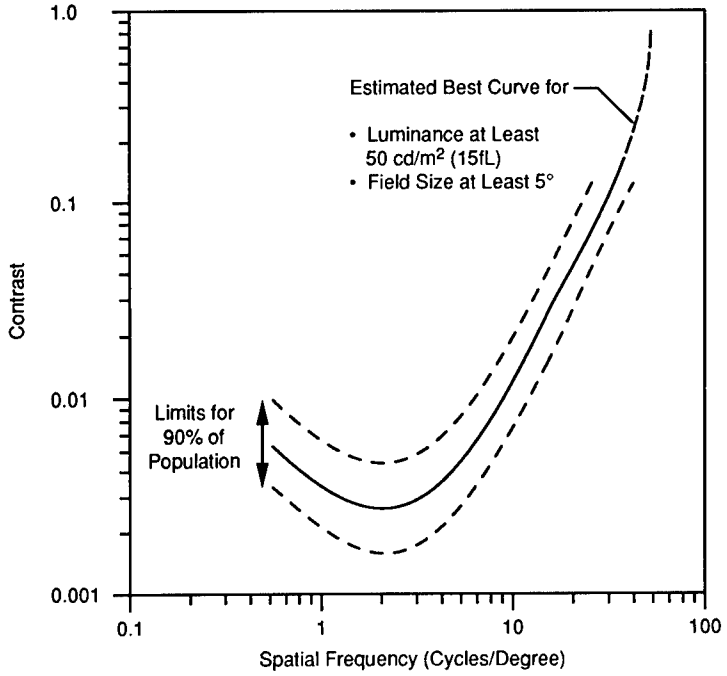


Fig. 2.5 Spatial contrast threshold function for sine-wave targets. Dashed lines indicate limits for 90% of the test population. (From Ref. 24, as developed by S. J. Briggs.)

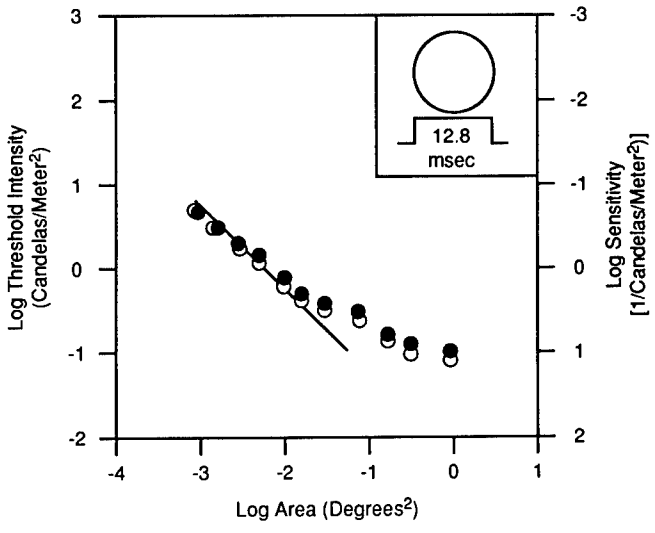


Fig. 2.6 Measurements indicating spatial summation in the human visual system. Log minimum detectable light intensity versus log target area in degrees^2. Red target disk is presented at fovea of dark adapted eye for 12.8 ms. Filled and open circles represent data from two observers. (From Ref. 22, based on data from Ref. 121.)

old intensity varies inversely with area (slope of -1) indicating complete spatial summation. For somewhat larger targets, there is a region where the slope is between -1 and 0, indicating only partial summation. For a large range of illumination, therefore, dimmer targets are easier to see if they are larger. The curve finally flattens for large targets, indicating a region of little or no spatial summation.

The region of complete spatial summation increases in size with increasing angle from the fovea. For targets fixated on the fovea, complete spatial summation occurs only for small targets (approximately 6 or 7 min of arc). Targets up to 1 deg in diameter are completely summed when viewed 15 deg from fixation. For targets fixated nearly 40 deg from the fovea, measurements[20] have indicated complete spatial summation of target diameters as large as 4.7 deg.

Temporal Summation.[26] Experimental evidence indicates that the eye-brain temporally integrates energy from the image for a period that depends on the intensity of the target, its background or adaptation level, target spatial frequency, and a host of other factors.[27] Figure 2.7 shows an observer's threshold detection capability for single pulse target disks of varying duration, spatial extent, and background intensity. Up to a pulse duration called the *critical duration* the detection threshold varies inversely (sensitivity increases) with pulse duration. This indicates a region of complete temporal summation similar to the case for spatial summation. For pulses that are longer than the critical

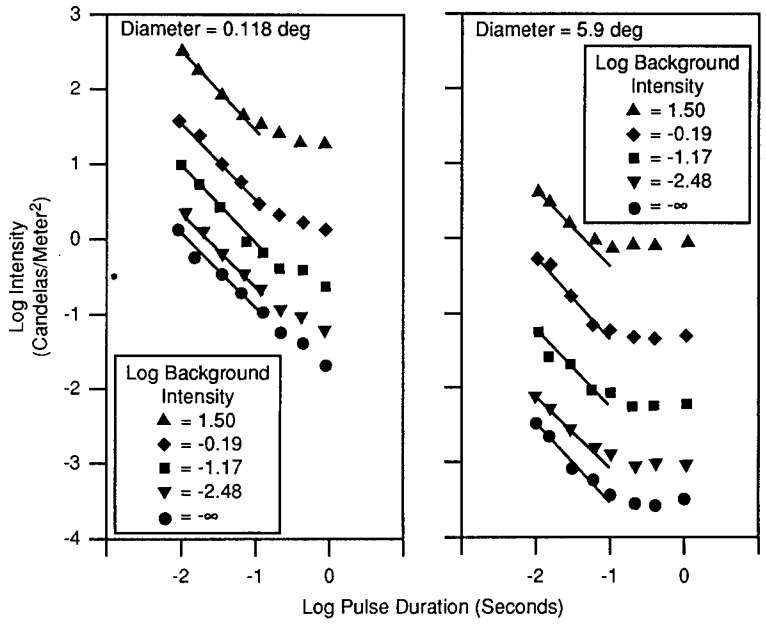


Fig. 2.7 Measurements indicating temporal summation. Log of the threshold intensity versus log of pulse duration for two different size targets at five different background levels. (From Ref. 26, based on data from Ref. 25.)

duration, the curves begin to level off, indicating a region of incomplete summation and, finally, when the slope is zero, a region of no summation analogous to the situation for spatial summation. For smaller targets, the region of complete temporal summation extends to longer durations than for larger targets.

Modeling the Observer. The properties of the human visual system are reflected in many places in performance prediction; they are not simply another MTF in the image processing chain. Spatial and temporal summation as well as assumptions about how the human performs detection are essential facets of any static performance model. Search models sometimes specifically include peripheral visual acuity and eye movement characteristics.

In the prediction methods to be discussed, the images of tactical vehicles or bar patterns used for system testing are assumed to be completely summed spatially. In reality, some targets of interest can be large enough that they are incompletely summed. The spatial summation of signal and noise is assumed to behave like a matched filter that has been matched to the image of the target after having passed through the imager. Similar models[1] have incorporated what was termed a *synchronous integrator* for the same purpose.

Although the time interval over which the eye-brain totally sums an input scene has been shown to vary with illumination level, for the purposes of target acquisition predictions using FLIRs,[28] the eye has typically been assumed to sum for a single, fixed integration time t_E routinely chosen to be between 0.05 and 0.2 s. Since typical display illumination levels for FLIRs do not vary over several orders of magnitude, the single fixed integration time is an adequate approximation. For typical imaging systems that produce 25 to 30 images per second, the length of t_E indicates that the human integrates several frames of information from an EO display.

Based on earlier work by de Vries,[29] Rose,[30] and Schade,[31] Rosell[1,32] applied fluctuation theory concepts to the detectability of objects in noisy television imagery. De Vries postulated that in order for an object to be detectable the signal to noise ratio (SNR) of the image must exceed some threshold value. Rosell and Willson[33,34] related the video SNR (SNR_V) to a displayed image SNR (SNR_{DI}). In transforming from SNR_V to SNR_{DI} they accounted for the spatial and temporal integration properties of the human visual system but did not account for the eye MTF or any noise sources in the visual system:

$$\text{SNR}_{DI} = \text{SNR}_V\left(\frac{2t_E F_R A_T}{A_D}\right)^{1/2} . \tag{2.17}$$

Here A_T is the target area, A_D is the area of the display, F_R is the frame rate of the sensor, and t_E is the integration time of the eye.

Rosell and Willson performed numerous psychophysical tests using noisy imagery, the results of which demonstrated that the probability of detection was a strong function of SNR_{DI}. Figure 2.8 shows results of testing the detectability of various size rectangles in white Gaussian noise, indicating a 50% probability of detection for an SNR_{DI} of approximately 2.8. This is one of the fundamental relations used in the pure detection prediction methodology to be discussed in a later section. A thorough discussion of the influence of noise on the human visual system is found in Peli.[35]

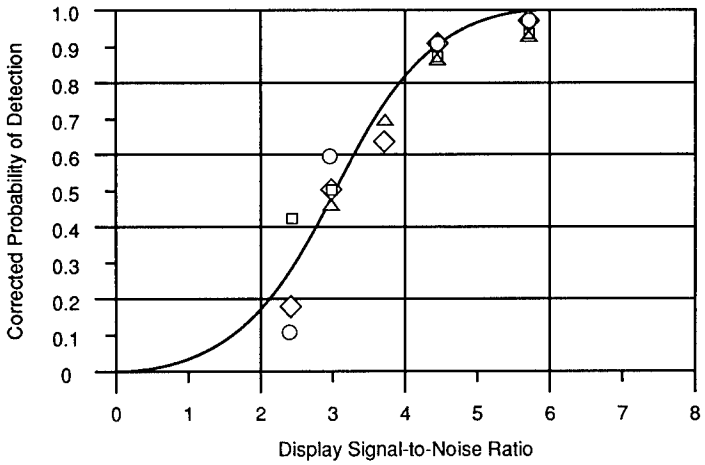


Fig. 2.8 Corrected probability of detection versus SNR_{DI} required for rectangular images of size (○) 4 × 4, (□) 4 × 64, (△) 4 × 128, and (◇) 4 × 180 scan lines (from Ref. 33).

The human visual system is a nonlinear device and strictly speaking should not be described by an MTF.[36] Again, an approximation is made for modeling, the objective being to describe the human within the confines of a linear system framework. Many researchers have commented on the suitability of a particular eye model for performance prediction, the appropriate choice still being a topic of current debate (see Holst and Taylor[37] for an example).

The MTF of the visual system as determined by Kornfeld and Lawson,[38] which is used in the static performance model to be discussed, is given by

$$H_{\text{eye}}(f_x) = \exp[-(\Gamma f_x)] \ , \tag{2.18}$$

where Γ is plotted in Fig. 2.9(a), and Fig. 2.9(b) illustrates the visual system MTF for several light levels.

2.4 SYSTEM PERFORMANCE MEASURES

The description of system level figures of merit for EO imagers is a necessary precursor to a discussion of their application in a performance prediction methodology. The discussion that follows describes FLIR performance measures in detail. Performance measures for TVs and image intensifiers, minimum detectable contrast (MDC) and minimum resolvable contrast (MRC), are similar as is their use in performance prediction. For derivation and discussion of those measures appropriate for TV sensors, see Leslie et al.[39] For detailed discussion of image intensifier performance measures, see Lawson,[40] and Refs. 41 and 42. For further discussion of FLIR performance measures, refer to Lloyd[2] and Ratches et al.[3]

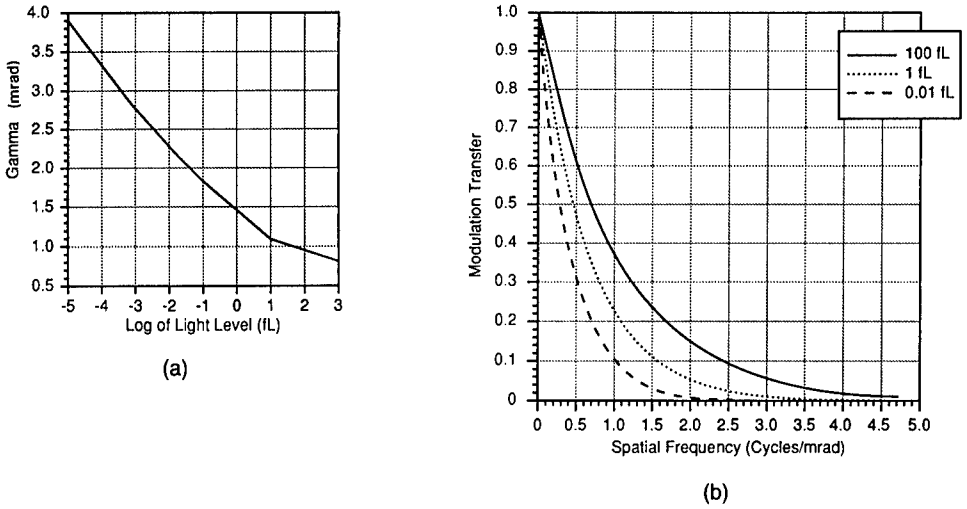


Fig. 2.9 (a) Γ, the light level dependent parameter in the eye MTF, as a function of light level (footlamberts). (b) The eye MTF of Kornfeld and Lawson as used in the NVL static performance model for three different light levels.

2.4.1 FLIR Performance Measures

Noise Equivalent Temperature Difference (NET or NETD). NET is a convenient expression of the sensitivity of a thermal system. It is the temperature difference, referenced to 300 K, between a large target (a rectangle or square at least 10 times the detector angular subtense) and its background, which is required to produce a peak signal to rms noise of one at a particular point in the signal processing chain. Customarily,[43] the signal is measured after the amplifier of the FLIR using a single-pole, low-pass filter with a 3-dB point of $(2\tau_d)^{-1}$ where the dwell time τ_d is the time it takes a point source to scan across a detector width.

As in Ratches et al.,[3] NET is equal to

$$\text{NET} = \frac{(\Delta f_n)^{1/2} 4 (f/\#)^2}{\pi \tau_o \tau_A A_d^{1/2} N_s^{1/2} \int_{\Delta\lambda} \frac{\partial L_\lambda}{\partial T} D_\lambda^* \, d\lambda} , \tag{2.19}$$

where

L_λ = spectral radiance (W cm^{-2} sr^{-1} μm^{-1})
$f/\#$ = system effective $f/\#$
τ_o = optics transmission
τ_A = atmospheric transmission over the laboratory path used to make the measurement
A_d = detector area (cm^2)
N_s = number of detectors in series
$\Delta\lambda$ = spectral band (μm)

$\partial L_\lambda/\partial T$ = derivative with respect to temperature of the spectral radiance
D^*_λ = spectral detectivity (cm $Hz^{1/2}$ W^{-1})
Δf_n = electronic noise bandpass
$= \int_0^\infty S(f_t)H^2_{\text{elect}}(f_t)H^2_B(f_t)H^2_{MD}(f_t)\,df_t$
$S(f_t)$ = normalized noise power spectrum [$S(f_t) = 1$ at f_0]
f_0 = measuring frequency at which $S(f_t)$ is normalized, customarily chosen as $1/(2\tau_D)$ in hertz
$H_{\text{elect}}(f_t)$ = electronics transfer function = $\{1/[1 + (f_t/f_0)^2]\}^{1/2}$
$H_B(f_t)$ = boost transfer function
$H_{MD}(f_t)$ = transfer function of the measuring device customarily taken to be a low-pass filter with 3-dB point at $1/(2\tau_D)$.

NET is commonly referenced to the bandwidth

$$\Delta f_n = \frac{\pi}{2}\Delta f_e = \frac{\pi}{4\tau_D}\,, \tag{2.20}$$

where

$$\tau_D = \frac{n_p \Delta x \Delta y \eta_{\text{scan}}}{\alpha\beta F_R \eta_{\text{ovsc}}}$$

and

τ_D = detector dwell time, the time it takes one detector element to scan a point in object space
n_p = number of detectors in parallel
Δx = angular subtense of the detector in the scan direction in milliradians
Δy = angular subtense of the detector in the orthoscan direction in milliradians
η_{scan} = scan efficiency
α = horizontal system field of view in milliradians
β = vertical system field of view in milliradians
F_R = frame rate
η_{ovsc} = overscan ratio = $\Delta y/\Delta y_{\text{sample}}$

where

$$\Delta y_{\text{sample}} = \frac{\beta}{n_p \times N_{\text{interlace}}}$$

and $N_{\text{interlace}}$ is the number of fields per frame.

NET has been widely used to specify system sensitivity or low-frequency response, but this figure of merit has certain drawbacks. Since it does not

include display or observer response characteristics, NET does not relate directly to human performance. Kennedy[44] suggests a "pixel" NET being equal to the square root of the variance of a sequence of samples taken from a particular detector. He also discusses ways to incorporate nonuniformity into the measure.

Minimum Detectable Temperature Difference (MDT or MDTD). Minimum detectable temperature difference is a laboratory measure of thermal system sensitivity, which, unlike NET, includes the human operator. It relates directly to the noise-limited detection performance of the system and will be used in the detection prediction method to be discussed in the following section.

MDT is the temperature difference between a square (or circular) target and a uniform background when viewed through a FLIR, which is required by a trained observer to just detect the target. The observer knows the approximate location of the target on the system display and is given as much time as needed to determine whether he or she can detect the target.

The derivation of MDT following Ratches et al.[3,45] assumes that the eye sums an input scene for a fixed integration time t_E, chosen to be 0.2 s. The spatial summation capability of the eye is modeled using a matched filter.

Calculation of the signal used in the MDT equation involves convolution of the target with the system PSF. The resulting degraded image of the target is convolved with a matched filter that has been matched to the degraded target. The signal used in the MDT equation is the peak of the degraded image after filtering with the matched filter multiplied by the number of frames in a visual response time.

The detector noise power is filtered by all system components after the preamplifier including the display and eye in order to arrive at a perceived noise. The result is convolved with the matched filter, which again has been matched to the degraded image of the target, to get the single frame noise power spectrum used in the MDT equation. Integrating the noise power spectrum over frequency, we get the variance of the noise, which is then multiplied by the number of frames in a visual response time. The square root of this quantity is used as the noise in the signal-to-rms-noise value in the MDT equation. The MDT is taken to be that temperature difference between target and background where the SNR is equal to a fixed SNR threshold, SNR_{T1}.

The equation for MDT is

$$\mathrm{MDT} = \frac{\mathrm{NET}\ \mathrm{SNR}_{T1}}{A_T \int_{-\infty}^{\infty} H_T^2 H_D^2\, d^2f} \left[\frac{\Delta y v}{\eta_{\mathrm{ovsc}} F_R\, t_E\, \Delta f_n}\right]^{1/2} \times \left[\int_{-\infty}^{\infty}\int_{0}^{\infty} S(f_x) H_{\mathrm{elect}}^2(f_x) H_{\mathrm{display}}^2(f_x,f_y) \times H_{\mathrm{eye}}^2(f_x,f_y) H_T^2(f_x,f_y) H_D^2(f_x,f_y)\, \partial f_x\, \partial f_y\right]^{1/2}, \tag{2.21}$$

where

A_T = target area (mrad^2)

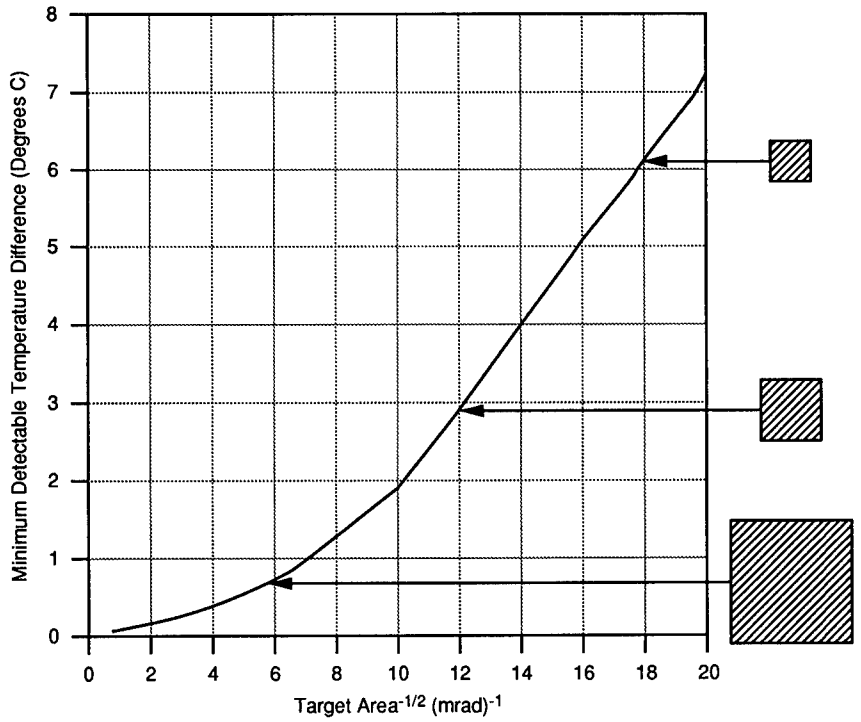


Fig. 2.10 A typical plot of MDT showing relative sizes of the MDT test pattern.

SNR_{T1} = threshold SNR for the MDT target
H_T = Fourier transform of the target (a circle or square)
H_D = total device and eye MTF
H_{display} = display MTF
H_{eye} = eye MTF.

The threshold value for SNR_{T1} of 2.25 is the result of an empirical fit of MDT measurements. MDT is measured for a set of targets of increasing area and is often plotted as ΔT versus 1/(target area) as in Fig. 2.10. The curve is not asymptotic for small targets since a target can be made detectable by increasing its temperature even when its angular subtense is smaller than an IFOV.

Minimum Resolvable Temperature Difference (MRT or MRTD). Following work on television systems by Schade,[46] Coltman and Anderson,[47] and others, Lloyd and Sendall[48] developed the thermal imager system performance measure, MRT. It is of key importance in the static performance prediction method to be described in the following section. MRT is a noise-limited threshold measure of system spatial resolution and thermal sensitivity, which includes the characteristics of the human. It is defined as the temperature difference between a uniform background and the bars of a four-bar pattern, each bar having a 7:1 aspect ratio (so the overall pattern will be a square), which is required by a trained observer to just resolve all four bars when viewing the pattern through the imager.

MRT was originally defined for a scanning FLIR, and in the measurement procedure the bars of the four-bar pattern are oriented perpendicular to the scan direction of the FLIR. The observer is allowed to adjust the gain and level settings of the FLIR as well as his or her position relative to the system display during the measurement in order to minimize the temperature difference required to just resolve the target. The just resolvable temperature differences are determined for several bar patterns of increasing fundamental frequency, and the results are plotted as in Fig. 2.11.

The MRT derivation is, in many ways, similar to the MDT derivation. Signal from a single bar is filtered by the system and then by a matched filter, which has been matched to the degraded image of the bar. Noise from an area of background equal to that of the bar is also filtered by the FLIR and matched filter. The temporal response characteristics of the eye are accounted for by multiplying the signal and noise variance by the number of frames in a visual response time.

Again following Ratches et al.,[3] the equation for MRT is

$$\mathrm{MRT}(f_x) = \mathrm{SNR}_{T2}\frac{\pi^2}{4\sqrt{14}}\frac{NE\Delta T}{H_{\mathrm{tot}}(f_x)}\left[\frac{\Delta y v f_x Q}{\Delta f_n F_R t_E \eta_{\mathrm{ovsc}}}\right]^{1/2}, \tag{2.22}$$

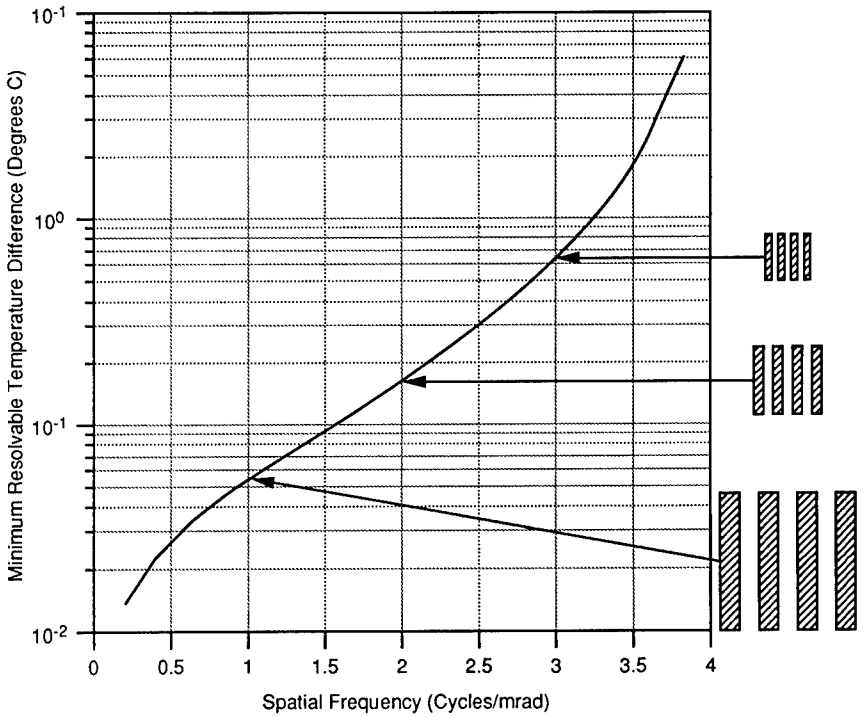


Fig. 2.11 A typical MRT plot indicating the relative sizes of the MRT four-bar pattern at different spatial frequencies.

where

SNR_{T2} = threshold SNR necessary to resolve the four-bar pattern also taken to be 2.25
H_{tot} = total system MTF
v = detector scan velocity in milliradians per second
f_x = target fundamental spatial frequency in cycles per milliradian

and the noise filtering term

$$Q = \int_0^\infty S(f_x)H_N^2(f_x)H_W^2(f_x)H_{\text{eye}}^2(f_x)\,df_x \ ,$$

where $H_N(f_x)$ is the noise filter function from the detector to display and $H_W(f_x)$ is the target filter function for a bar of width W, and, hence, $\text{sinc}(\pi W f_x)$.

As can be seen from Fig. 2.11, the temperature difference required to resolve the four bars increases as the bars get smaller. Unlike MDT, the MRT curve is finally asymptotic at a spatial frequency near $(\Delta x)^{-1}$ since even very large temperature differences will not allow resolution of bar patterns where one cycle of the bar pattern is smaller than a detector angular subtense. Because of this and because other component MTFs limit the visible contrast at high frequencies, for first-generation common module FLIR systems, one cannot usually measure a resolvable frequency higher than approximately $0.8(\Delta x)^{-1}$.

It is implicit in the measurement of MRT that the performance of the observer is limited by "well-behaved" noise on the display.[48] It was realized that the noise-limited case was "significant because an infrared imaging system exhibits its ultimate sensitivity when noise is visible to an observer. In a well designed system, noise is visible only with high system gain, as is used under adverse environmental or scene conditions. Under more favorable conditions, systems are contrast limited, and then only the modulation transfer function is important as a measure of resolution."[48] Performance was also observed to be limited by clutter, spatial frequency content, or observation time.

Several authors[49–52] have pointed out the discrepancies between predicted and measured MRTs, the predicted MRT often being higher than the measured MRT at high frequencies and lower than predicted at low frequencies. Suggested solutions to the problem have often centered around the treatment of the observer eye-brain in the model.[50–52]

2.5 STATIC PERFORMANCE PREDICTION

The laboratory and field experiments that serve as the empirical foundation of the static performance model were first performed in the late 1950s using image intensifiers,[53] and since then experimentation has been extended to include FLIRs. After the basic prediction methods are described and examples are given, other prediction methods are described, and finally some recent model updates are covered.

It should be emphasized that the result of the prediction of probability of detection, recognition, or identification using the model discussed is to be

compared with test results using an ensemble of observers. The calculation does not indicate the probability that a single observer in a specific scenario will be able to perform a given task to a specific level.

Reports by Greening[6,54,55] provide an excellent historical perspective on the development of target acquisition modeling. Several models are compared in the reports, and relevant laboratory and field data are reviewed.

2.5.1 Detection[56]

As described earlier, three basic types of detection are pertinent to target acquisition modeling: pure detection, discrimination detection, and detection as it occurs in the search procedure.

Pure Detection. The fundamental mechanism of target acquisition is pure detection: the perception of an object against a locally uniform background. It is assumed that the approximate position of the target is known, that is, no search for the target is required.

The methods used for performance prediction of all EO sensors are essentially identical. In this section the procedures and examples have been specialized to the FLIR case. In practice, the procedure for calculation of the probability of pure detection at a particular range to the target can be broken into four steps as follows (see Fig. 2.12). (This procedure is easily iterated to determine the probability of detection P_D as a function of range.)

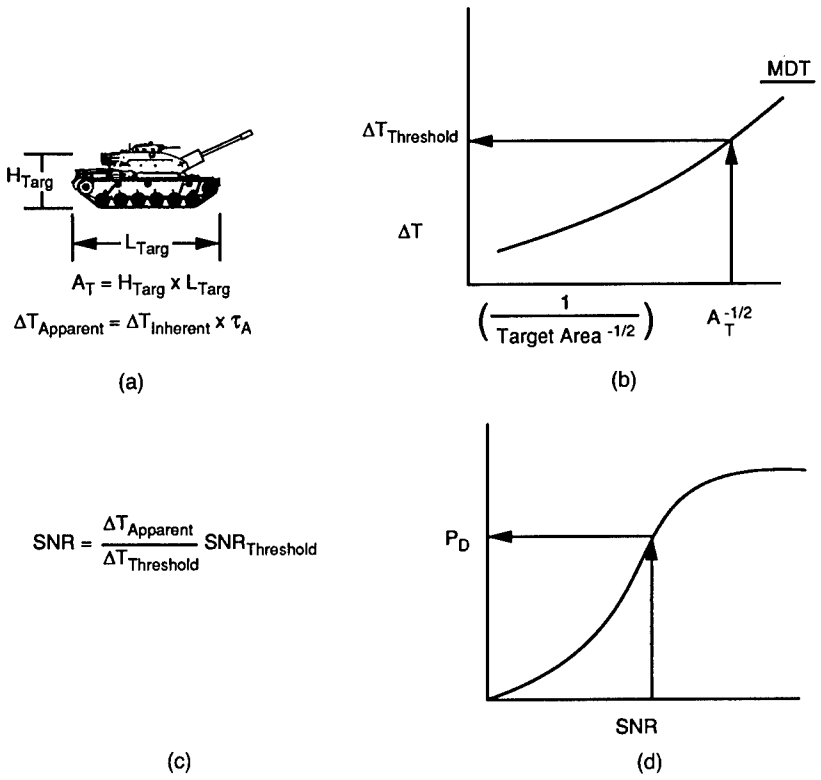


Fig. 2.12 Diagram of the pure detection calculation procedure.

1. Determine the target's inherent ΔT. Calculate the target area and the target's projected angular subtense at range R. Using knowledge of the atmospheric transmission, calculate the apparent ΔT of the target at range R.
2. Calculate or measure the system demand function, MDT. Using the target's projected angular subtense at range R and the MDT, determine the threshold ΔT required to detect the target.
3. Compute the signal to noise of the target seen by the observer as follows:

$$\text{SNR} = \frac{\text{target apparent } \Delta T}{\text{threshold } \Delta T} \times \text{SNR}_{T1} \ , \tag{2.23}$$

   where the threshold SNR for pure detection, SNR_{T1}, is 2.25.
4. Determine P_D using the empirical relationship between SNR_D and P_D shown in Fig. 2.13.

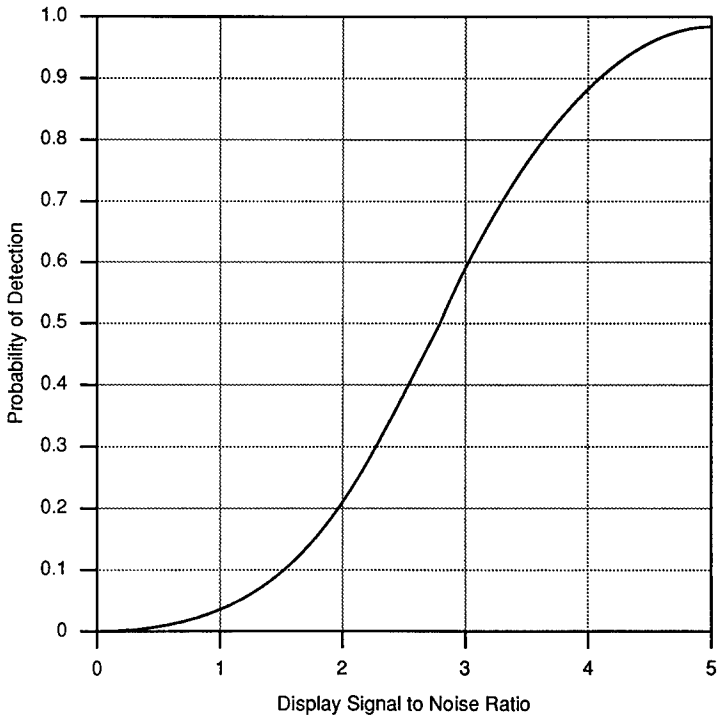


Fig. 2.13 Relation between P_D and SNR used in the pure detection prediction methodology.

Table 2.3 TOW System Parameters

System	
Horizontal FOV	2.2 deg
Vertical FOV	1.1 deg
Frame rate	30 frames/s
Interlace	2 fields/frame
Scan efficiency	0.75
System magnification	12.0
Spectral band	7.7–11.75 μm
Optics	
Effective focal length	12.0 in.
Effective aperture diameter	4.5 in.
f/#	2.667
Average optical transmittance	0.57
Diffraction wavelength	10.0 μm
Blur spot size	0.007 mrad^2
Detector	
Detector IFOV$_x$	0.133 mrad
Detector IFOV$_y$	0.200 mrad
Detectors in parallel	60
Detectors in series	1
Limiting system noise	Detector
Peak D^*	5.1×10^{10} cm Hz$^{1/2}$ W^{-1}
Display	
Display type	EO mux LED array
Average luminance	10.0 mL
LED subtense$_x$	0.0625 mrad
LED subtense$_y$	0.3130 mrad

The lack of field data relating displayed SNR and P_D meant that researchers had to rely on the results of lab tests by Rosell and Willson[33] for the establishment of the empirical relation shown in Fig. 2.13.

Example. A pure detection prediction example for an actual FLIR system follows. Table 2.3 lists system parameters of the tube-launched, optically tracked, wire-guided (TOW) infrared target acquisition system to be used in this example. Figures 2.14 and 2.15 are plots of further input to the MDT calculation in the model.

Pure detection performance (as well as higher order discrimination performance and search performance presented in later sections) are to be predicted for an M60 tank front for two atmospheric conditions.

Step 1 of the pure detection prediction example requires the calculation of the solid angle subtended by the target in the reference frame of the FLIR. The physical size of the M60 target front is 3.6 m wide by 3.2 m high, resulting in a total area of 11.52 m^2. At a range of 6 km, the projected angular subtense is 0.32×10^{-6} sr.

The performance of the imager in two atmospheres will be compared: The first is a "good" atmosphere corresponding to a visibility range of 23 km, relative humidity of 50%, and air temperature of 15°C, which is termed the "U.S. Standard, Spring–Summer" atmosphere in the LOWTRAN 7 atmospheric code,[12] and the second is a "poor" atmosphere with a 5-km visibility

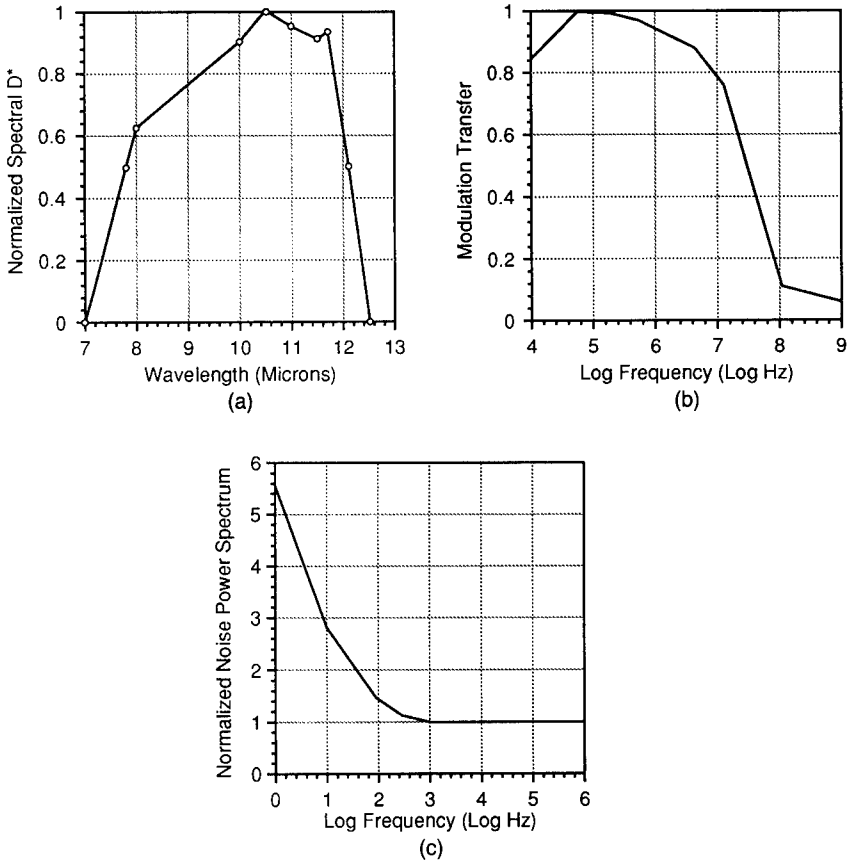


Fig. 2.14 (a) Spectral D^*, (b) system electronics response, and (c) normalized noise power spectrum for the example system.

range, 75% relative humidity, and an air temperature of 27°C, called the "Tropical Spring–Summer" atmosphere. Spectral transmission for the two atmospheres is plotted in Fig. 2.16.

For this example, the inherent ΔT of the M60 target is taken to be 1.25°C. The apparent target ΔT will be a function of range due to the attenuation of the signal by the atmosphere. Total transmission integrated over the spectral band of the TOW system, as determined using LOWTRAN 7, is plotted versus range in Fig. 2.17. This is used to determine the apparent temperature at a particular range. At a range of 6 km the atmospheric transmission is 51% and 7.5%, leading to apparent ΔT's of 0.64° and 0.093°C for the good and poor atmospheres, respectively.

The second step in the prediction procedure involves calculation of the system MDT. NET, one of the parameters in MDT, is calculated first using Eq. (2.19). Using the FLIR system descriptors above, the NET is calculated as 0.17°C, which compares favorably with laboratory NET measurements of the TOW system. MDT is then calculated using Eq. (2.21). The resulting MDT is plotted as a function of inverse target size (mrad^{-1}) in Fig. 2.18.

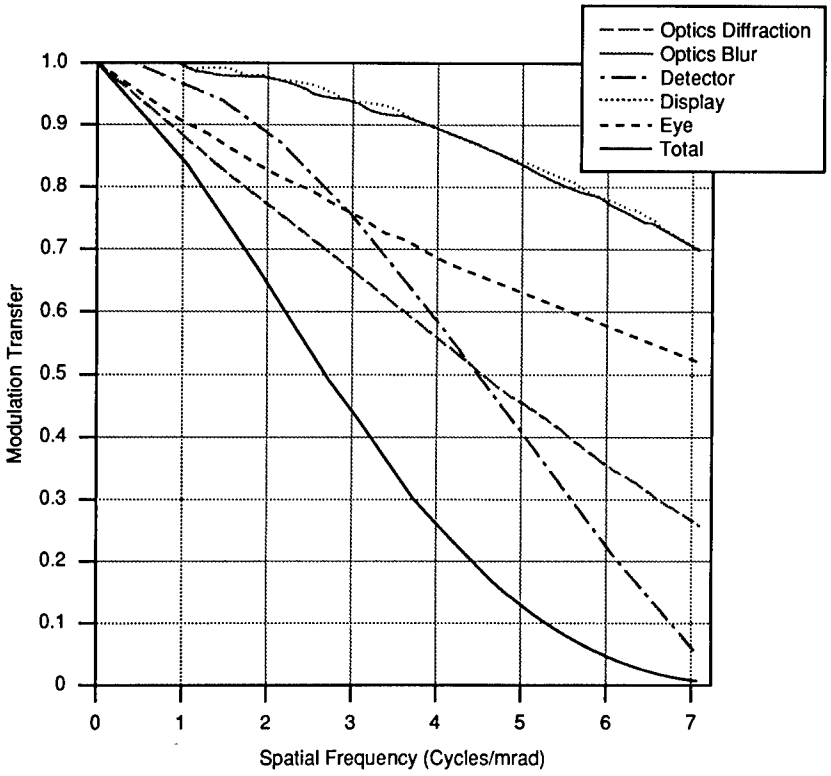


Fig. 2.15 Optics, detector, display, eye, and total system MTFs for the example system.

Knowing the target subtense and having calculated the system MDT, the threshold temperature difference required to just detect the target is read off the graph of MDT. The target size (taken as the square root of the area) at 6 km subtends 0.57 mrads. The corresponding inverse target size is 1.77 mrad^{-1}. From the MDT it is determined that the threshold temperature difference is 0.058°C. Once the apparent temperature difference at a particular range has been determined, the SNR of the target is calculated as in step 3. For step 4 the probability of detection is then simply read off the graph of SNR_D versus P_D in Fig. 2.13.

Using the "poor" atmosphere, at a range of 6 km the SNR of the target using Eq. (2.23) is 3.6. From Fig. 2.13 the corresponding probability of pure detection is 77%. The procedure can be iterated for many ranges in order to arrive at a graph of probability of detection versus range as has been done for Fig. 2.19.

Although a number of laboratory experiments have investigated threshold detection capability, the same cannot be said for field experiments. Several factors combine to make the field testing for pure detection difficult, and as a

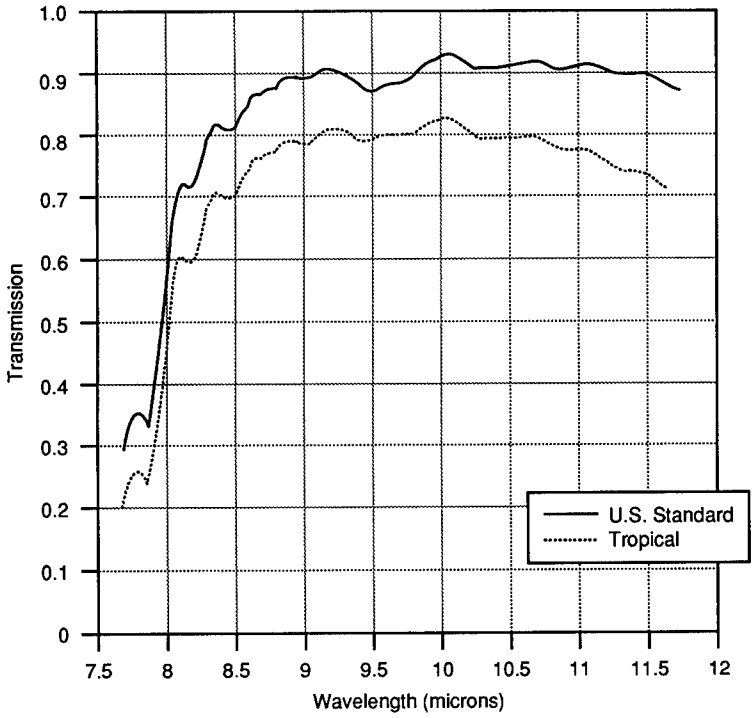


Fig. 2.16 Spectral transmission for the two atmospheres used in the example.

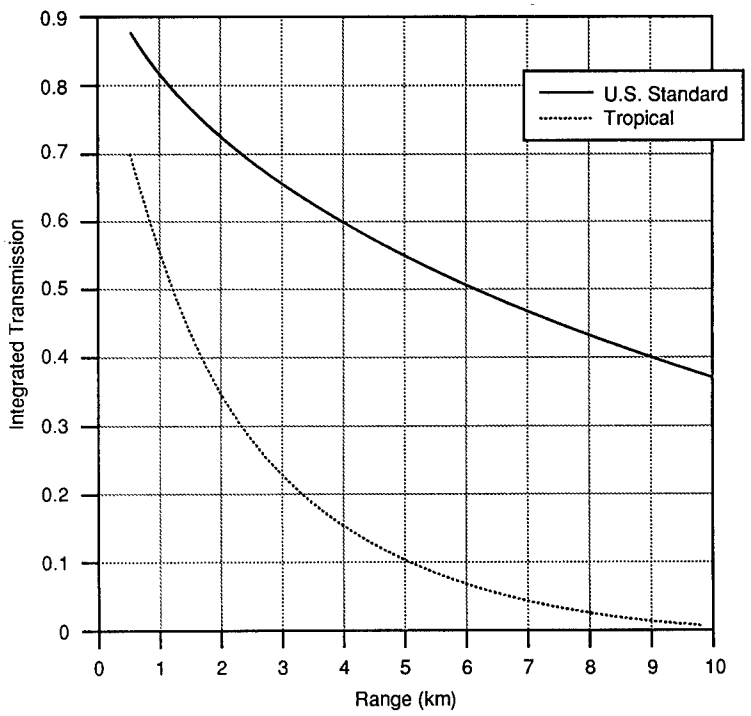


Fig. 2.17 Total integrated transmission versus range for the example atmospheres.

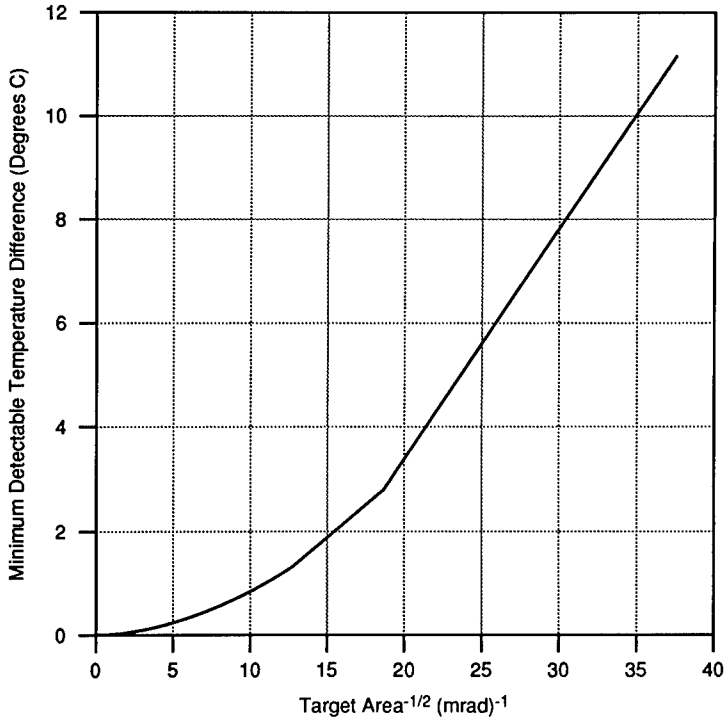


Fig. 2.18 MDT for the example system.

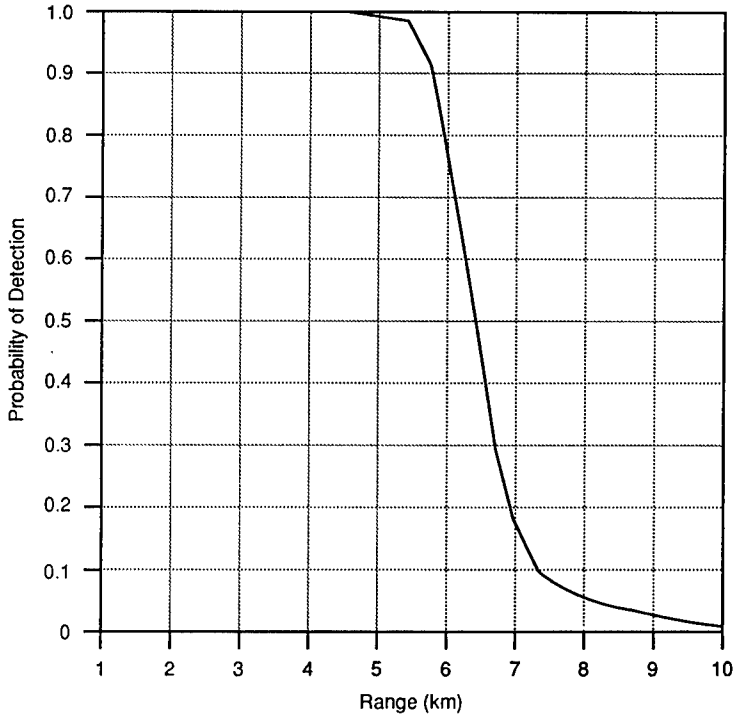


Fig. 2.19 Probability of pure detection versus range for the "poor" atmosphere using the example system.

result there has not been a great deal of field validation of the prediction method for pure detection.

2.5.2 Higher Order Discrimination and the Johnson Methodology

> Intuitively, it would seem that there must be some relationship between the number of lines resolved at the target and the corresponding decisions of detection, recognition and identification.[53]

In the late 1950s John Johnson investigated the relationship between the ability of an observer to resolve bar targets through an imaging device and their ability to perform the tasks of detection, recognition, and identification of military vehicles through the same imager. It is worthwhile to review his experiment briefly, along with what has become known as the *equivalent bar pattern approach*, before discussing higher order discrimination tasks since the empirical relationship that he determined serves as the foundation for the prediction methodology to be discussed.

In the laboratory, scale models of eight different vehicles and one soldier were placed against a bland background. Observers viewing the targets through image intensifiers were asked to detect, determine the orientation of, recognize, and identify the targets. Air Force bar charts whose bars had the same contrast as the scale targets were also viewed through the same imager, and the maximum resolvable bar pattern frequency was determined as a function of contrast of the target. The maximum number of resolvable cycles across the target critical dimension was determined for each different task using

$$N = H_{\text{targ}} \cdot f_x \ ,$$

where

N = the number of just resolvable cycles across the target critical dimension

H_{targ} = the critical dimension of the military target in milliradians, usually chosen to be the minimum dimension, which for tactical army vehicles is often the height

f_x = the highest resolvable bar pattern fundamental spatial frequency in cycles per milliradian.

(While reading the literature on target acquisition, attention should be paid to different units of spatial frequency where one line pair = one cycle = a bar and a space between the bars in a multiple bar pattern.)

Table 2.4 shows the results of the experiment in which it was found that the number of just resolvable line pairs across the minimum dimension of the target required to perform a particular discrimination task was within 25% of a fixed number of cycles. As long as the bar target contrasts equaled the target-to-background contrasts, the resulting number of cycles was found to be independent of target-to-background contrast and scene light level. In this manner the ability of an observer to perform discrimination tasks was related to their ability to resolve bar patterns, and, therefore, the prediction of the discrimination performance of the observer looking through an EO device was simplified to the prediction of their ability to resolve bar patterns through the EO system.

Table 2.4 Number of Cycles Required across a Target's Critical Dimension for Various Discrimination Tasks (from Ref. 53)

Target	Resolution per Minimum Dimension			
Broadside View	Detection	Orientation	Recognition	Identify
Truck	.90	1.25	4.5	8.0
M-48 tank	.75	1.2	3.5	7.0
Stalin tank	.75	1.2	3.3	6.0
Centurion tank	.75	1.2	3.5	6.0
Half-track	1.0	1.5	4.0	5.0
Jeep	1.2	1.5	4.5	5.5
Command car	1.2	1.5	4.3	5.5
Soldier	1.5	1.8	3.8	8.0
105 howitzer	1.0	1.5	4.8	6.0
Average	1.0 ± 0.25	1.4 ± 0.35	4.0 ± 0.8	6.4 ± 1.5

Results of several field tests[45] using thermal imagers have been used to establish what has been termed the *target transform probability function* (TTPF). Johnson reported in Table 2.4 what was essentially the 50% point (to be referred to as N_{50}) on a curve defining what portion of an ensemble of observers was able to perform a particular discrimination task. The shape of the TTPF curve can be approximated by

$$P = \frac{(N/N_{50})^E}{1 + (N/N_{50})^E} , \tag{2.24}$$

where

$$E = 2.7 + 0.7(N/N_{50}) . \tag{2.25}$$

The TTPF curve when plotted as a function of N/N_{50} can be used for all discrimination tasks by simply associating a particular N_{50} with the 50% probability of performing a particular task. Such a curve is plotted in Fig. 2.20.

Later field testing using FLIRs indicated that the N_{50} for identification was closer to 8 cycles than 6.4 as determined by Johnson. This could be caused by several factors, perhaps primarily the differences between I^2 and FLIR imagery. It was also noticed that recognition tasks varied widely in difficulty so that N_{50} varied over the range of 3 to 4 cycles. This led to a distinction being made between "conservative" (4-cycle) and "optimistic" (3-cycle) recognition tasks. It should be stressed that these N_{50} values should only be taken as representative values. When an analyst wishes to predict the performance of a specific task they should make a judgment about the difficulty of the task and vary N_{50} accordingly.

The procedure for the prediction of all of the higher order discrimination tasks is essentially the same except for the number of resolvable cycles required to perform the task. The prediction methodology being described is general in that the recommended cycle criteria are the result of averaging of test results incorporating several situations, observers, clutter levels, target aspects, etc. Predictions using this technique will not reflect the performance of a specific

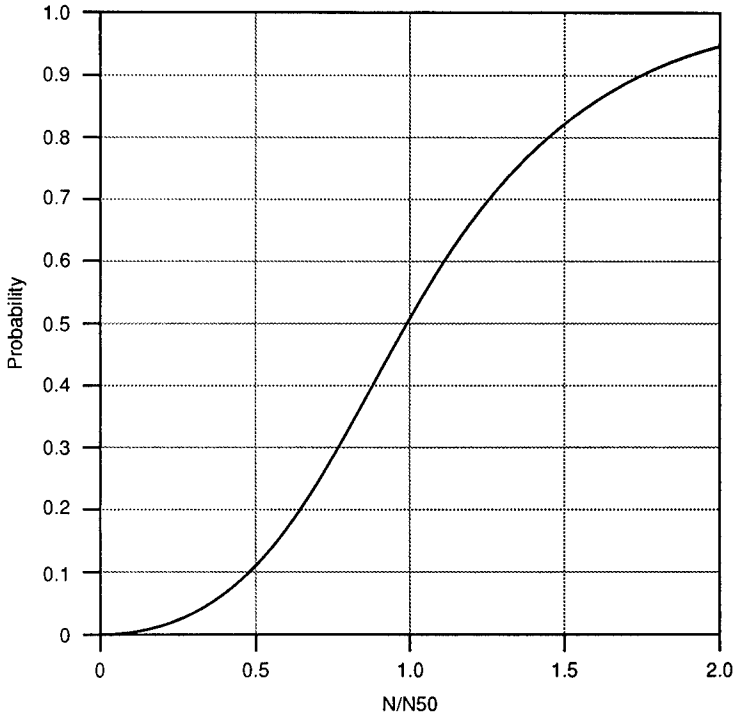


Fig. 2.20 Target transfer probability function curve.

observer trying to discriminate a particular target signature, rather the prediction is of the performance of an ensemble of observers viewing a set of target signatures. Predictions for recognition are thought to be accurate to $\pm 20\%$ in range.

The steps involved in the discrimination prediction methodology are as follows:

1. Determine the target critical dimension, range, and inherent ΔT. Using knowledge of the atmospheric attenuation as a function of range, calculate the apparent ΔT of the target at the range to the target as described in the section on pure detection.
2. Calculate or measure the system MRT. From the apparent ΔT and the MRT determine the maximum resolvable spatial frequency of the sensor at this apparent ΔT, f_x, in cycles per milliradian.
3. Using the angular subtense of the target critical dimension, H_{targ}/R, calculate the maximum number of resolvable cycles across the target, N, using

$$N = f_x \frac{H_{\text{targ}}}{R} .$$

4. Determine the probability of performing a task from the TTPF curve.

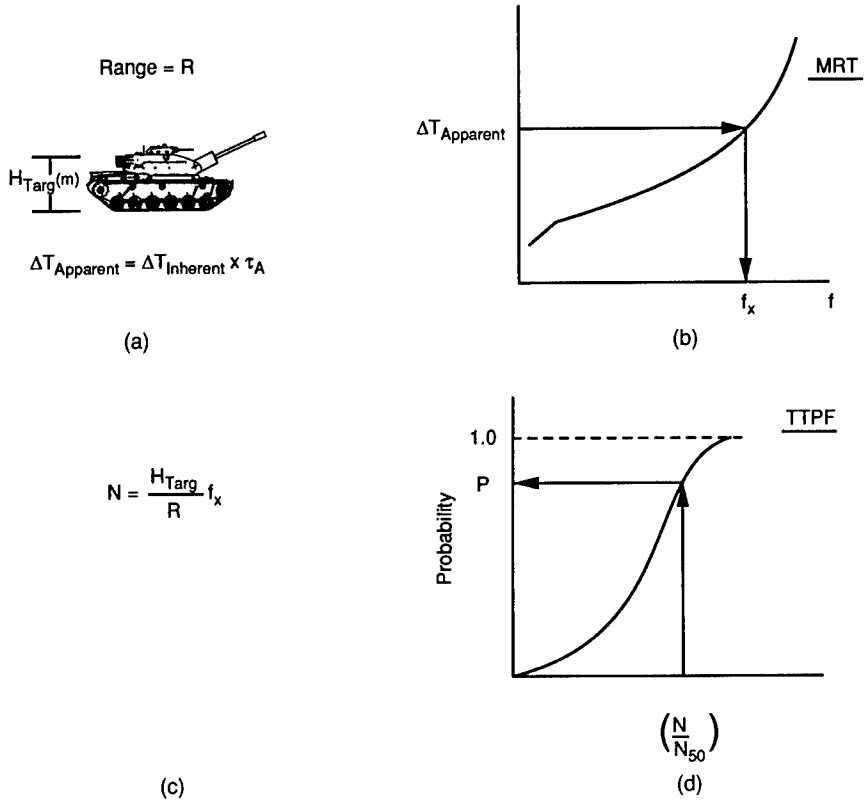


Fig. 2.21 Schematic diagram of the discrimination performance prediction methodology.

This procedure is depicted schematically in Fig. 2.21. It is easy to iterate the procedure to arrive at a prediction as a function of range.

Example. Using the same 3.6- × 3.2-m M60 tank front target with a ΔT of 1.25°C and the same "good" and "bad" atmospheres as described in Sec. 2.5.1, we will consider the recognition performance (N_{50} = 4 cycles) at a range of 3 km. For the "good" atmosphere the transmission at 3 km is 65% resulting in an apparent ΔT of 0.81°C. The MRT for the example TOW system calculated using Eq. (2.22) is shown in Fig. 2.22. The maximum resolvable frequency corresponding to an apparent ΔT of 0.81°C is approximately 5.05 cycles. The corresponding number of resolvable cycles across the critical dimension of the M60 target (height = 3.2 m) is approximately 5.4 cycles. Using the TTPF curve of Fig. 2.20, this leads to approximately a 75% probability of recognition (N_{50} = 4 cycles). The same procedure using the bad atmosphere at 3 km leads to a 53% probability of recognition. Figure 2.23 shows recognition performance versus range for the two atmospheric conditions.

It was thought that the higher order discrimination predictions of the model would be more accurate if the portion of the equivalent bar pattern used in the prediction (for example, the four bars and spaces used for recognition) subtended approximately the same solid angle as the target. A target length correction was applied in the original version of the model. It has the most

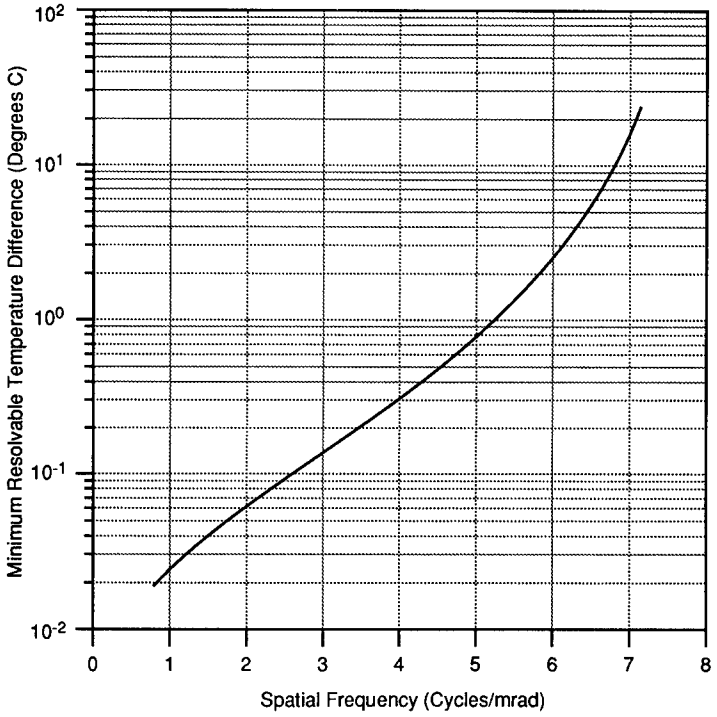


Fig. 2.22 MRT for the example system.

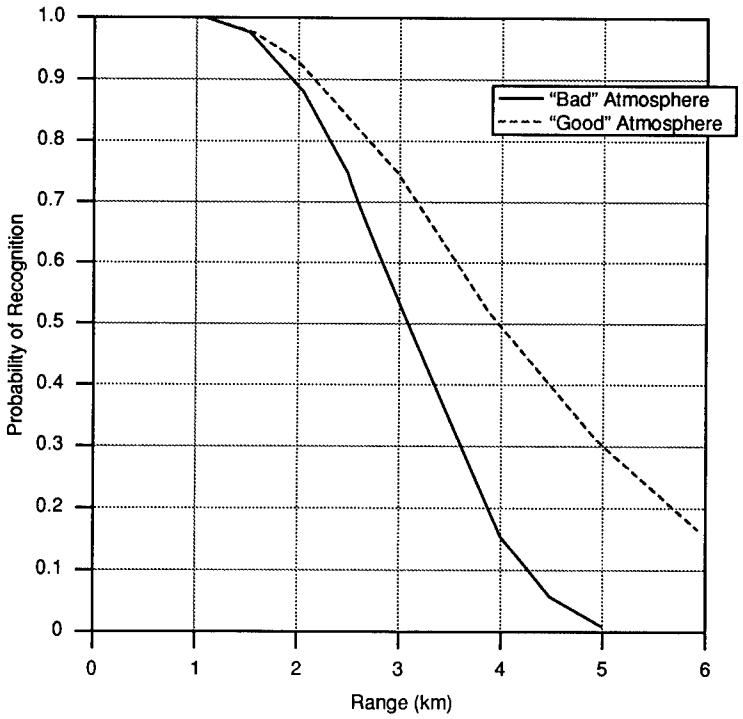


Fig. 2.23 Recognition performance versus range for both atmospheres.

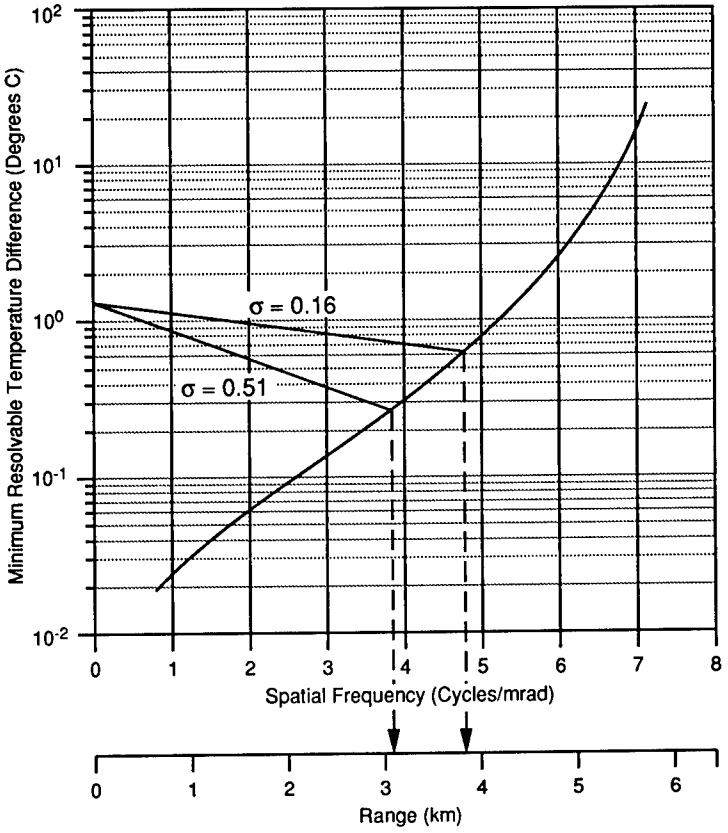


Fig. 2.24 Loadline calculation for the example system.

impact on predictions when the general shape of the target is very different in overall shape from that portion of the bar pattern used, for example, very different from a square for 4-cycle recognition, or different from a 2:7 aspect ratio for 1-cycle detection.

For recognition of tactical ground vehicles where the target aspect does not vary much from a square, this length correction does not have much effect. When the target is very long, such as is the case for a ship, the correction becomes significant.

Loadline Procedure. Once the MRT has been either measured or calculated, the process of estimating range performance can be completed using a relatively simple graphic technique sometimes called a *loadline procedure*. Besides its simplicity, the technique has the advantage of allowing the analyst to compare rapidly the essential relationships between several imagers, imager designs, cycle criteria, atmospheres, and/or targets all on the same graph. The steps involved in the procedure, using the TOW MRT as an example, are as follows (refer to Fig. 2.24):

1. Plot the measured or calculated MRT(s) on a semilog graph.
2. Knowing the target critical dimension H_{targ} and the number of cycles

required to perform the discrimination task to the desired level N, convert the frequency axis into a range axis using

$$R = f\frac{H_{\text{targ}}}{N} .$$

3. Knowing the atmospheric extinction and assuming that Beer's law adequately approximates the attenuation, apply it to the target inherent ΔT to get the range dependence of the apparent temperature difference (Beer's law is accurate for monochromatic radiation but is less accurate for broadband transmittance because of the structure in the atmospheric absorption spectrum.):

$$\Delta T_{\text{apparent}} = \Delta T_{\text{inherent}} \cdot \exp(-\gamma R) ,$$

where γ is the atmospheric extinction coefficient (km^{-1}). The exponential attenuation when plotted on a semilog plot will result in a straight line with negative slope of γ (the "loadline") that intersects the MRT curve at some range (= spatial frequency). The point of intersection corresponds to the maximum range at which the task can be performed to the desired level. Keep in mind that for each new H_{targ} or N on the graph, a new range axis and atmospheric loadline must be drawn even if the atmospheric transmission is the same.

Example. For the TOW example system the "good" atmosphere was approximated by an atmosphere with a γ of 0.16 ($\approx$85% transmission per kilometer), and the "poor" atmosphere by a γ of 0.51 ($\approx$60% transmission per kilometer). For a 4-cycle (50% probability) conservative recognition calculation, the poor atmosphere results in approximately 3.1 km and the good atmosphere about 3.8 km. These values differ only slightly from those predictions that make use of a more detailed atmosphere, yet we realize that the loadline is only a convenient tool to be used for approximate calculations.

2.5.3 Discussion

Most of the targets used in the field tests, which served as validation for the modeling methods for both detection and higher order discrimination tasks, were high-contrast targets. The majority of the field trials tested the observer's recognition and identification capability; there has been relatively little field testing of either pure or discrimination detection prediction. To date, the amount of data used to support the two detection prediction methods remains small.[3,56]

The similarity of a target signature to other signatures in a given target signature set strongly influences the N_{50} required for recognition. For a given battlefield scenario or field test, the task of recognition is going to be more difficult if the signature set consists of signatures that appear similar rather than disparate. The dependence of task difficulty on the expected composition of the target set has led van Meeteren[57] to propose, for the purposes of modeling, that a standard set of targets be decided on by the modeling community.

Target aspect can have a large effect on the resolution required for recognition and identification.[58] Researchers noticed that many targets were harder

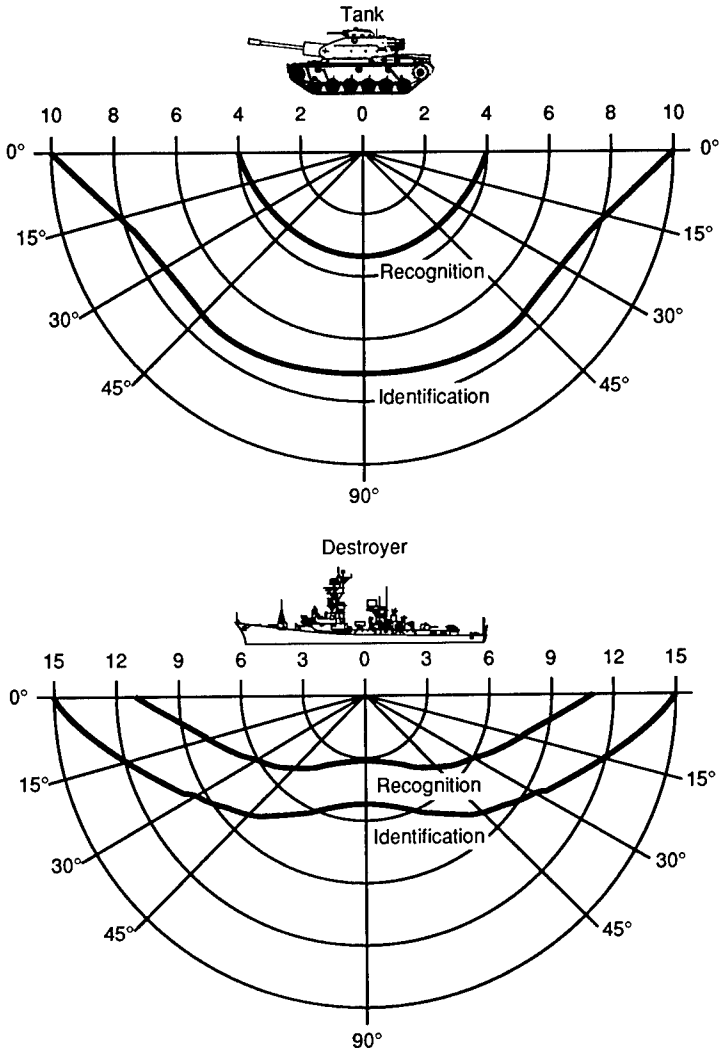


Fig. 2.25 Polar diagram indicating the anisotropy of the cycle criteria for different vehicles (from Ref. 3).

to recognize from the front than from the side, and they began to make polar diagrams such as those in Fig. 2.25 to show the anisotropy in the cycle criteria for particular targets. Often there is an absence of distinctive features when viewing a target from the front; however, distinguishing features appear when a near front or side aspect is viewed. In a laboratory test using 70 scale models,[58] the N_{50} required for recognition of ground vehicles did not increase by more than approximately 30% in going from the side to the front aspect. However, for targets with larger length-to-width ratios such as battleships and aircraft carriers the anisotropy in N_{50} for recognition and identification was more pronounced, sometimes being as much as a factor of 5 between the side and front aspects.

The authors of Ref. 3 summarized the limitations of their model in the 1975 model documentation. They mentioned the lack of validation for detection and higher order discrimination tasks using low ΔT targets, and they expressed the need for a "complex target model" and a more suitable way to choose the critical dimension of the target than simply the minimum dimension. They planned to incorporate the target aspect variation of P_{rec} into a future version of the model.

It was realized that the resolution of an imaging system in two dimensions should be accounted for in a performance prediction methodology. Prediction of the MRT perpendicular to the scan direction, MRT_Y, as well as a "45-degree MRT" were included in the model even though they were hardly ever measured in the laboratory and were not used in predictions.

2.5.4 Other Static Performance Models

Section 2.5.2 described a model that is one of the most widely used performance models and that has served as the basis for several other similar models. The model is being enhanced in order to improve the prediction accuracy for modern thermal imagers. Before the more recent imagers and model enhancements are described, it is instructive to compare the form of some other static performance models.

One of the reasons for trying to account for the system resolution in both directions in a performance model is related to the typical target description used in the models. As related earlier, for some targets N_{50} has a strong dependence on target aspect. This led system analysts to develop models that used the number of resolvable areas, or pixels, on target, instead of the number of resolvable bars across a critical dimension as a criterion for whether an observer could perform a discrimination task.[59–61] It was assumed that the target area more accurately reflected the information from the image than did a critical dimension. The resolvable pixel was a function of resolution in both the scan and orthoscan directions. This approach has the advantage of not requiring the analyst to select a target critical dimension. The whole target projected area can then be used, a fact that is more important for targets with nonsquare aspect ratios such as ships.

Moser's[60] work was one of the early efforts to relate empirically the number of pixels on target to the probability of performing a discrimination task. Through laboratory testing using black-and-white ship silhouettes presented as a mosaic of square tiles or "pixels," he determined that the approximate number of pixels needed to perform ship classification (distinguish a commercial vessel from a warship) was approximately 66. The number of pixels required for identification (deciding on the class of ship) was around 400. These numbers were used as discrimination criteria in a prediction method that involved resolvable pixels.

O'Neill's[61] work also included an investigation of the number of resolvable pixels required to perform discrimination tasks. He used an active range-gated low-light-level TV in the laboratory to image scale models of ships. The number of resolvable pixels required to perform discrimination tasks similar to those in Moser's study were reported to be significantly higher: 400 resolvable pixels versus 66 square pixels for classification and 1000 versus 400 for identification.

However, the imagery in the two studies was significantly different. The targets in Moser's study were black-and-white tiled silhouettes whereas O'Neill's were TV images of models.

Another major conclusion of O'Neill's work was that there was not a close relation between image detail and bar pattern resolution. As the SNR decreased, the number of resolvable pixels required to perform a given discrimination task increased. Flaherty's work[62] on image intensifiers also showed evidence for increasing N_{50} with decreasing light levels (decreasing SNRs).

Rosell and Willson measured threshold signal-to-noise values (SNR_D) for bar patterns of varying spatial frequency. They found that the threshold varied, being as high as 4 at lower spatial frequencies and near 1.5 at high frequencies. In an approach similar to that described in Sec. 2.5.2, Sendall and Rosell[52] advocate the use of a variable threshold SNR_D, fixed at 2.5 below 400 TV lines per picture height and decreasing above that frequency.

The point was made by Rosell[63] that system performance can be limited either by noise or by system MTF ("aperture limited"). It was observed that more resolution is required to perform a particular discrimination task using a sensor that is noise limited rather than aperture limited, and that a future prediction method should account for the possibility of both cases. Rosell has also suggested that performance predictions would be more accurate if some account were taken of the variation of threshold SNR with bar pattern frequency.[52]

Van Meeteren[64,65] and Vos and van Meeteren[66] examined observer performance of discrimination tasks viewing projected images of tactical ground vehicles through an image intensifier. He formulated a model that related the performance of various discrimination tasks with the ability of an observer to detect various size disks. For his chosen set of six targets he found that 50% probability of recognition is equivalent to detection of a disk with a diameter of 0.7 m.

Blumenthal and Campana[67,68] investigated several image quality metrics and their relation to subjective ranking of image quality. Results of testing suggested that image quality was inversely proportional to the size of a barely detectable object, leading the investigators to promote an "aperiodic detection model" that, like van Meeteren's approach, relates detectability of a circle or square with the discriminability of military targets.

Task[69] investigated 19 display system figures of merit, most of which were broadband. He did not come up with a specific performance model. Yet, he did find that the figures of merit that had the highest correlation with observer performance of discrimination tasks involving tactical military vehicles were those measures that had been logarithmically transformed so that they were weighted properly in terms of visual information content.

Moser questioned the basic assumptions of these performance methodologies, namely, whether "the amount of identifying information conveyed by a target image is a function of the projected area of the target."[60] He pointed out that most of the projected area of a ship is relatively featureless, and most of the identifying information is in the superstructure. He went on to say "Perhaps a better measure of information content would be the resolvable perimeter of the target or the number of resolvable angles along its perimeter."[60] One performance model somewhat along those lines has been described

by Overington.[5,70] He postulates that detection and higher order discrimination performance can be related to the resolvability of fractions of a target's perimeter, the smaller the resolvable fraction of the perimeter, the higher the order discrimination task that is possible.

2.5.5 Recent FLIR Systems and Performance Models

The limitations of the one dimensional (1-D) modeling approach described in Sec. 2.5.2 became more pronounced as FLIR system designs deviated more and more from typical first-generation systems. Some features of more modern FLIR systems that can significantly influence system performance and which are not addressed in the 1-D model are now discussed in more detail.

Sampling. The influence of spatial sampling[71–74] was neglected in the previous MRT derivation. The image was not sampled along a scan line for first-generation FLIR systems and so, at least for prediction of MRT parallel to scan (bars oriented perpendicular to the scan direction), sampling effects were not a factor. Still, for newer scanning and staring sensors that are sampled in two dimensions, sampling can play a larger role.

The resolution of an imaging system has commonly been equated with its IFOV size when in reality an observer's ability to resolve objects in an image can be limited by a host of factors including the finite spacing between samples. In most EO imaging systems, the image is sampled both spatially and temporally. As explained in Sec. 2.3.3, a parallel scanned FLIR spatially samples an input scene perpendicular to the scan direction by virtue of the spacing of the detectors on the array. For detector arrays that employ charge transfer devices to read signals off the focal plane (see the second-generation system depicted in Fig. 2.3), each detector temporally samples the scene as it scans across the FOV.

Previous discussion of FLIR components showed that many tended to low-pass filter an input signal resulting in a band-limited signal. Periodic sampling[16,71] of a band-limited signal, either in time or space, creates replicas of the baseband signal in the frequency domain centered on integer multiples of the sampling frequency f_s. The Nyquist frequency, f_N, is defined to be $f_s/2$. See Fig. 2.26 for 1-D frequency-domain representations of the baseband and sidebands for three different sampling rates.

If the sampling rate is chosen such that the limiting frequency of the signal being sampled is greater than f_N, the signal is said to be undersampled. The sidebands and baseband will overlap as in Fig. 2.26(a), causing what is termed *aliasing*. A spatial frequency f_x greater than f_N and sampled at f_s will result in an attenuated output signal at $f_s - f_x$. Within the overlapping frequency band there is an ambiguity since it is impossible to know whether the output frequency resulted from an input frequency of f or $f_s - f$.

The aliased signal can sometimes result in jagged edges in an image. Also, undersampled systems can exhibit a shift variance, a dependence of the system PSF on position in the image, which can lead to difficulties in the definition of MRT. The appearance of the four-bar MRT target will depend on the relative phasing of the sampling grid and the target, a phenomenon that has been called the *sample-scene phase effect*.[75] The shift variance does not allow the strict application of linear system theory to systems modeling.

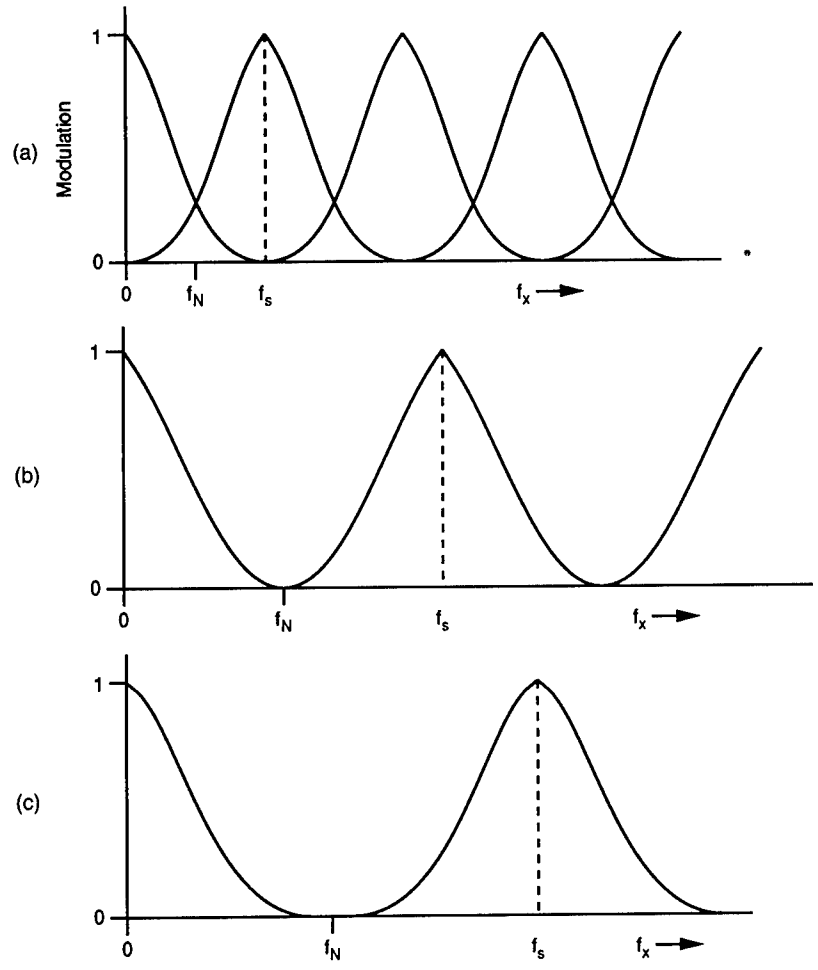


Fig. 2.26 Frequency-domain representations of the influence of different sampling rates corresponding to (a) undersampling, (b) critical sampling, and (c) oversampling.

If f_s is increased to the point that the limiting frequency of the band-limited signal is just equal to f_N, as shown in Fig. 2.26(b), then the signal is said to be critically sampled. Here the sampled signal will not exhibit any aliasing or sample-scene phase effects if the signal is properly reconstructed. A perfect low-pass postsample filter cutting off frequencies greater than f_N will completely remove any sidebands, leaving only the original baseband signal, exactly reconstructing the input to the sampler. Good approximations to a perfect low-pass filter will cut off near f_N with little attenuation of the high baseband frequencies and little contribution from the sidebands. Increasing the sampling rate still further results in oversampling [Fig. 2.26(c)], where the baseband and sideband do not overlap at all. Again, no aliasing or phasing effects will be seen after proper reconstruction.

One way in which systems analysts have tried to account for the influence of sampling in performance prediction is to assume that the MRT asymptotes

at f_N. Calculations involving frequencies greater than f_N are not allowed. Other methods of modeling the effects of sampling have been mentioned in the literature. As one way to account for the influence of sample-scene phasing, Park[73] describes an MTF that has been averaged over all possible phases of the sampling grid with respect to the scene. Rosell[76] also recommends an MTF that is an average of the MTFs corresponding to the best and worst sample-scene phases. These averaged MTFs are perhaps appropriate for field performance predictions where all phases are possible, but, for prediction of laboratory measurements of MRT where only one phase is used, they are not suitable. Kennedy[44] advocates treating the aliased part of the signal spectrum as a contribution to the noise power spectrum.

Although a few investigations of the influence of sampling on observer performance have been performed,[13–15] a complete understanding of the influence of sampling rate and the related presample and postsample filtering on observer performance remains a topic for future research.

Nonuniformity. For imaging systems employing multiple detectors, the response to an input signal can vary from detector to detector. If the response variation is not corrected, the resulting imagery will exhibit a *fixed-pattern noise*. Several sources have been mentioned as possible contributors to the nonuniformity, for example, nonuniform thermal reference sources, electrical and optical crosstalk, and nonlinear detector channel transfer functions. For staring sensors the uncorrected nonuniformity can result in a stationary "windshield mud" pattern superimposed on the imager. For a parallel scanned array, the uncorrected image will have pronounced streaks along the scan direction.

The 1-D system performance model discussed in Sec. 2.5.2 was not developed or validated using imagery where the fixed pattern noise was a significant contributor to the visible noise. In the derivation of system-level figure of merit, it was assumed that "well-behaved" temporal noise limited observer performance. Fundamental perception experiments will have to be done to quantify the influence of various amounts of fixed pattern noise on an observer's ability to perform various system discrimination tasks.

Incorporation of the fixed pattern noise into a performance modeling scheme requires the realization that the temporal summation of the eye incoherently sums white noise while it coherently sums the noise due to nonuniformities. For details on several efforts to include nonuniformity effects in performance models see Cantella,[77] Rosell,[78] Kennedy,[44] Mooney,[79] and D'Agostino.[80]

Charge Transfer Devices. The signal from each detector of first-generation systems was read off the focal plane and out of the dewar on a separate lead. In contrast, modern systems usually read signals off the focal plane using some sort of charge transfer device, the signals from which are then multiplexed and read out of the dewar.

If a charge-coupled device (CCD) is used to read signals off the focal plane, then the transfer function due to integration of the charge for the duty cycle of the CCD is given by

$$H_{\text{integ}}(f_x) = \text{sinc}(\pi \Delta x_{\text{sample}} f_x) \ , \tag{2.26}$$

where Δx_{sample} is the angular subtense of a single sample in milliradians. This replaces the first-generation single-pole temporal rolloff of the detector. One can account for the charge transfer inefficiency using

$$H_{\text{CCD}}(f) = \exp\left\{-n\varepsilon\left[1 - \cos\left(2\pi\frac{f}{f_{\text{sample}}}\right)\right]\right\} , \tag{2.27}$$

where

n = number of cells in the CCD shift register
f = input signal spatial frequency
f_{sample} = frequency at which the CCD samples the signal
ε = charge transfer efficiency.

These MTFs apply along the direction of the readout. If a multiplexing stage is used after the readout CCDs, the response of the multiplexer is accounted for using the CCD equations with appropriate changes for the sampling frequency, number of transfers, and transfer efficiency.

Reconstruction. Often the only reconstruction performed on the signal before final reconstruction at the display is a simple sample-and-hold circuit that implies that the filtering done on the signal has the form

$$H_{\text{recon}}(f_x) = \text{sinc}(\pi\Delta x_{\text{sample}} f_x) . \tag{2.28}$$

Higher order reconstruction filters (e.g., linear interpolation, cubic spline, Butterworth) would require substitution of the appropriate function for $H_{\text{recon}}(f_x)$.

Image Processing. Several real-time image processing operations can be implemented in the digital processing chain of current imagers, many of which are nonlinear operations and as such cannot be inserted into the present linear systems formulation of performance models as simple MTFs. Automatic gain and level controls are typical features of current imagers. Histogram modification techniques can be used to display optimally a signal with a large dynamic range. Although their influence on performance can be significant, their incorporation into performance models is not straightforward.

Horizontal and Vertical Resolution. First-generation parallel scanning detector arrays were standardized around only a few designs. The resolution of the systems, as evidenced by typical system MRTs, perpendicular to scan was around half that parallel to scan. Second-generation FLIR systems were built that had nearly equal resolution along the scan direction and perpendicular to it. It seemed logical that the improved resolution perpendicular to the scan direction should result in an improvement in predicted performance, a prediction that was not possible using a 1-D model.

Some experimental evidence exists corroborating the hypothesis that increased vertical resolution improves observer performance. The results of much of the related testing are complicated by the fact that vertical and horizontal resolution parameters are not treated separately. Results of tests where the number of TV raster lines subtending the height of a target (and therefore the vertical sample spacing) was varied indicate better discrimination capability

for an increased number of raster lines. Precise quantification of the relative importance of vertical and horizontal resolution has been difficult since experimental results were often confounded by the presence of a strong raster effect. For description of some of these experiments see Williams and Erickson[81] and studies summarized in Farrell and Booth.[24] Only a few studies have treated sample spacing and IFOV size variations along each dimension separately.[13–15]

Many researchers have proposed two-dimensional variations of the equivalent bar pattern approach.[59,82,83] Instead of calculating the number of resolved bars across a critical dimension of a target, some approaches use the number of resolvable pixels on target calculated using a formula such as[59]

$$N_{\text{pixels}} = 2N_{\text{cycles}_x} 2N_{\text{cycles}_y} = \frac{4f_x f_y L_{\text{targ}} H_{\text{targ}}}{R^2} , \tag{2.29}$$

where

f_x, f_y = just resolvable frequencies determined from the MRT in each direction
R = range in kilometers
$L_{\text{targ}}, H_{\text{targ}}$ = target length, height in meters.

The shape of a resolvable pixel will, in general, reflect any asymmetry in the threshold resolution capability of the imager. The size of the resolvable pixel, $1/(4f_x f_y)$, will change depending on the apparent thermal contrast. A TTPF curve relating number of resolvable pixels on target to probability of performing a task is then used.

A second, similar approach is to measure the MRT in each direction and then average the two. A TTPF curve using cycles across a particular dimension of a target can then be used (the dimension is sometimes chosen to be the square root of the area), the N_{50} point being adjusted from that in Fig. 2.20 depending on the averaging technique used and the choice of target dimension.

An updated version[84,85] of the static performance model discussed in detail in Sec. 2.5.2 was issued in June 1990. The model update consisted of two parts: "FLIR90," which predicts laboratory measures, and "Acquire," which predicts field performance. In FLIR90, predicted or measured horizontal and vertical MRTs are averaged at a particular temperature using

$$f_{\text{eff}} = (f_x \cdot f_y)^{1/2} . \tag{2.30}$$

The "effective" MRT is used along with modified values of N_{50} for the different discrimination tasks to predict range performance. The N_{50} for a particular task using the more recent 2-D version of the model is found by multiplying the original "1-D" N_{50} values by 0.75. The amount of shift was determined by requiring the range predictions for the 2-D model to predict correctly the results of the field tests, which served as validation for the original model.

The target model used in this version of the model uses a rectangular target characterized by $A_T^{1/2}$ instead of the smallest dimension as in the previous version. This target description is thought to be appropriate for tactical army targets that are normally near square, yet the description may need refinement

for targets with drastically different aspect ratios such as ship targets. The influence of aliasing is handled in the model update by assuming the MRT cuts off at f_N. While the MRT predictions still tend to be slightly optimistic at low spatial frequencies and pessimistic at high spatial frequencies, the agreement with MRT measurements is significantly better than the original static performance model.

The original version of the performance model relied on NET to completely describe the system noise. Measurement of prototype second-generation systems showed that MRT measurements could be dominated by different types of noise such as pronounced readout noise or the streaking due to nonuniformity. Along with FLIR90 a framework is provided for the description of noise terms in the MRT that account for what was termed *directional noise*, that is, measured noise or nonuniformity defined in a three-dimensional coordinate system (temporal, horizontal spatial, and vertical spatial),[80] and a standard laboratory measurement procedure[86] is specified. The noise analysis methodology isolates system noise into eight directional components. Each component is multiplied by the appropriate temporal or spatial integration factors, and horizontal and vertical components are summed to get correction factors in the MRT equation. The influence of each of the noise components on both MRT and field performance is being actively investigated.

There have been some efforts to quantify the performance of staring systems. Cantella[77] expresses typical system-level descriptors such as MRT in terms of electron density at the focal plane. Nonuniformity is accounted for as an additive noise term. Aliased signal and noise is not dealt with. Rosell[78,87] develops expressions for NET and MRT of staring arrays that include the influence of nonuniformity in a similar manner. Shaham and Woody[88] present an MRT that includes an aliased noise term. For the systems they considered, the sampling rate is assumed to be high enough that signal aliasing can be neglected. Kennedy[44,59] treats aliased signal and noise as well as other characteristics such as nonuniformity in his 2-D model. A more recent update[89,90] of the Night Vision and Electro-Optics Directorate model, FLIR92, includes the capability to model staring systems. Sample-scene phasing effects are accounted for using an averaged MTF. Predictions are not allowed using MRT frequencies beyond the Nyquist frequency in order to account for aliasing.

It has been pointed out that most of the so-called two-dimensional models are really two directional instead. That is, threshold resolution performance is predicted or measured along two directions 90 deg apart; resolution of the imager along intermediate directions is not used in the models. Such two-directional models would predict no advantage for hexagonal sampling schemes, which have been shown to be the best sampling scheme for natural scenes.[72]

Initial steps have been taken to add to the validation data base for the 2-D models using the field test results of second-generation thermal imagers. Given recent improvements in detector technology and digital processing, the rapid pace of development of EO sensor technology will continue to provide systems analysts with a challenge for the future.

2.6 SEARCH PERFORMANCE MODELING[91–96]

Bloomfield[92] categorizes several situations in which an observer would need to perform visual search. One condition requiring search has been termed

competition: the circumstance where the target and other nontargets (for example, other vehicles or parts of the background) resemble each other to some degree. Partial obscuration of the target by smoke, defilade, or camouflage will increase the degree of competition and, therefore, the difficulty of the search task. Another situation where search is needed arises when the target is so close in luminance to the background that it is near the threshold of detectability, a condition that has been termed, logically, a *threshold situation*. Also, search would be needed even if the target could be easily discerned from the background but, by virtue of its very small size, was difficult to find.

2.6.1 The Search Process

The previous discussion of static performance focused on prediction of tasks where the target position is approximately known and the time required to perform the task is not a critical variable. For static performance tasks it is assumed that the observer can take as long as needed to perform a given task. In this section emphasis is on the prediction of search tasks where target position is unknown and time is a critical variable.

A general search task will often involve an EO system, the LOS of which can be panned through specified angles in elevation and azimuth either automatically or manually. The segment of object space that can be imaged by virtue of this panning forms the system field of regard (FOR). The search field is defined as "the area of object space which is physically covered by the device in quest of the target, to the exclusion of other areas."[11] Since the FOR will frequently include regions where the target cannot possibly exist (i.e., a tank in the sky or on a very steep slope), the search field will often be smaller than the system FOR.

The process of location of a target using an EO system can often be conveniently partitioned into two main processes: physical search, or panning the system FOV to examine the FOR, and display search (sometimes called FOV search), which involves visually sampling the EO display. Figure 2.27 is a flow diagram of a simplified search procedure involving display search for a single target in a search field. It has been noticed in several search experiments that when an image is first presented to an observer, they scan the whole image quickly to orient themselves in the image. Any contextual features of the scene found during the orientation that can be used to reduce the size of the search field will lead to reduced mean search times. For example, an observer looks for ground vehicles first along or near a road.

The search process proceeds with a series of visual interrogations of fixation points and decisions as to the presence of a target. If the target has not been found after an investigation of the FOV, the sensor LOS is repositioned, the observer orients him- or herself in the portion of the search field being displayed, and the process of display search repeats itself.

Eye Movements and Visual Search.[95,97,98] The human eye searches a display by aiming the fovea at different locations in an image, rapidly moving from one fixation point to another. Each fixation lasts about a third of a second on the average. The large jumps are called *saccades*; a saccade plus a fixation is termed a *glimpse*. The saccades last only a few milliseconds and can reach speeds of 1000 deg/s. During the saccadic motion, vision is greatly reduced.

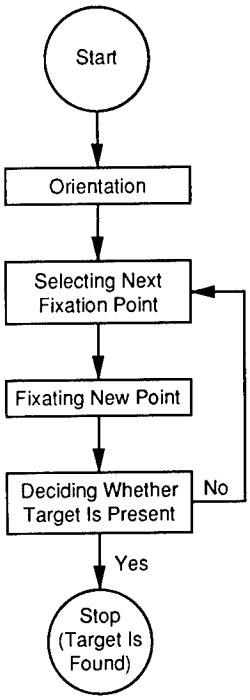


Fig. 2.27 Search flow diagram (after Ref. 97).

The speed and length of the saccade are dependent on the task, being faster for situations where, for example, the allowed time to perform the search is limited to a short interval, and being physically shorter when the density of information in the image is higher. For more complicated scenes, the fovea will tend to interrogate areas of high information content. The choice of the areas to interrogate will be governed in part by the observer's expectations and by the information desired from the image. Obviously the characteristic motion of the eye will play a central role in the ability of an observer to perform visual search; therefore, models of visual search must explicitly or implicitly account for some aspects of typical eye motion.

2.6.2 Factors That Influence Search Performance

As emphasized in Table 2.2, numerous factors influence an observer's ability to acquire a target. Only some of the major considerations are discussed here.[92,99–101]

For competition search, as would be expected, the more the target differs from the nontargets in contrast,[102] shape, or size,[103–105] the faster the target will be found. This assumes that other factors such as the number of targets and the target density are kept constant. Any other factors that would act to increase the target discriminability would also lead to decreased mean search times.

Also, as might be expected, increasing the number of nontargets, their homogeneity, or the size of the search field will generally result in increased mean

search times. Any change to the displayed image that would require an observer either to examine more objects, or to examine the objects in greater detail, will lengthen the search process.

Kaplan and Carvellas[106] found that an observer, while first training on a set of targets (in their case, letters of the alphabet), will scan more slowly while searching through a list for one of many targets. Neisser[107] points out that after learning the target set thoroughly an observer can scan through a list looking for one of many targets in a target set just as fast as when scanning for a single target. Yonas and Pittinger[108] state that even though the scan rate for multiple targets may be the same after learning, the error rates increase. Practice also can improve the performance of tasks such as threshold detection that do not involve learning the shapes or characteristic features of targets.

If the observer has been prebriefed so that he or she knows probable positions or states of the target, mean search times will be shorter. On the other hand, if a target does not appear very much like the mental image of the target that has been learned by the observer, the search process will be more difficult. If the target is camouflaged or partially occluded, the target discriminability from background decreases and the search process becomes correspondingly more difficult.[109]

Search strategies for the motion of the sensor LOS in a physical search task have been shown to have a large influence on total search performance. The recommendations of one study[110] call for a regular physical search pattern that produces systematic and comprehensive coverage of the search field and the use of a variable scan rate that is adjusted to the difficulty of terrain being investigated.

2.6.3 Search Modeling[6,111,112,113]

The mathematical forms of military search models often have their roots in basic perception experiments, many of which were performed to investigate nonmilitary applications of search such as reading or assembly line quality control. Some field testing for specialized military scenarios exists (see Ref. 114, for example), but usually the cost of carrying out extensive, well-instrumented field validation trials in a realistic military environment is prohibitive. Because of this, for specific scenarios such as the tactical ground-to-ground case, only a small amount of field validation exists. Although efforts to model search performance have existed since Koopman's efforts[115] during World War II, visual search prediction still is considered to be in its infancy, and no generally accepted overall model of visual search yet exists.

Despite the lack of maturity of models of visual search, their careful use can still provide valuable insight. The search models put more weight on factors such as EO system FOV, FOR, and gimbal panning capability than do static performance models. Such considerations often turn out to be the major cost drivers of an EO system.

Search is a complex process and the modeling of search is very difficult. In principle, the models must include many factors, either implicitly or explicitly. For instance, in performing a physical search it is important to have a clear line of sight to the target. However, the target could be obscured by terrain,

smoke, or defilade, and in some more complex models, estimates of the probability of occurrence of each of these conditions would be an input to the model.

The search process has been described using several different mathematical formulations.[111] Although it is known that the human searches through a displayed image spending more time on areas of high information content where he or she knows a target is more likely to be, it often is assumed that the scan path that connects a fixation sequence is random. The random sampling of the image can be assumed to proceed with or without memory, that is, whether or not the same spot in an image can be interrogated more than once. The time between fixations can be chosen to vary or remain fixed.

Many models make assumptions of how to calculate the probability of detecting a target given that a fixation lands on or near a target. Based on empirical results, Lamar[114] formulated a relation between threshold contrast and visual angle of the target from the fovea called the *visual lobe*. It is "a set of probability contours, mapping out the probability of detecting the target at various eccentricities from the point of fixation."[111] The lobe can incorporate many factors that would influence a single glimpse detection probability such as peripheral visual acuity and display characteristics.

Several search models have been formulated that incorporate different assumptions and hypothetical constructs. Reference 6 provides an excellent survey of many. In this chapter, we do not explore the more complex of these models. Instead, we present a simpler model in more detail. The model was developed at the Night Vision and Electronic Sensors Directorate (NVESD) and is presently the most commonly used search model for military applications.

The NVESD Search Model. As an example of a search model that predicts both physical and display search, the main features of the NVESD search model[112] are presented. The model predicts the ensemble probability of locating a single target in a search field using an EO imager. Most of the field tests that serve as validation for the model employed thermal imagers searching for ground targets. The model is used widely within the U.S. Army community as an important component of wargames. It is used by other countries as well. Parameters used in the search model are calculated using results of the static performance model discussed previously.

Display Search. For this submodel it is assumed that an observer is viewing the display of an EO imaging system with a fixed LOS. The visual search process is assumed to proceed as a series of random, independent glimpses, each glimpse being characterized by a probability of detection. For each combination of observer, target, and background, a conditional probability is assumed that the observer will acquire the target given the target has not been found previously and given that the target will eventually be found. For P_i equal to the probability of finding the target in the i'th fixation, the conditional probability of acquiring the target in j fixations is given by

$$P(j) = 1 - \prod_{i=1}^{j} (1 - P_i) \ . \tag{2.31}$$

If P_i is not a function of i ($P_i = P$) and P is small, j can be approximated by t/t_f where t_f is the mean fixation time (≈ 0.3 s), and then

$$\begin{aligned} P(j) = P(t) &= 1 - (1 - P)^{t/t_f} \cong 1 - \exp(-Pt/t_f) \\ P(t) &= 1 - \exp(-t/\tau_{\mathrm{FOV}}) \ , \end{aligned} \tag{2.32}$$

where τ_{FOV} ($= t_f/P$) is the mean acquisition time given that the target is found. This is the formula for the cumulative probability of the first arrival time for a Poisson process and is a standard search equation seen throughout the literature.[116]

For an ensemble of observers the equation is

$$P(t) = P_\infty[1 - \exp(-t/\tau_{\mathrm{FOV}})] \ . \tag{2.33}$$

Here the single glimpse probability of finding a particular target has been assumed to be either zero or some value P_0 for every member of the ensemble, and $\tau_{\mathrm{FOV}} = t_f/P_0$. The asymptotic probability P_∞ is the portion of observers in the ensemble that can find the target given unlimited time. It is a measure of the difficulty of acquiring a target, given it is in the FOV. Probability P_∞ arises due to the experimentally observed fact that not all observers in an ensemble will always be able to find the target given sufficient time. It differs from P_D for discrimination detection in that P_∞ involves trying to locate the target, whereas in static detection it is assumed that the position of the target is approximately known. The $P(t)$ is interpreted as the portion of the ensemble of observers that can find the target in time t. See Fig. 2.28 for plots of $P(t)$ for different τ_{FOV} and P_∞.

The equation for P_∞ is very similar to the equation for static detection probability, differing only in the N_{50} values used:

$$P_\infty = \frac{[N/(N_{50})_D]^E}{1 + [N/(N_{50})_D]^E} \ , \tag{2.34}$$

where, as before, N is the number of resolvable cycles across the target, E is $2.7 + 0.7\,[N/(N_{50})_D]$, and now $(N_{50})_D$ is the number of cycles required for 50% acquisition probability.

Recall that for static performance prediction, N_{50} reflects the difficulty of performing static discrimination tasks where the approximate location of the target is known. But $(N_{50})_D$ reflects the difficulty of finding a target in various levels of clutter. The determination of the appropriate $(N_{50})_D$ for a particular search scenario is made using only an approximate set of guidelines. Representative values of $(N_{50})_D$ are shown in Table 2.5. Note that these values of $(N_{50})_D$ are appropriate for use with the 1-D static performance model described earlier. As mentioned previously, the N_{50} values appropriate to the 2-D model were 0.75 times those for the 1-D case. This factor also should be applied to the $(N_{50})_D$ values quoted above if the 2-D static performance model is used to provide input to this search model.

After analysis of field test results, the single glimpse probability of search detection, P_0, was proportional to P_∞. This seems intuitive since one would

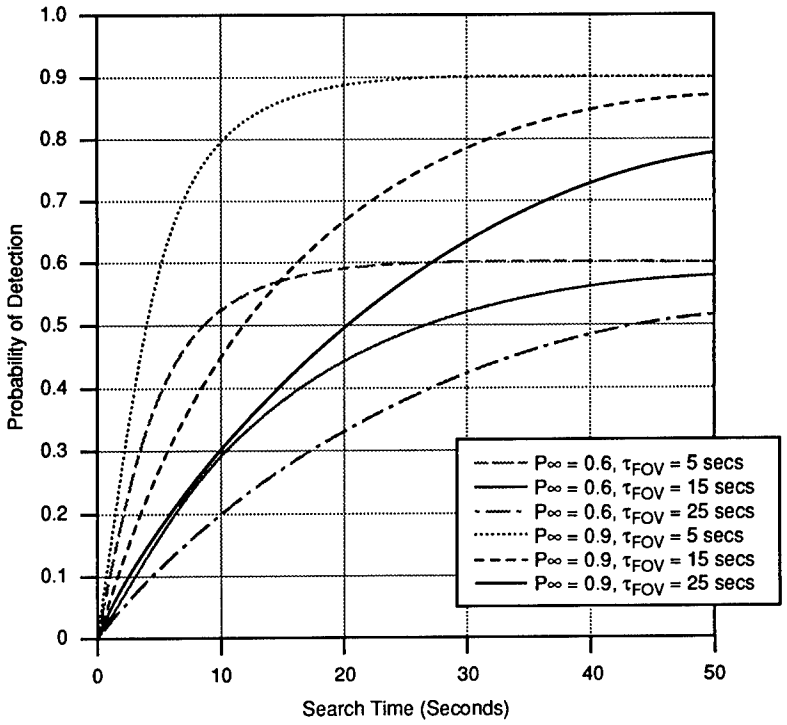


Fig. 2.28 Search detection probability as a function of time for different τ_{FOV} and P_∞.

expect the single glimpse probability of locating the target to be larger for those easier search tasks in which a higher portion of the observers was able to find the target. The approximation that is used for the case of $P_\infty \leq 0.90$ and low to medium clutter is that

$$\frac{1}{\tau_{FOV}} = \frac{P_0}{t_f} = \frac{P_\infty}{3.4} \,. \tag{2.35}$$

For $N/(N_{50})_D < 2$, P_∞ can be approximated by $N/2(N_{50})_D$, and substitution into the equation above gives

$$\frac{1}{\tau_{FOV}} = \frac{P_0}{t_f} = \frac{1}{6.8}\frac{N}{(N_{50})_D} \,. \tag{2.36}$$

Table 2.5 Approximate Number of Cycles Required to Perform Target Acquisition, $(N_{50})_D$, in Various Levels of Clutter

Task	Example	$(N_{50})_D$
Highly conspicuous target	Bright source, movement, zero clutter	$\ll 0.5$
Low clutter	Target in field, on road	0.5
Medium clutter	Tank in a desert of tank-size bushes	1.0
High clutter	Vehicle in array of similar vehicles	2.0

Using these assumptions, one can evaluate τ_{FOV} using Eq. (2.35) or (2.36) for use in Eq. (2.33) for FOV search above.

The steps involved in the calculation of the ensemble probability of FOV search performance for a particular time t are summarized as follows:

1. Calculate N, the number of resolvable cycles across the target critical dimension, using the procedure described in Sec. 2.5.2 on static performance modeling.
2. Estimate a value for $(N_{50})_D$, the number of resolvable cycles across the target critical dimension required for 50% of the observer population to acquire the target using guidance from Table 2.5.
3. Using the equation

$$P_\infty = \frac{[N/(N_{50})_D]^E}{1 + [N/(N_{50})_D]^E} , \tag{2.37}$$

   determine the asymptotic search probability P_∞.
4. Determine τ_{FOV} using

$$\frac{P_0}{t_f} = \frac{1}{\tau_{FOV}} = \begin{cases} \dfrac{P_\infty}{3.4} , & \text{for } P_\infty \le 0.9 \\ \dfrac{1}{6.8}\dfrac{N}{(N_{50})_D} , & \text{for } P_\infty > 0.9 \end{cases} . \tag{2.38}$$

5. Calculate the ensemble acquisition probability at time t using

$$P(t) = P_\infty[1 - \exp(-t/\tau_{FOV})] . \tag{2.39}$$

Though the form of this model is very simple, its predictions have been shown to agree with existing field test results.[112] Indeed, the model's simplicity is one of its major strengths. However, there are some significant limitations to the model, some of which could benefit from further lab and field testing. The FOV is not an explicit parameter in the search performance equations although perhaps it should be. It is, however, a parameter required to calculate the number of resolvable cycles across target. Efforts have been made to extend the procedure to the multitarget scenario,[117] to multiple observers and dynamic scenarios,[118] and to cued detection.[119] In its present form the model is not applicable to search with the unaided eye, or even with optically aided eyes (using binoculars, for example). Ideally, the relation between search performance and the amount of clutter should be quantified explicitly and should be backed by extensive testing. Other important areas of future research also remain, for example, search prediction for moving targets and target groupings.

Physical Search. An equation similar to the one used for FOV search is used to predict the ensemble probability of target acquisition as a function of time for the case of an observer systematically panning a thermal imager FOV through the search field. The equation expressing the probability of acquiring the target while performing FOR search is

$$P(t) = P_\infty[1 - \exp(-t/n\tau_{FOV})] \ , \tag{2.40}$$

where n is the number of device FOVs in the search FOR. This equation is applicable for the case where the target is not found quickly.

Reference 112 describes a slightly more complicated method of calculating physical search probability as a function of time. It involves the calculation of the probability of detecting the target in an integer number of scans through the FOR. If one desires a prediction for a noninteger number of scans, it is assumed that the detection probability can be linearly interpolated from the two adjacent probabilities, which correspond to integer numbers of scans.

The procedure for calculating the probability of performing FOR search detection is very simple and is summarized as follows:

1. Determine P_∞ and τ_{FOV} as in the FOV search procedure described in the previous section.
2. Calculate the number of device FOVs required to cover the FOR using

$$\frac{\text{FOR}}{\text{FOV}} = n$$

and substitute into

$$P(t) = P_\infty[1 - \exp(-t/n\tau_{FOV})] \ ,$$

which is the equation for the ensemble probability of finding the target.

As for the FOV search case this is a relatively simple equation that describes the FOR search performance for the case of systematic panning of the FOR.

Example. For the purposes of this example, we will assume that a medium clutter level exists. (Refer back to Sec. 2.5.1 for a description of the example system.) For a target with an inherent ΔT of 1.25°C, at 6 km the averaged transmission of the "good" atmosphere is 51% leading to an apparent ΔT of 0.64°C. From the MRT the maximum resolvable spatial frequency corresponding to the apparent ΔT is 4.8 cycles/mrad. The number of resolvable cycles across the target is then 2.56 cycles. For medium clutter the $(N_{50})_D$ value, from Table 2.5, is 1. The resulting P_∞ value is 0.99, and τ_{FOV} is 2.65 s.

The "poor" atmosphere results in an apparent temperature difference of 0.093°C, a maximum resolvable frequency of 2.51 cycles/mrad, and 1.34 cycles across the critical dimension of the target. The P_∞ for this case is 0.74, and τ_{FOV} is 4.52 s. Probability of display search detection versus time is plotted for both atmospheres in Fig. 2.29.

Assuming the search field is 12 deg^2 (= 5 × FOV), and using the equation above for an FOR search, the probability of FOR search versus time is calculated. Figure 2.30 shows the FOR search probability versus time for both atmospheres.

Acknowledgments

I wish to thank the many colleagues who read through this chapter and gave helpful comments and suggestions. Special thanks goes to Walt Lawson and

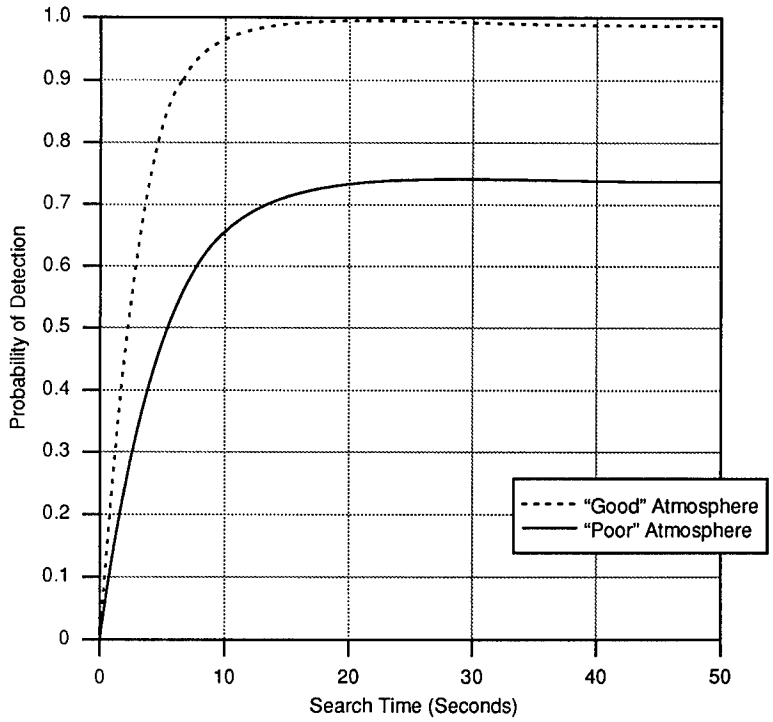


Fig. 2.29 Probability of FOV search detection versus time for both "good" and "poor" atmospheres.

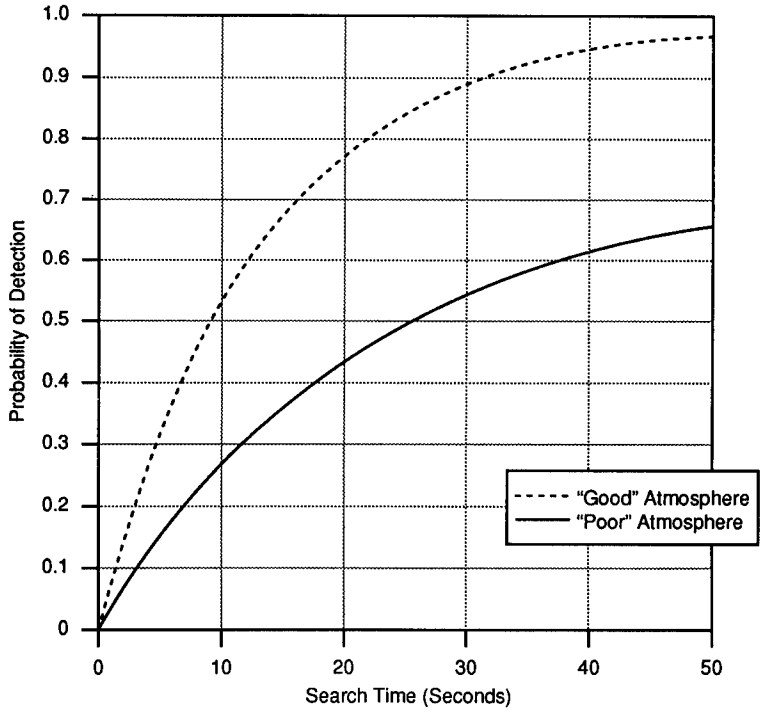


Fig. 2.30 FOR search probability versus time for both "good" and "poor" atmospheres.

Jim Ratches who have not only provided reviews of this chapter, but have long been sources of friendly advice and support. The views, opinions, and/or findings contained in this chapter are those of the author and not necessarily those of the U.S. Army.

References

1. F. Rosell and G. Harvey, Eds., "The fundamentals of thermal imaging systems," NRL Report 8311, ADA 073 763 (1979).
2. J. M. Lloyd, *Thermal Imaging Systems,* 2nd ed., Plenum Press, New York (1975).
3. J. A. Ratches et al., "Night vision laboratory static performance model for thermal viewing systems," ECOM-7043 (1975).
4. L. M. Biberman, Ed., *Perception of Displayed Information,* Plenum Press, New York (1973).
5. I. Overington, *Vision and Acquisition,* Pentech Press, London (1976).
6. C. P. Greening, "Mathematical modeling of air-to-ground target acquisition," *Human Factors,* **18**(2), 111–148 (1976).
7. D. L. Schumaker, J. T. Wood, and C. D. Thacker, *FLIR Performance Handbook,* DCS Corporation (1988).
8. D. B. Jones et al., Eds., "Air-to-ground target acquisition source book: a review of the literature," ADA 015079, Office of Naval Research, Arlington, VA (1974).
9. W. D. Bliss in "Air-to-ground target acquisition source book: a review of the literature," D. B. Jones et al., Eds., ADA 015079, Office of Naval Research, Arlington, VA (1974).
10. "Department of Defense dictionary of military and associated terms," Joint Chiefs of Staff, Washington, DC (1987).
11. Quadripartite Working Group on Army Operational Research, *Search and Target Acquisition Nomenclature,* Quadripartite Advisory Publication, Special Working Party on the Modeling of Target Acquisition (May 1990).
12. F. X. Kneizys et al., "Users guide to LOWTRAN 7," AFGL-TR-88-0177, Air Force Geophysics Laboratory, Hanscom AFB, MA (Aug. 1988).
13. J. D. Howe et al., "Thermal model improvement through perception testing," *Proceedings of IRIS Specialty Group on Passive Sensors,* Infrared Information Analysis Center, ERIM, Ann Arbor, MI (1989).
14. J. D'Agostino et al., "An experimental study of the effects of sampling on FLIR performance," *Proceedings of the IRIS Specialty Group on Passive Sensors,* Infrared Information Analysis Center, ERIM, Ann Arbor, MI (1990).
15. L. Obert et al., "An experimental study of the effect of vertical resolution on FLIR performance," *Proceedings of the IRIS Specialty Group on Passive Sensors,* Infrared Information Analysis Center, ERIM, Ann Arbor, MI (1990).
16. J. W. Goodman, *Introduction to Fourier Optics,* McGraw-Hill, New York (1968).
17. G. Westheimer, "The eye as an optical instrument," in *Handbook of Perception and Human Performance: Vol. I. Sensory Processes and Perception,* K. R. Boff, L. Kaufman, J. P. Thomas, Eds., John Wiley & Sons, New York (1986).
18. G. A. Fry, "The eye and vision," in *Applied Optics and Optical Engineering, Vol. II.* R. Kingslake, Ed., Academic Press, New York (1965).
19. C. H. Graham, Ed., *Vision and Visual Perception,* John Wiley & Sons, New York (1965).
20. K. R. Boff, L. Kaufman, and J. P. Thomas, Eds., *Engineering Data Compendium, Human Perception and Performance,* John Wiley & Sons, New York (1986).
21. F. S. Werblin, "The control of sensitivity in the retina," *Scientific American* **228**(1), 70–79 (1973).
22. D. Hood and M. Finkelstein, "Sensitivity to light," in *Handbook of Perception and Human Performance: Vol. I, Sensory Processes and Perception,* K. R. Boff, L. Kaufman, J. P. Thomas, Eds., John Wiley & Sons, New York (1986).
23. R. Kingslake, Ed., *Applied Optics and Optical Engineering, Vol. II,* p. 202, Academic Press, New York (1965).
24. R. J. Farrell and J. M. Booth, *Design Handbook for Imagery Interpretation Equipment,* D180-19063-1, Boeing Aerospace Company, Seattle, WA (1984).

25. H. B. Barlow, "Temporal and spatial summation in human vision at different background intensities," *Journal of Physiology* **141,** 337–350 (1958).

26. A. B. Watson, "Temporal sensitivity," in *Handbook of Perception and Human Performance: Vol. I. Sensory Processes and Perception.* K. R. Boff, L. Kaufman, J. P. Thomas, Eds., John Wiley & Sons, New York (1986).

27. C. H. Graham, Ed., *Vision and Visual Perception,* p. 130, John Wiley & Sons, New York (1965).

28. J. M. Lloyd, *Thermal Imaging Systems,* 2nd ed., p. 132, Plenum Press, New York (1975).

29. H. de Vries, "The quantum character of light and its bearing upon threshold of vision," *Physica* **10,** 7 (1943).

30. A. Rose, "The sensitivity performance of the human eye on an absolute scale," *Journal of the Optical Society of America* **38**(2), 196–208 (1948).

31. O. H. Schade, "The resolving-power functions and quantum processes of television cameras," *RCA Review* **28**(3), 460–535 (1967).

32. F. A. Rosell, "Limiting resolution of low-light-level imaging sensors," *Journal of the Optical Society of America* **59**(5), 539–547 (1969).

33. F. A. Rosell and R. H. Willson, "Performance synthesis (electro-optical sensors)," Technical Report AFAL-TR-71-137, AD 884829, Air Force Avionics Laboratory, Wright Patterson Air Force Base, OH (May 1971).

34. R. A. Rosell and R. H. Willson, "Performance synthesis of electro-optical sensors," EOTM 575, Night Vision Laboratory, U.S. Army Electronics Command, Ft. Belvoir, VA (Feb. 1975).

35. D. G. Peli, "Effects of visual noise," PhD dissertation, Cambridge University (1981).

36. T. N. Cornsweet, *Visual Perception,* Academic Press, Orlando, FL (1970).

37. G. C. Holst and A. R. Taylor, "What eye model should we use for MRT testing?," *Proceedings of the SPIE,* **1309,** 67–75 (1990).

38. G. H. Kornfeld and W. R. Lawson, "Visual-perception models," *Journal of the Optical Society of America* **61**(6), 811–820 (June 1971).

39. J. A. Leslie et al., "Reference and user's manual for NV&EOL, visionics division TV performance model," AMSEL-NV-TR-0054, Night Vision and Electro-Optics Laboratory, Ft. Belvoir, VA (1987).

40. W. R. Lawson, "Image intensifier evaluation," Master's thesis, University of Rochester, New York (1968).

41. J. A. Leslie et al., "Reference and user's manual for NV&EOL, visionics division image intensifier performance model," Release 1.2, Night Vision and Electro-Optics Laboratory, Ft. Belvoir, VA (1985).

42. "The image intensifier MRC and MDC model," Final Report, DCS Corporation (Dec. 1991).

43. "NV&EOL image evaluation methodology," Technical Memorandum, Night Vision and Electro-Optics Laboratory, Ft. Belvoir, VA.

44. H. V. Kennedy, "Modeling second-generation thermal imaging systems," *Optical Engineering* **30**(11), 1771–1778 (Nov. 1991).

45. J. A. Ratches, "Static performance model for thermal imaging systems," *Optical Engineering* **15**(6), 525–530 (1976).

46. O. H. Schade, Sr., "An evaluation of photographic image quality and resolving power," *Journal of the Society of Motion Picture and Televison Engineers* **73,** 81–120 (Feb.1964).

47. J. W. Coltman and A. E. Anderson, "Noise limitations to resolving power in electronic imaging," *Proceedings of the IRE* **48**(5), 858–865 (1960).

48. J. M. Lloyd and R. L. Sendall, "Improved specifications for infrared imaging systems," *Proceedings of IRIS Imaging Symposium,* pp. 109–129, Infrared Information Analysis Center, ERIM, Ann Arbor, MI (1970).

49. J. A. Ratches and W. R. Lawson, "FLIR modeling fundamentals," in *FLIR Modeling Workshop I: Performance Summary Measures,* J. A. Ratches, C. D. Nguyen, Eds., C2NVEO Report (1990).

50. Y. J. Shaham, C. D. Capps, and L. M. Candell, "FLIR performance modeling considerations," *Proceedings of IRIS Imaging Symposium,* Infrared Information Analysis Center, ERIM, Ann Arbor, MI (1980).

51. K. C. Hepfer, "A new TIS technical performance model," *Proceedings of IRIS Imaging Symposium,* Infrared Information Analysis Center, ERIM, Ann Arbor, MI (1981).

52. F. A. Rosell, "Review of the current periodic sensor models," *Proceedings of IRIS Imaging Symposium,* Infrared Information Analysis Center, ERIM, Ann Arbor, MI (1981).

53. J. Johnson, "Analysis of image forming systems," *Proceedings of the Image Intensifier Symposium,* October 6–7, 1958, AD 220 160, U.S. Army Engineer Research and Development Lab, Ft. Belvoir, VA, pp. 249–273 (1958).

54. C. P. Greening, "Target acquisition model evaluation final summary report," NWC TP 5536, U.S. Naval Weapons Center, China Lake, CA (June 1973).

55. C. P. Greening, "Approaches to modeling visual target acquisition," NWC TP 5698, Naval Weapons Center, China Lake, CA (Sep. 1974).

56. W. R. Lawson and J. A. Ratches, "Modeling detection or the detection game," *Proceedings of IRIS Imaging Symposium,* Infrared Information Analysis Center, ERIM, Ann Arbor, MI (1980).

57. A. van Meeteren, "Characterization of task performance with viewing instruments," *Journal of the Optical Society of America A* **7,** 2016–2023 (1990).

58. J. Johnson and W. R. Lawson, "Performance modeling methods and problems," *Proceedings of IRIS Imaging Symposium,* pp. 105–123, Infrared Information Analysis Center, ERIM, Ann Arbor, MI (1974).

59. H. V. Kennedy, "Two-dimensional modelling of FLIR systems," *Proceedings of 1983 Meeting of IRIS Specialty Group on Infrared Imaging,* Infrared Information and Analysis Center, ERIM, Ann Arbor, MI (1983).

60. P. M. Moser, "Mathematical model of FLIR performance," NAVAIRDEVCEN Technical Memorandum NADC-20203:PMM, ADA 045247 (Oct. 19, 1972).

61. G. J. O'Neill, "The quantification of image detail as a function of irradiance by empirical tests," NAVAIRDEVCEN Technical Memorandum NADC-202139:GJO (Jan. 10, 1974).

62. R. T. Flaherty, "Evaluation of the static recognition model with respect to target subtense, target equivalence and learning," Master's thesis, George Mason University (1981).

63. F. A. Rosell, "Levels of visual discrimination for real scene objects vs. bar pattern resolution for aperture and noise limited imagery," *Proceedings of the IEEE National Aerospace and Electronics Conference,* Report 75CN0956-3 NAECON, pp. 327–334 (1975).

64. A. van Meeteren, "Prediction of realistic visual tasks from image quality data," *Proceedings of the SPIE* **98,** 58–64 (1977).

65. A. van Meeteren, "Characterization of task performance with viewing instruments," *Journal of the Optical Society of America A* **7,** 2016–2023 (1990).

66. J. J. Vos and A. van Meeteren, "PHIND: an analytic model to predict target acquisition distance with image intensifiers," *Applied Optics* **30**(8), 958–966 (1991).

67. A. H. Blumenthal and S. B. Campana, "An improved electro-optical image quality summary measure," *Proceedings of the SPIE* **310,** 43–52 (1981).

68. A. H. Blumenthal and S. B. Campana, "Development of an image quality model for object discrimination," *Proceedings of the SPIE* **467,** 24–32 (1983).

69. H. L. Task, "An evaluation and comparison of several measures of image quality for television displays," PhD thesis, AMRL-TR-79-7, ADA 069 690, Aerospace Medical Research Laboratory, Wright Patterson Air Force Base, OH (Jan. 1979).

70. I. Overington, "Towards a complete model of photopic visual threshold performance," *Optical Engineering* **21**(1), 3–13 (1982).

71. *A Symposium on Sampled Images,* IS 10763, Perkin Elmer Corporation (1971).

72. R. Legault, "The aliasing problems in two-dimensional sampled imagery," in *Perception of Displayed Information,* L. M. Biberman, Ed., pp. 279–312, Plenum Press, New York (1973).

73. S. K. Park, R. Schowengerdt, and M.-A. Kaczynski, "Modulation-transfer-function analysis for sampled image systems," *Applied Optics* **23**(15), 2572–2582 (Aug. 1984).

74. F. O. Huck, N. Halyo, and S. K. Park, "Aliasing and blurring in 2-D sampled imagery," *Applied Optics* **19**(13), 2174–2181 (July 1980).

75. S. K. Park, "Image sampling, reconstruction, and the effect of sample-scene phasing," *Applied Optics* **21**(17), 3142–3151 (Sep. 1982).

76. R. A. Rosell, "Solid state imaging sensors," EOTM-577, Westinghouse Electric Corporation (1977).

77. M. J. Cantella, "IR focal plane array system performance modeling," *Proceedings of the SPIE* **327,** 40–58 (Jan. 1982).

78. F. A. Rosell, "Periodic model for platinum silicide staring arrays," *Proceedings of IRIS Imaging Symposium,* Infrared Information Analysis Center, ERIM, Ann Arbor, MI (1987).

79. J. Mooney, "Effect of spatial noise on the minimum resolvable temperature of a staring sensor," *Applied Optics* **30**(23), 3324–3332 (Aug. 1991).

80. J. D'Agostino, "The modeling of spatial and directional noise in FLIR90, part 1: a 3-D noise analysis methodology," *Proceedings of IRIS Passive Sensors,* Infrared Information Analysis Center, ERIM, Ann Arbor, MI (1991).

81. L. G. Williams and J. M. Erickson, "FLIR operator requirements study," AFAL-TR-76-9, ADA 024 288, Air Force Avionics Laboratory, Wright Patterson Air Force Base, OH (1976).

82. Y. J. Shaham, "Two dimensional recognition range model," *Proceedings of the SPIE* **890,** 92–94 (1988).

83. R. Ruger, "Extended range performance model for thermal imagers," *Proceedings of the SPIE* **972,** 238–245 (1988).

84. L. B. Scott, "C2NVEO thermal imaging systems performance module, FLIR90, user's manual, FLIR90 version 0," M5008986, U.S. Army CECOM C2NVEO, Ft. Belvoir, VA (1990).

85. D. Tomkinson, "Acquire user's guide," C2NVEO Report, Ft. Belvoir, VA (1990).

86. C. M. Webb, P. A. Bell, and G. P. Mayott, "Laboratory procedure for the characterization of 3-D noise in thermal imaging systems," *Proceedings of IRIS Passive Sensors,* Infrared Information Analysis Center, ERIM, Ann Arbor, MI (1991).

87. F. A. Rosell, "Predicting the performance of staring arrays," *Proceedings of the SPIE* **1762** (in press).

88. Y. J. Shaham and L. M. Woody, "MRT for second generation systems," *Proceedings of IRIS Imaging Symposium,* Infrared Information Analysis Center, ERIM, Ann Arbor, MI (1981).

89. "NVEOD FLIR92 thermal imaging systems performance model, analyst's reference guide," U.S. Army Night Vision and Electro-Optics Directorate, Ft. Belvoir, VA (1992).

90. L. B. Scott and J. D'Agostino, "NVEOD FLIR92 thermal imaging systems performance model," *Proceedings of the SPIE* **1689,** 194–203 (1992).

91. A. Morris and E. P. Horne, Eds., *Visual Search Techniques: Proceedings of a Symposium Sponsored by the Armed Forces-NRC Committee on Vision,* Publication 712, AD 234 502, National Academy of Sciences, National Research Council, Washington DC (1960).

92. J. R. Bloomfield, "Visual search," PhD thesis, University of Nottingham (Oct. 1970).

93. *Visual Search,* AD 754 327, National Research Council, National Academy of Sciences, Washington DC (1973).

94. J. N. Clare and M. A. Sinclair, Eds., *Search and the Human Observer,* Taylor and Francis Ltd., London (1979).

95. Special issue on "Visual Search and Eye Movements," *Human Factors* **21,** 3 (1979).

96. *Proceedings of the Second International Conference on Visual Search,* Durham, UK (1990).

97. L. G. Williams, "A study of visual search using eye movement recordings, first annual report," AD 629 624, Honeywell, Minneapolis, MN (1965).

98. A. Ford, C. T. White, and M. Lichtenstein, "Analysis of eye movements during free search," *Journal of the Optical Society of America* **49,** 287–292 (1959).

99. "Visual search," in *Engineering Data Compendium, Human Perception and Performance,* K. R. Boff, L. Kaufman, J. P. Thomas, Eds., John Wiley & Sons, New York (1986).

100. Chapter 5 in *Air-to-Ground Target Acquisition Source Book: A Review of the Literature,* D. B. Jones et al., Eds., ADA 015079, Office of Naval Research, Arlington, VA (1974).

101. J. R. Bloomfield, "Studies on visual search," in *Human Reliability and Quality Control,* C. G. Drury, J. G. Fox, Eds., pp. 31–43, Taylor and Francis, London (1975).

102. R. M. Boynton and D. E. Ross, "The effect of background luminance and contrast upon visual search performance," *Illumination Engineering* **66,** 173–186 (1971).

103. L. G. Williams, "The effect of target specification on objects fixated during visual search," *Perception and Psychophysics* **1,** 315–318 (1966).

104. S. W. Smith, "Time required for target detection in complex abstract visual display," AD 256 039, The University of Michigan, Ann Arbor (Apr. 1961).

105. R. M. Boynton and W. R. Bush, "Recognition of forms against a complex background," *Journal of the Optical Society of America* **46**(9), 758–764 (1956).

106. I. T. Kaplan and T. Carvellas, "Scanning for multiple targets," *Perceptual and Motor Skills*

21, 239–243 (1965).

107. U. Neisser, "Visual search," *Scientific American* **210,** 94–102 (1964).
108. A. Yonas and J. Pittinger, "Searching for many targets: an analysis of speed and accuracy," *Perception and Psychophysics* **13,** 513–516 (1973).
109. J. R. Bloomfield, "Visual search with embedded targets: color and texture differences," *Human Factors* **21**(3), 317–320 (1979).
110. J. H. Banks et al., "Improved search techniques with passive night vision devices," Technical Research Report 1169, AD 722 236, U.S. Army BSRL (1971).
111. J. R. Bloomfield, "Theoretical approaches to visual search," in *Human Reliability and Quality Control,* C. G. Drury, J. G. Fox, Eds., pp. 19–29, Taylor and Francis, London (1975).
112. W. R. Lawson, T. W. Cassidy, and J. A. Ratches, "A search prediction model," *Proceedings of IRIS Specialty Group on Imaging,* Infrared Information Analysis Center, ERIM, Ann Arbor, MI (June 1978).
113. T. H. Monk, "Search," Chap. 9 in *Sustained Attention in Human Performance,* J. S. Warm, Ed., pp. 294–321, John Wiley & Sons, New York (1984).
114. M. E. Franklin and J. H. Whittenburg, "Research on visual target detection, part I, development of an air-to-ground detection/identification model," AD 619 275, Human Sciences Research, McLean, VA (June 1965).
115. B. O. Koopman, "Search and screening," OEG Report 56, AD 214 252, Chief of Naval Operations, Washington, DC (1946).
116. E. S. Krendel and J. Wodinsky, "Search in an unstructured visual field," *Journal of the Optical Society of America* **50**(6), 562–568 (1960).
117. S. R. Rotman, E. S. Gordon, and M. L. Kowalczyk, "Modeling human search and target acquisition performance: I. first detection probability in a realistic multitarget scenario," *Optical Engineering* **28**(11), 1216–1222 (1989).
118. S. R. Rotman, "Modeling human search and target acquisition performance: II. simulating multiple observers in dynamic scenarios," *Optical Engineering* **28**(11), 1223–1226 (1989).
119. B. Blecha et al., "Aided vs. unaided search experiments to further our understanding," *Proceedings of Passive Sensors IRIS,* Infrared Information Analysis Center, ERIM, Ann Arbor, MI (1991).
120. W. E. Woodson and D. W. Conover, *Human Engineering Guide for Equipment Designers,* 2nd ed., University of California Press, Berkeley (1964).
121. C. H. Graham and N. R. Bartlett, "The relation of size of stimulus and intensity in the human eye: II, intensity thresholds for red and violet light," *Journal of Experimental Psychology* **24,** 574–587 (1939).

CHAPTER 3

Optomechanical System Design

Daniel Vukobratovich
Optical Sciences Center/University of Arizona
Tucson, Arizona

CONTENTS

3.1 INTRODUCTION

The quality of optical systems depends on the performance of individual surfaces, the surfaces of optical elements (windows, prisms, lenses, and mirrors), and detectors. Optomechanical design is concerned with maintaining the correct position and shape of these surfaces. Optomechanics is different from conventional mechanical engineering due to its emphasis on strain rather than stress. In optomechanics, very small strains, typically about a part per million, are important. This chapter is intended to address the most common optomechanical problems encountered in IR systems. The development is from simple components to more complex systems analysis.

3.2 WINDOWS

As zero-power elements, windows are the simplest components of an optical system. Windows are intended to provide protection for the optical system while it is being subjected to a wide variety of environmental loading conditions. This variety of loading conditions complicates the design of these "simple" elements.

3.2.1 Thermal Effects on Flat Windows

A window exposed to thermal gradients deforms from a stress-free condition. Thermal deformation of the window can induce an error in the wave front passing through the window. For axial and radial thermal gradients, approximations are used to estimate the magnitude of the wave-front error induced. As a rule of thumb, the window must be as thin as possible, and as oversized with respect to the clear aperture as reasonable, to minimize thermal gradient effects. Insulating the edge of the window reduces radial gradient effects. Use of active temperature control of the window edge may be necessary in some cases.

Axial Temperature Gradients. A linear temperature gradient through the thickness of a window causes the window to deform into a spherical shape. If the window is initially a plane parallel plate with unconstrained edges, the radius of the curvature of the thermal-gradient-induced spherical shape is given by:

$$R = \frac{K}{\alpha q}, \tag{3.1}$$

where

R = thermally induced radius of curvature
α = thermal coefficient of expansion of the window material
q = heat flux per unit area absorbed by the window
K = thermal conductivity of the window material.

The ratio of thermal coefficient of expansion to thermal conductivity is a material parameter called the *thermal distortion parameter,* and is given in units of meters per watt. For minimum thermal-gradient-induced distortion, the thermal distortion parameter should be as small as possible.

After distortion into a spherical shape, a circular window acts as a weak meniscus lens. The power of this meniscus lens is given by[1]:

$$\frac{1}{f} = \frac{n-1}{n}\left(\frac{\alpha}{K}\right)^2 hq^2 \ , \tag{3.2}$$

where

$1/f$ = power of the distorted window (reciprocal of the focal length)
n = index of refraction of the window material
α/K = window material thermal distortion parameter
h = window thickness.

Radial Temperature Gradients. A linear radial temperature gradient from center to edge of a circular window induces a change in optical pathlength for a wave front passing through the window. Assume that the window has no edge loading (simple support) and that the value of the thermo-optic coefficient (variation in window index of refraction with temperature) is constant for small temperature gradients. Then, if second-order terms are neglected, the optical path difference is estimated by[1]:

$$\text{OPD} = \left[(n-1)(1+\nu)\alpha + \frac{dn}{dT}\right] h\Delta T \ , \tag{3.3}$$

where

OPD = optical path difference induced in the wave front
ν = Poisson's ratio for the window material
dn/dT = thermo-optic coefficient
ΔT = temperature gradient from center to edge in the window.

3.2.2 Pressure Effects on Flat Windows

Pressure differentials induce deflection in an initially flat window. Deflection due to pressure in the window causes a change in the optical pathlength in the wave front passing through the window. Stress is induced in the window by the pressure differential. Generally, stress in the window due to the pressure differential is a more serious concern in the performance of IR systems than the optical path difference.

Pressure-Induced Window Deformation. The maximum optical path difference for a wave front passing through a circular window at normal incidence, when the window is deformed by a pressure differential, is given by[2]:

$$\text{OPD} = 8.89 \times 10^{-3}\left[\frac{(n-1)\Delta P^2 d^6}{E^2 h^5}\right] \ , \tag{3.4}$$

where

OPD = maximum optical path difference due to the pressure differential

ΔP = axial pressure differential in the window
d = window diameter
E = elastic modulus of the window material.

Window Stress Due to Pressure Differentials. For brittle IR materials, the maximum tensile stress due to a pressure differential should not exceed

$$\sigma_{\max} = \frac{\sigma_F}{\mathrm{SF}}, \tag{3.5}$$

where

$\sigma_{\max}$ = maximum allowable tensile stress
σ_F = fracture stress of the window material
SF = safety factor.

A common practice in optomechanical engineering is to use a safety factor of 4. An increase in effective inertia loading by a factor of 2 is used to allow for suddenly applied loading, and another safety factor of 2 is used for stress concentration effects near the window edge.

For a simply supported circular window, the minimum ratio of window thickness to diameter, or aspect ratio, is given by:

$$\frac{d}{h} = 2\left[\frac{8\sigma_F}{3\Delta P(3 + \nu)\mathrm{SF}}\right]^{1/2}, \tag{3.6}$$

where d/h is the window aspect ratio and σ_F is the maximum allowable tensile stress.

For a clamped circular window, the minimum aspect ratio is given by:

$$\frac{d}{h} = 2\left[\frac{8\sigma_F}{3\Delta P(1 + \nu)\mathrm{SF}}\right]^{1/2}. \tag{3.7}$$

The normal conservative design practice is to assume a simply supported window. This assumption should be used for both circular and rectangular windows.

For a rectangular window, the stress equation is used to find the minimum thickness necessary to withstand a given external pressure. If the window is simply supported, the minimum thickness is given by:

$$h = b\left\{\frac{\Delta P}{\sigma_F}\frac{3}{4[1 + 2(b/L)^3]}\right\}^{1/2}, \tag{3.8}$$

where

h = minimum window thickness
b = unsupported window width
L = unsupported window length.

If the window is clamped, the minimum thickness is given by:

$$h = b\left\{\frac{\Delta P}{\sigma_F}\frac{1}{2[1 + (b/L)^4]}\right\}^{1/2} . \tag{3.9}$$

3.2.3 Dome Stress Due to Pressure Differentials

Use of a dome streamlines the window and permits a very wide field of view. For design, dome deformation is normally not considered as important as the strength of the dome. Normally, dome stress at any radius is evaluated using the Lamé pressure vessel equations given by[3]:

$$\sigma_m = \sigma_h = -\Delta P\frac{R_o^3(R_i^3 + 2R^3)}{2R^3(R_o^3 - R_i^3)} , \tag{3.10}$$

$$\sigma_r = -\Delta P\frac{R_o^3(R^3 - R_i^3)}{R^3(R_o^3 - R_i^3)} , \tag{3.11}$$

where

σ_m = meridional membrane stress
σ_h = hoop membrane stress
σ_r = radial wall stress
ΔP = external pressure
R_o = outer radius of the dome
R_i = inner radius of the dome
R = radius of the dome.

If the included angle of the dome exceeds 180 deg (hyper-hemisphere), as a rule of thumb assume that the meridional membrane stress is 100% greater than the hoop membrane stress. Dome geometry is shown in Fig. 3.1.

If the ratio of the dome radius to thickness is 10 or more, then the dome is said to be thin. For thin domes, a simplified set of equations is used to determine stresses. For an externally pressurized dome these stresses are given by:

$$\sigma_m = \Delta P\left(\frac{R}{h}\right)\frac{1}{1 + \cos\phi} , \tag{3.12}$$

$$\sigma_h = \Delta P\left(\frac{R}{h}\right)\left(\frac{1}{1 + \cos\phi} - \cos\phi\right) , \tag{3.13}$$

where h is the dome thickness and ϕ is the angle from the vertex of the dome.

Both the Lamé equations and the simplified equations hold for a self-weight loaded dome. A dome exposed to high-acceleration loading can be tested using external pressure to produce an equivalent stress condition in the dome material. In the preceding equations, the external pressure is replaced by an equivalent external pressure found using the following equation:

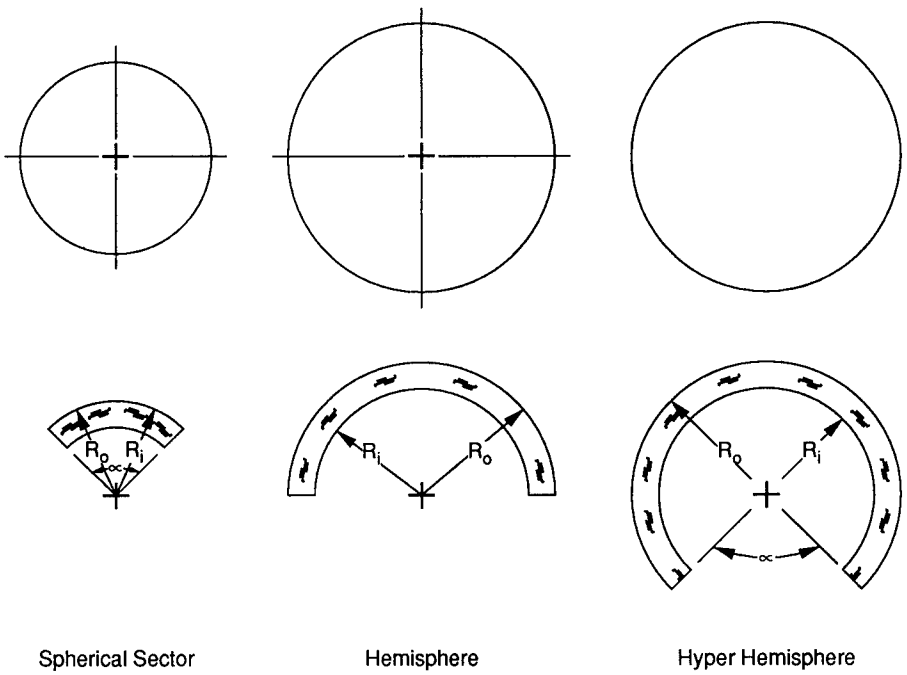


Fig. 3.1 Dome geometry.

$$\Delta P = \frac{a}{g}\rho h \ , \tag{3.14}$$

where

ΔP = equivalent external pressure
g = acceleration due to the Earth's gravity
a = acceleration
ρ = dome material density.

A dome under external pressure may fail through collapse by elastic buckling. Elastic buckling of domes is a complex problem that lacks a simple theoretical solution. The following equation was developed from extensive model testing:

$$P_{CR} = \frac{0.8E}{(1 - \nu^2)^{1/2}}\left(\frac{R_o - R_i}{R_o}\right)^2 \ , \tag{3.15}$$

where

P_{CR} = critical collapse pressure
E = elastic modulus of the dome material
ν = Poisson's ratio for the dome material.

If the dome loading is uniform on the projected area (aerodynamic loading), the stresses in a thin dome are estimated by:

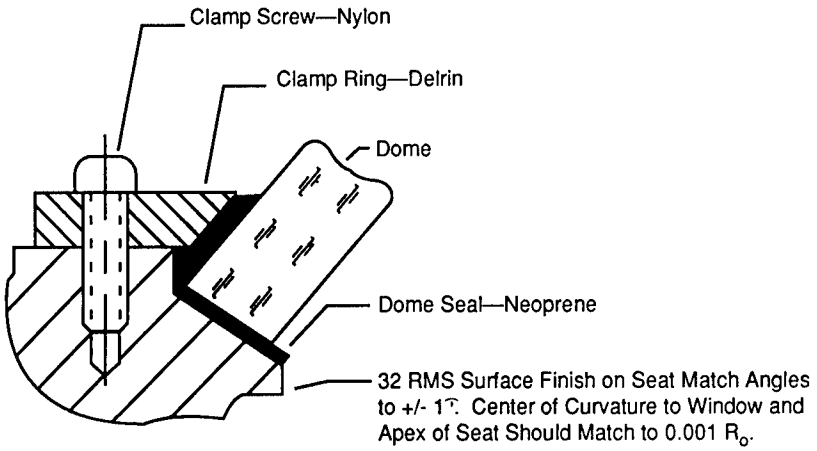


Fig. 3.2 Typical hermetically sealed dome mount.

$$\sigma_m = \frac{P}{2}\left(\frac{R}{h}\right) , \tag{3.16}$$

$$\sigma_h = \frac{P}{2}\left(\frac{R}{h}\right)\cos(2\phi) , \tag{3.17}$$

where P is the magnitude of the uniform loading on the projected area of the dome.

Domes are typically mounted in a conical seat. To avoid collapse by elastic buckling under the loading of external pressure, the apex of the conical seat and the center of curvature of the dome must be coincident to a tolerance of $0.001\,R_o$. For minimum stress in the area of the seat, the angle of the seat must match the angle of the dome edge to a tolerance of $\pm$ 1 arcmin. In addition, the dome seat should have a surface finish of 32 rms or better.[4] A surrounding external ring with a conical inner edge tangential to the outer surface of the dome is used to retain the dome in its seat. If a gasket is required to seal the dome, the gasket is placed in the conical surface of the seat. Use of a compliant bezel material, such as nylon or delrin, reduces stress in the dome and allows relative motion of the dome with respect to the seat during temperature changes. A typical dome mounting is shown in Fig. 3.2.

A hemispheric dome may be mounted against a flat seat with a surrounding bezel. Typically, the area between the bezel and the dome is filled with an adhesive to restrain the dome. A zinc sulfide dome 2.4 mm thick, with an external radius of about 50 mm, mounted in this fashion is used to protect an IR seeker in the nose of a homing 107-mm mortar shell. During firing,[5] this dome sees an acceleration of $11 \times 10^3\,g$, and survives a stress of 26 MPa.

3.2.4 Window Strength

Most optical window materials are brittle, and fail by fracture. Failure occurs as a result of the enlargement of small surface flaws under tensile stress; these

flaws become cracks that propagate across the surface until catastrophic fracture occurs. No significant plastic range of deformation occurs before failure in window materials. Failure is entirely due to surface fracture. Window strength is therefore dependent on surface condition and surface area. The larger the flaw in the window surface, the lower the strength. A larger window will have a greater chance of having a larger flaw, and is usually weaker than a smaller area window.

Nondestructive testing techniques normally cannot be used to locate the largest flaw in the surface of the window. Window strength is determined by scaling from tests of small samples. Because uncertainty exists concerning the actual window surface condition, the strength of the window is given in terms of probability of failure.

Two probability distributions are used to determine the strength of a window: the Gaussian (normal) and Weibull. For the range of strength of interest in optical windows, both distributions are a good representation of window strength.[6] The probability of failure of a window at an applied stress is given by the Gaussian distribution:

$$P_F = 1 - \int_{-\infty}^{\sigma_a} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\sigma_a - \bar{\sigma}}{\sigma_d}\right)^2 \right] d\sigma \ , \tag{3.18}$$

where

P_F = probability of failure
σ_a = applied stress
$\bar{\sigma}$ = mean stress of the distribution
σ_d = standard deviation of the distribution.

The probability of failure of a window at an applied stress is given by the Weibull distribution:

$$P_F = 1 - \exp\left[-\left(\frac{\sigma_a}{\sigma_0}\right)^m \right] \ , \tag{3.19}$$

where m is the Weibull modulus and σ_0 is the scale factor.

Both of these failure probability equations must be scaled to the actual surface area and the surface condition of the window. If the stress for failure is known for a certain surface area, then the stress required for failure for another surface area is found by[7]:

$$\frac{\sigma_1}{\sigma_2} = \left(\frac{S_2}{S_1}\right)^{1/m} \ , \tag{3.20}$$

where S_1, S_2 are the respective surface areas and σ_1, σ_2 are the respective stresses. If the stress required for failure is known for a certain flaw depth, then the stress required for failure at another flaw depth is given by[8]:

$$\frac{\sigma_1}{\sigma_2} = \left(\frac{a_2}{a_1}\right)^{1/2} \ , \tag{3.21}$$

where a_1, a_2 are the respective flaw depths.

Tables 3.1 and 3.2 provide representative Gaussian and Weibull statistics for common window materials. Unfortunately, scaling data on surface area and flaw size are not common in a technical discussion of window strength.

Table 3.1 Gaussian Strength Properties of Glass

Material Type	$\bar{\sigma}_1$ (MPa)	$\bar{\sigma}_d$ (MPa)	$\bar{\sigma}_N$ (MPa)	$t_{N0.5}$ (s)	s (mm^2)	a (μm)
1. Fused silica (Amersil)		5.52	103	43.7×10^3	471	5
2. Borosilicate (GE FN 7052)			100	32.3	471	6
3a. Borosilicate (Pyrex) (Corning 7740)		61.4	245	251	660	Unabraded
3b. Borosilicate (Pyrex) (Corning 7740)	236.7	56.6			660	2
4. Silicon	137.0	32.4				2

Table references by material type:

1, 2 J. E. Burke, R. H. Doremus, W. B. Hillig, and A. M. Turkalo, "Static fatigue in glasses and alumina," in *Ceramics in Severe Environments*, W. W. Krigel and H. Palmour, Eds., Plenum Press, New York (1971).

3a, 3b R. H. Doremus, "Fracture and fatigue of glass," *Treatise on Materials Science and Technology*, Vol. 22, Academic Press, New York (1982).

4 F. M. Anthony and A. K. Hopkins, "Actively cooled silicon mirrors," *Proceedings of the SPIE* **297**, 196–203 (1981).

Table 3.2 Weibull Strength Properties of Glasses

Material Type	m	σ_0 (MPa)	s (mm^2)	a (μm)	σ_F (MPa)	t_f (s)	P_F	n
1. Schott Zerodur	14	11.3	50	10	90			
2a. ULE (Corning 7971)	4.5	40.4		Polished	15.8	60	10^{-3}	29.2
2b. ULE (Corning 7971)	6.8	11.6		Unpolished	11.7	60	10^{-3}	15.1
3. Silicon (single crystal)	4.54	346.5						
4. AL_2O_3 (single crystal)	5–15	30–50						
5. CaF_2	3	5						
6. SrF_2	3	4						
7. ZnSe	6–9	23–27						
8. ZnS	4–9	15–29						
9. Soda lime (Kimble R-6)	6.18	128.8	660	2				16.0
10. Pyrex (Corning 7740)	4.99	257.2	660	2				27.4

Table references by material type:

1 *Zerodur Glass Ceramics*, Schott Optical Glass, Duryea, PA (1982).

2a, 2b D. C. Cranmer, "Mechanical properties of fused silica," The National Bureau of Standards (now The National Institute of Standards and Technology), U.S. Department of Commerce (1987).

3 C. P. Chen and M. H. Leipold, "Stress rate and proof testing of silicon wafers," *Journal of the American Ceramic Society* **68**, C54–C55 (1985).

4–8 N. C. Fernelius, G. A. Graves, and W. L. Knecht, "Characterization of candidate laser window materials," *Proceedings of the SPIE* **297**, 188–195 (1981).

9, 10 R. H. Doremus, "Fracture statistics: a comparison of the normal, Weibull, and type 1 extreme value distributions," *Journal of Applied Physics* **54**, 193–198 (1983).

A window that is safe, as determined by a probability of failure calculation as shown previously, may still fail under long-term loading due to static fatigue. In the presence of a corrosive environment, and under constant tensile loading, surface flaws may propagate slowly into catastrophic cracks. If the size of the largest or critical surface flaw is known, and the tensile stress applied to this flaw is also known, the time to failure can be estimated. Unfortunately, nondestructive testing for surface flaws is usually not practical for windows. Instead, windows are proof tested by subjecting the window to a higher stress than the design stress. If the window does not fail during the proof test, the test establishes an upper limit for the critical flaw, which can in turn be used to estimate the time to failure.

Velocity of crack propagation is related to the stress intensity in the window. If the crack velocity is related to stress intensity by a power law,[9]

$$V = AK_I^n \ , \tag{3.22}$$

then the time to failure after proof testing is given by[10]:

$$t_F \cong \frac{2K_{IC}^{2-n}}{\sigma_a^2 A Y^2 (n-2)} \left(\frac{\sigma_p}{\sigma_a}\right)^{2-n} \ , \tag{3.23}$$

where

V = crack velocity
A = a material parameter
n = a material parameter
K_I = stress intensity
t_F = time to failure after proof testing
K_{IC} = critical stress intensity for the material
σ_a = service tensile stress
σ_p = proof tensile stress
Y = a constant, normally set equal to $\sqrt{\pi}$.

If the crack velocity is related to the stress intensity by an exponential law,[11]

$$V = \exp -(a/b) \exp(K_I/b) \ , \tag{3.24}$$

$$t_F \simeq \frac{2b^2 \exp(a/b)}{\sigma_a Y^2} \left[\frac{K_{IC}}{b}\left(\frac{\sigma_p}{\sigma_a}\right) + 1\right] \exp\left[-\frac{K_{IC}}{b}\left(\frac{\sigma_p}{\sigma_a}\right)\right] \ , \tag{3.25}$$

where a is a material parameter and b is a material parameter.

Table 3.3 provides static fatigue material parameters for a variety of window materials. Unlike strength calculations based on probability distributions, static fatigue material parameters are independent of surface area and surface condition.

Vibration of a window can induce static fatigue. The service lifetime of a window exposed to cyclic loading is related to the service lifetime under static loading by[12]:

Table 3.3 Fracture Mechanics Properties of Glasses

	$K_I = a + b\,InV$			$K_I = AK_I^n$	
Material Type	K_{IC} (MN/m$^{3/2}$)	a (kN/m$^{3/2}$)	b (kN/m$^{3/2}$)	A (m/s)	n
1. Fused silica	0.741	693.1	13.42		36.1
2. 96% silica (Corning 7900)	0.700	649.0	13.88		
3. Vycor (Corning 7913)	0.715	650.7	12.98		49–52
4. Aluminosilicate (Corning 1723)	0.846	795.0	17.97		46
5. Borosilicate (Pyrex)	0.760	643.1	14.14		16.0
6. Dense flint (Schott SF-1)	0.624	571.7	15.57		38
7. Borosilicate crown (Schott BK-7)	0.862	794.6 V > 10^{-6} m/s 582.6 V < 10^{-7} m/s	25.3 11.77		
8. Borosilicate crown (Schott UBK-7)	0.886	787.2 V > 10^{-6} m/s 597.6 V < 10^{-7} m/s	26.74 13.48		
9. Soda lime	0.750	831.0	24.3	0.0313	11.65
10. Sapphire	0.300	1305	26.9		
11. ZnSe (single crystal)	0.33			1.8×10^{18}	40
12. ZnSe (polycrystalline)	0.90			2.0	40
13. ULE (Corning 7971)		605	11.6		

Table references by material type:

1–8 S. M. Wiederhorn, A. G. Evans, and D. E. Roberts, "A fracture mechanics study of the Skylab windows," in *Fracture Mechanics of Ceramics*, R. C. Bradt, D. P. H. Hasselman, and F. F. Lange, Eds., Plenum Press, New York (1974).

9, 10, 13 W. P. Barnes, Jr., *Mechanical Engineering Aspects of Optical Systems,* SPIE Engineering Update Series Course U13 Notes, SPIE, Bellingham, WA (1986).

11,12 S. W. Freiman et al., "Influence of microstructures on crack propagation in ZnSe," *Journal of the American Ceramic Society* **58,** 406–409 (1975).

$$t_c = J^{-1}\left(\frac{\sigma_s}{\sigma_c}\right)^n t_s \ , \tag{3.26}$$

where

t_c = cyclic time to failure
t_s = static time to failure
σ_s = failure stress under static loading
σ_t = cyclic mean stress
n = material parameter (from Table 3.3)
J^{-1} = amplitude factor.

For sinusoidal loading of the window,

$$\sigma(t) = \sigma_c + \sigma_0 \sin(\omega t) \ . \tag{3.27}$$

The amplitude factor is given by:

$$J = \sum_{i=0}^{(n/2)_I} \left\{ \left[\frac{n!}{(n-2i)!(i!)^2} \right] \left(\frac{\sigma_0}{2\sigma_c} \right)^{2i} \right\} \ , \tag{3.28}$$

where

σ_0 = cyclic stress amplitude
ω = cyclic period
J = amplitude factor
$(n/2)_I$ = integer part of $n/2$.

If the cyclic mean stress is zero, the amplitude factor is given by:

$$J = \frac{1}{(4\pi)^{1/2}} \frac{\Gamma[(n + 1)/2]}{\Gamma(n/2 + 1)} . \tag{3.29}$$

A useful approximation for the amplitude factor when the cyclic mean stress is zero and the material parameter is greater than nine (as it normally is for common window materials) is:

$$J^{-1} = (2\pi n)^{1/2}\left(1 + \frac{1}{4n} + \frac{1}{32n^2}\right) , \tag{3.30}$$

where Γ denotes the Gamma function.

Note that the service life of a window under cyclic stress is longer than under pure static stress. The increased lifetime results from the window being stressed in tension during only part of the cycle.

3.2.5 Window Fundamental Frequency

Application of the cyclic fatigue equations presented in Sec. 3.2.4 requires knowledge of the fundamental frequency of the window. The fundamental frequency of a simply supported circular window is given by[13]:

$$f_n = \frac{\pi}{4} \frac{1}{r^2} \left[\frac{gEh^2}{12\rho(1 - \nu^2)}\right]^{1/2} , \tag{3.31}$$

where

f_n = natural frequency of the window
g = gravitational acceleration at the earth's surface
E = elastic modulus of the window material
h = window thickness
ν = Poisson's ratio for the window material
ρ = window material density.

The fundamental frequency of a simply supported rectangular window is given by:

$$f_n = \frac{\pi}{2} \left[\frac{gEh^2}{12\rho(1 - \nu^2)}\right]^{1/2} \left(\frac{1}{b^2} + \frac{1}{L^2}\right) , \tag{3.32}$$

where b is the window width and L is the window length.

3.2.6 Window Mounting

Because windows are zero-power optical elements, window position usually does not influence the optical performance of the system. Tilt and translation are acceptable when mounting the window. It is not uncommon to introduce a deliberate tilt into a window with respect to the optical axis to control ghost reflection. However, window deformation introduces optical aberration. Window mounts must therefore be designed to minimize window deformation.

Use of a compliant adhesive material between the window edge and mount reduces mounting deformation. The adhesive bond is limited to the edge of the window and restrains the window both axially and radially. Hard contact occurs between the window and mount. (The thin mounting ring surrounding a window is sometimes called a *bezel.*) A ring of flexible adhesive between the window and mount reduces thermal-induced stress in the window due to thermal coefficient of expansion mismatch between the window and mount. Any irregularity in the surface of the mount in contact with the window induces a corresponding deformation in the window. The optical effect on the window due to mount-induced deformations is difficult to calculate. A rule of thumb is to require that the surface to which the window is mounted have no irregularity greater than the elastic deformation of the window or mount under the window clamping stress. If no retainer is used to hold the window in place, the clamping force is set equal to the window weight in calculating the surface deformation. The permissible tolerance for the mounting surface irregularity is given by:

$$P = W + F \ , \tag{3.33}$$

$$\delta_t = \frac{ph}{SE} \ , \tag{3.34}$$

where

δ_t = irregularity tolerance for the window mount
W = window weight
h = window thickness
S = area of contact between mount and window
E = window material elastic modulus
F = window clamping force.

The required edge thickness of adhesive around the window is determined using the equations given in Sec. 3.4.2.

Mechanical contact between the window and mount is another scheme for mounting windows. Contact should not be made between the window edge and mount unless the window edge is tapered. Contact of the mount against a window edge that is normal to the plane of the window surface can create large hoop stresses. These large hoop stresses arise from a thermal coefficient of expansion mismatch between window edge and window mount. A temperature drop causes the window mount to contract around the window creating these hoop stresses. Large temperature-induced stresses in the window are reduced by tapering the edge of the window and utilizing a matching taper in the window mount. Use of a tapered edge permits the window to move relative

to the mount as the mount expands or contracts. Reduction of contact stresses in the window mount requires that the taper angles of the window and mount be matched to 1 arcmin. In addition, a high-quality surface is necessary on the window mount, typically an 8 to 32 finish. Because the window moves relative to the mount, a flexible retainer is required. Flexibility is provided by milling a groove in a conventional retainer. The area under the groove acts as a flexure and permits the retainer to move up and down to accommodate the motion of the window.

Isolation of the window from expansion and contraction of the mount is provided by O-ring seals between the window and mount. O-ring seals are located around the circumference on the upper edge of the window, and on the lower edge of the window, directly opposed to the upper edge. Each is compressed to about 50 to 70% of the nominal full compression. A space is created between the window and mount due to this lack of compression. The remaining flexibility of the O-rings created by this partial compression is used to take up the change in clearance between window and mount as the temperature changes. O-ring seals of up to 70 durometers are used for this application.

Figure 3.3 shows details of typical window mounts.

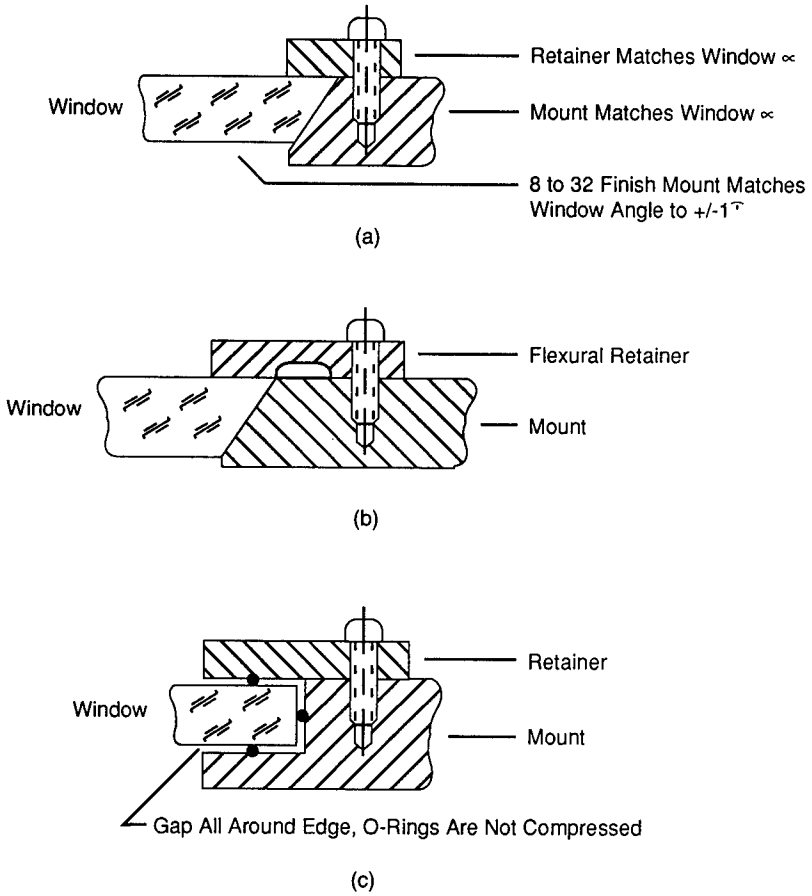


Fig. 3.3 Typical window mounts: (a) hard mount, (b) athermal hard mount, and (c) O-ring mount.

3.3 PRISMS

This section emphasizes the principles of mounting prisms rather than prism design.

3.3.1 General Principles of Prism Mounting

The fundamental rules for mounting prisms are the same as for any other precision optical component. The prism must be held in a well-defined position without deformation. Unlike windows, prisms are very stiff and difficult to deform. An additional difference between windows and prisms is the requirement for the precise location of the prism with respect to angle. Since prisms are zero-power optical elements, precise location of the prism with respect to translation is rarely required.

In designing a prism mount, deformation of the optically active prism faces must be avoided. Deformation can be greatly reduced if mounting surfaces are selected on the sides of the prism that are at right angles to the optically active prism surfaces. Stray light can enter a prism through surfaces that are used in total internal reflection. Control of stray light may require an additional opaque shield around the prism to limit transmission of light into the prism. In some cases, such as Porro prism assemblies, a shallow cut across the hypotenuse of the prism is a useful means of suppressing stray light at very oblique angles. Overconstraint of the prism should be avoided by the use of kinematic mounting principles.

3.3.2 Prism Bonding

Prisms, with their high stiffness and large flat areas, are well suited for direct bonding to metal mounts. The classic method involves the use of a large circular bond against a side of the prism at right angles to the optically active surface. Yoder[14] has developed an equation that is used to size the area of the metal to glass bond:

$$S = \frac{aWSF}{\sigma} , \tag{3.35}$$

where

S = bond area
a = acceleration of the prism (in dimensionless g)
W = prism weight
SF = a safety factor (Yoder[14] suggests a safety factor of 2)
σ = shear strength of the adhesive.

Control of the adhesive thickness and prism angle with respect to the mount is achieved by placing three symmetrically located shims or spacers between prism and mount during the bonding operation. Plastic or metal shims and thin wires are all possible choices for the spacers. Plastic shims allow for better expansion or contraction of the prism with respect to the mount, while metal shims provide better angular stability. A semiflexible adhesive such as 3M EC-2216™ permits expansion or contraction of the mount with respect to the

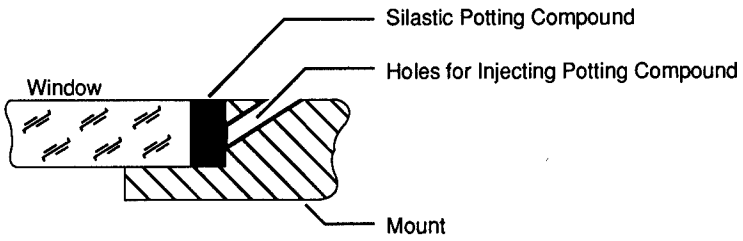


Fig. 3.4 Bonded window mount.

prism. Large bond areas where the distance between bond edge and center exceeds 10 mm may require several weeks to cure completely. The curing rate is improved by breaking the large single bond into a series of smaller bonds of equivalent total area. Figure 3.4 shows details of a typical bonded window mount.

3.3.3 Kinematic Mounting of Prisms

Kinematic mounting of prisms requires the selection of no more than six points to locate the prism uniquely without overconstraint.[15] To accommodate changes in temperature, it is desirable to use less than the full six kinematic locating points, so that the prism is free to expand or contract with respect to the mount. Kinematic points should be selected to maintain the angular position of the prism. Locating angular defining kinematic points may require violation of the basic principles of prism mounting since in many cases the ideal location of these points is in contact with an optically active surface. If direct contact between mounting points and an optically active surface is required, the flatness and coplanarity of the mounting points are given by Eqs. (3.33) and (3.34).

Retention of the prisms in a kinematic mount is provided by spring preloads.[16] These spring preloads should directly oppose the locating points. If the irregular shape of the prism prevents direct opposition of the locating points by the retaining springs, two alternate retainer designs are used. A central retainer, equidistant from the locating points, provides an even distribution of support. The other approach uses a broad area of contact between prism and retainer. A retainer contoured to fit the irregular surfaces of the prism is used; contact between retainer and prism is made through an intermediate elastomer pad. This last design, with its large areas of contact, is good for environments subjected to high accelerations or mechanical shock. Figure 3.5 shows a kinematic prism mount.

Flexure mounting of the prism accommodates large changes in temperature. Flexure mounting is similar to kinematic mounting in that mounting points are chosen using kinematic principles. Each mounting point is connected to a flexure. Each flexure has the necessary degrees of compliance and restraint to locate the prism while providing for expansion or contraction of the mount with respect to the prism.[17]

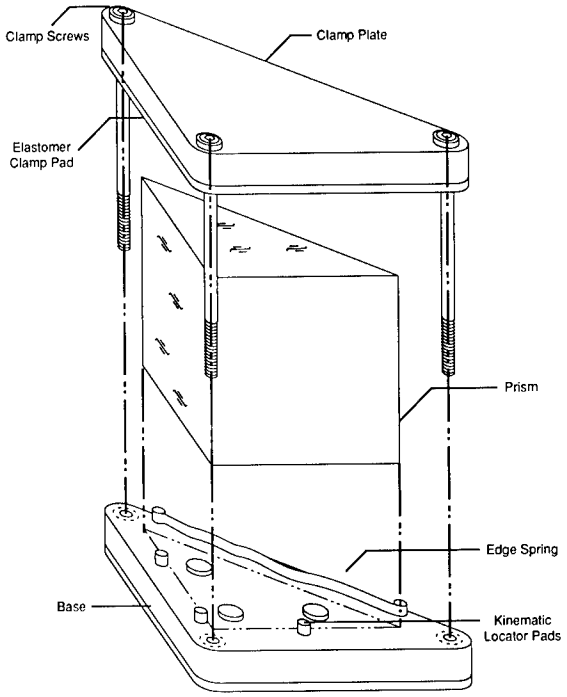


Fig. 3.5 Kinematic prism mount.

Flexures are very similar to springs. Flexures provide controlled motion through elastic deformation, while springs provide controlled force through elastic deformation. For optical mounting purposes, flexures are normally made in the form of cantilever springs, providing either translation or rotation. For example, a classic flexure mount for a prism consists of three flexures bonded to one face of the prism. Each flexure is compliant in a plane parallel to the prism face, and is stiff in the direction normal to the prism face. The axis of compliance of each of the three flexures is directed toward the center of gravity of the prism in the plane of the flexures. This flexure arrangement allows the mount to expand or contract relative to the prism without introducing any rotation of the prism. Figure 3.6 shows a flexure mount for a prism.

3.3.4 Large Contact Area Mounts for Prisms

High accelerations may create local stresses in bonded or kinematic mounted prisms that are high enough to induce failure in the prism material. Mounting stresses are reduced if the contact area between prism and mount is increased. In the limiting case, the mounting area is the same size as the prism face. For this limiting case, the maximum acceleration tolerable before prism failure is given by:

$$a = \frac{\sigma_F}{\rho h}\,, \tag{3.36}$$

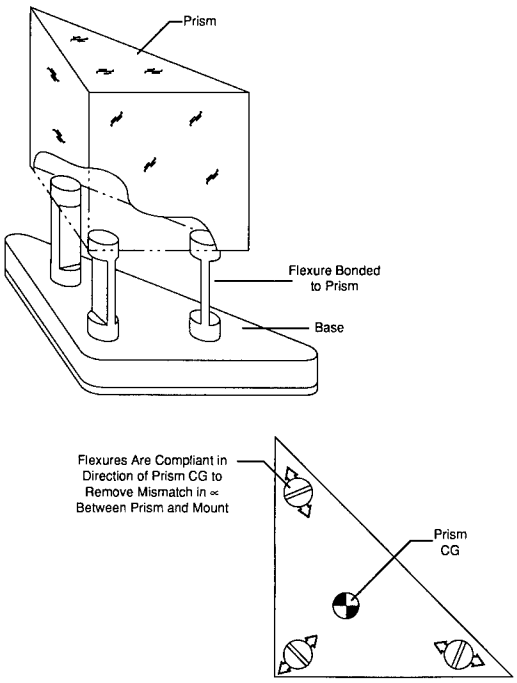


Fig. 3.6 Prism flexure mount.

where

h = prism thickness
σ_F = prism material yield strength
a = acceleration (in dimensionless g)
ρ = prism material density.

For a typical prism material such as BAK-4 glass, with a density of 3.1 kg m^{-3}, and a tensile yield strength of approximately 7 MPa, a prism thickness of 50 mm tolerates about 446 g.

Although the large area of contact increases the chances of survival of the prism under high acceleration conditions, this advantage is offset by some very serious disadvantages. The mounting surface for the prism must be flat and have the same tolerance as the optical surface of the prism. Large optically flat mounting surfaces are expensive to produce. A change in temperature can induce large stresses in the interface between prism and mount. Finally, the great strength of the large area of contact is applicable only to accelerations perpendicular to the mounting surface. An acceleration parallel to the mounting surface induces large shear stresses.

Normally, a large contact area mount is designed using semikinematic principles. Clamping forces are applied to the prism using spring-loaded plates that directly oppose the mounting surface; this method places the glass in compression. An elastomer pad can be used to replace the spring preload in the clamp. Multiple contacting surfaces are used, with support provided for

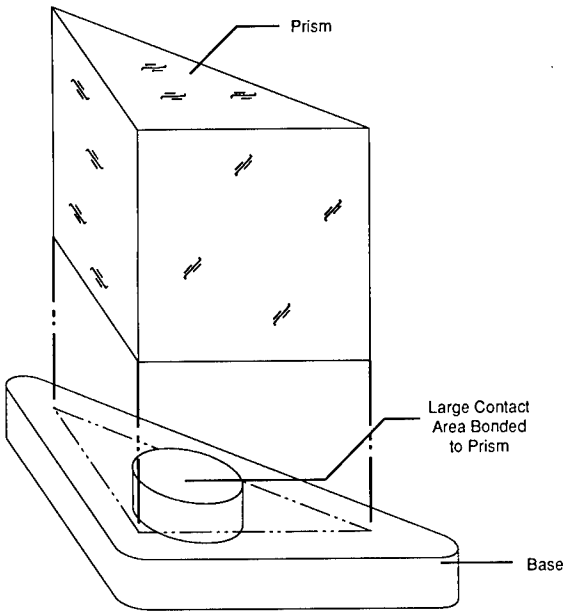


Fig. 3.7 Large contact area prism mount.

all prism surfaces except the input and output faces of the prism. Figure 3.7 shows a large contact area mount for a prism.

3.4 LENS MOUNTING

Distortion of the lens surfaces can lead to performance degradation, therefore requiring a low-stress mounting design. At the same time, the lens mount must maintain the position of the lens to much tighter tolerances than are customary for normal mechanical systems.

3.4.1 General Considerations in Lens Mounting

Although tolerances of less than 100 nm are routinely achieved in the fabrication of optical surfaces, mechanical tolerances on the mounting surfaces of lenses are literally several orders of magnitudes larger than optical tolerances. Assuming normal optical fabrication shop practice, the smallest mechanical tolerance that is readily and economically produced on a lens is about 25 μm. Smaller tolerances, particularly with respect to centering, require very large increases in production time and cost.[18]

Lens Centering. Centering of the lens is the most important of all mechanical tolerances. Centering is defined as the error in location between the optical axis of the lens and the mechanical axis of the lens. Traditionally, lens centering is defined with respect to an optical measurement. The lens is placed on a hollow centering bell and illuminated with parallel light. The focus of the lens is found and the lens and bell are then rotated. As the bell rotates, an error in centering of the lens causes the focus to describe a circle. The

diameter of the circular motion of the focus divided by the focal length of the lens is an angular measure of the centering of the lens and is given in arc minutes.[19] This technique is limited to lenses transmitting in the visible and with focal lengths and centering errors that are easy to measure. Very short or very long focus lenses, or lens materials that do not transmit in the visible, are not suited for use with this method. A wide variety of centering techniques exist, and are discussed in the literature on lens fabrication.[20] More and more often, designers are specifying the required centering and leaving the test method up to the fabrication shop.

One such method of specifying centering is to specify the locations of the centers of curvature of the two lens surfaces with respect to the mechanical axis of the lens. Using standard mechanical tolerance practice, the outer cylindrical surface of the lens is defined as a datum surface. The locations of the centers of curvature of the lens surfaces are then determined with respect to surfaces perpendicular to the datum surface. Normally, two perpendicular surfaces are used and each perpendicular surface is coincident with the lens surface at the mechanical axis of the lens. (Note that the perpendicular surfaces are not tangent to the vertices of the lens surfaces, since these vertices may lie off the mechanical axis of the lens.) Two lines are defined with one line passing through the lens surface center of curvature and the center of the perpendicular surface. The other line is the mechanical axis of the lens. The angle between these two lines defines the centering of the surface.[21]

Center of Gravity and Weight Estimates for Lenses. Lenses consist of three basic shapes: the right circular cylinder, the truncated cone (or beveled disk), and the spherical segment. These three basic shapes are combined to find the weight and center of gravity of the lens.[22]

Weight and center of gravity of the disk or right circular cylinder are given by:

$$W_D = \frac{\pi}{4}\rho h D^2 , \tag{3.37}$$

$$\mathrm{CG}_D = \frac{h}{2} . \tag{3.38}$$

Weight and center of gravity of the truncated cone or beveled disk are given by:

$$W_c = \frac{\pi}{12}\rho h(D - Dd + d^2) , \tag{3.39}$$

$$\mathrm{CG}_c = \frac{h}{3}\frac{D/2 + d}{D/2 + d/2} . \tag{3.40}$$

Weight and center of gravity of the spherical segment are given by:

$$h_s = R - \left(R^2 - \frac{D^2}{4}\right)^{1/2} , \tag{3.41}$$

$$W_s = \pi\rho h_s^2\left(R - \frac{h_s}{3}\right) , \tag{3.42}$$

$$CG_x = \frac{h_s}{4}\frac{(4R - h_s)}{(3R - h_s)} . \tag{3.43}$$

In the preceding equations

W = lens weight
ρ = lens material density
h = thickness
h_s = sagittal thickness of a spherical segment, measured from flat side to vertex
D = diameter
d = smaller diameter (for a beveled disk)
CG_D = center of gravity location of a disk
CG_c = center of gravity location of a beveled disk, measured from the smaller diameter
CG_s = center of gravity location of a spherical segment, measured from the flat side
R = radius of curvature of a spherical segment.

Once the weights and center of gravity locations of the individual basic shapes of the lens are known, the overall weight and center of gravity location is found from:

$$W_L = \sum_{i=1}^{n} W_i , \tag{3.44}$$

$$CG_L = \frac{\sum_{i=1}^{n} CG_i W_i}{W_L} , \tag{3.45}$$

where

W_L = overall lens weight
W_i = weight of a basic shape (negative for concave spherical shapes)
CG_L = center of gravity of the lens, with respect to the datum (normally taken as one of the vertices of the lens)
CG_i = center of gravity of the individual basic shapes, taken with respect to the same datum.

In many cases, an approximate method is more suitable for finding the weight and center of gravity. Most lenses are combinations of spherical and cylindrical shapes; the beveled disk is not commonly used. If the curvature of the spherical surface is such that the sagittal thickness is less than half the radius of curvature, an approximate method for finding the weight and center of gravity of a spherical segment is used:

$$W_s \cong \pi\rho h_s \frac{D^2}{8} = \pi\rho h_s^2 R , \tag{3.46}$$

$$CG_s \cong \frac{h_s}{3} , \tag{3.47}$$

where W_s is the weight of the spherical segment.

3.4.2 Bonded Lens Mounts

Bonding the lens into its mount is a simple and economical way to mount lenses. Two different bonding techniques are used: a thin, high-strength direct bond, and a thick, semiflexible bond.

Direct Bonding of Lens Mounts. In direct bonding, a thin, high-strength bond is produced between lens and mount. To produce maximum accuracy, mechanical contact should exist between the optical surfaces of the lens and the mount. Adhesive should be limited to the circumference of the lens, or to the area outside the circle of contact between lens surface and mechanical seat in the lens cell.

Mechanical clearance considerations may limit the adhesive thickness to below that desired for optimum properties. If this is the case, a groove can be machined into the cell to accommodate the adhesive. Injection holes are used to place the adhesive into the groove. For a typical lens diameter of 25 to 50 mm, and for normal optical adhesives, six equally spaced radial injection holes, 1 to 1.5 mm in diameter are sufficient. Spread of the adhesive is checked by looking through the lens. If the lens material is not transparent in the visible, adhesive bond progress is monitored by placing additional "witness" holes between the injection holes. Adhesive is injected until it begins to flow out the witness holes.

The enormous number of possible adhesives, and the rapid change in availability of commercial adhesives, makes it impossible to tabulate the properties of adhesives here. Several important properties should be considered in selecting an adhesive for optical applications[23]:

1. Most adhesives shrink during curing. This may cause a shift in lens position after the adhesive sets. One solution is to lock the lens in position mechanically, and then inject the adhesive around the lens. Another solution is to adjust the position of the lens continuously in smaller and smaller increments as the adhesive sets.
2. Shrinkage of the adhesive during curing produces residual stress in the lens. Some of this stress is reduced by increasing the cure time of the adhesive.
3. Ultraviolet setting adhesives are very convenient for lens bonding. These adhesives remain relatively fluid until illuminated by a strong uv light source. These uv curing adhesives shrink during curing. Residual stress in uv curing adhesives is reduced by lengthening curing time.
4. Many adhesives evolve a gas during curing and may outgas in a vacuum. The cyanoacrylic adhesives are particularly bad in this respect and should not be used in a vacuum environment.
5. There are solvents available that dissolve any adhesive. These solvents are often toxic and slow to take effect. Heat can also cause some adhesives to lose strength, allowing removal of the lens from the mount.
6. Hard, high-strength bonds are not flexible. A change in temperature can induce very large stresses between lens and mount.

Semiflexible Bonding of Lenses into Mounts. Semiflexible adhesives are used in relatively thick bonds to isolate the lens from temperature effects, and in some cases, against mechanical shock and vibration. The adhesive forms a thick ring between the outer diameter of the lens and the inner diameter of the lens cell. If properly designed, this thick ring of adhesive maintains the correct centering location of the lens as the temperature changes. At the same time, the thick adhesive layer maintains the lens in a low stress condition. Axial location of the lens is normally obtained via mechanical contact between the optical surface of the lens and the seat of the cell.[24]

The optimum thickness of adhesive between lens and cell to produce a zero stress condition in the lens as the temperature changes is given by[25]:

$$h_r = \frac{D_g(\alpha_m - \alpha_g)}{2(\alpha_r - \alpha_m)} , \tag{3.48}$$

where

h_r = required adhesive radial thickness
D_g = lens outer diameter
α_g = lens thermal coefficient of expansion
α_m = lens cell thermal coefficient of expansion
α_r = adhesive thermal coefficient of expansion (for the silastic adhesives often used in this application, this is about 200×10^{-6} m/m K).

The thickness of adhesive given by Eq. (3.48) is relatively large in comparison with normal adhesive bonding practice. Adhesive thicknesses of several millimeters are not uncommon. The large adhesive thickness used in this method requires that the lens be mechanically centered with respect to the mechanical axis of the lens cell before adding the adhesive. This mechanical centering is achieved through the use of radial screws or shims. In either case, a release agent is placed on the mechanical adjustment to allow removal after curing of the adhesive.

Under the weight of the lens the very thick layer of semiflexible adhesive deflects causing the lens to move with respect to the cell. The radial deflection of the center of a lens mounted in a semiflexible adhesive mount is given by[26]:

$$\delta_L = \frac{W_L}{(\pi/2)d_G(h_G/h_r)(E_r/(1 - \nu_r^2) + G_r]} , \tag{3.49}$$

where

δ_L = decentering of the lens with respect to the mount
W_L = lens weight
d_G = lens diameter
h_G = length of adhesive in contact with the side of the lens
E_r = elastic modulus of the adhesive
ν_r = Poisson's ratio of the adhesive
G_r = shear modulus of the adhesive.

Figure 3.8 shows a semiflexible adhesive bonded lens.

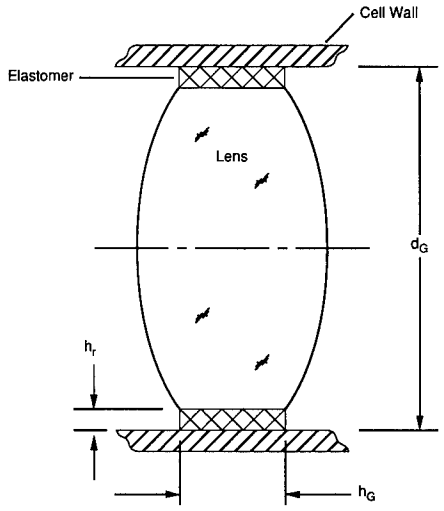


Fig. 3.8 Semiflexible adhesive bonded lens.

3.4.3 Shimmed Mounts

Assembly of the lens into its cell is made relatively easy by large radial clearances between the outer diameter of the lens and the inner diameter of the cell. Such large clearances limit assembly precision. Assembly precision can be reestablished by placing a shim between the lens and bore of the cell.

A metal shim is easy to place between the cell and lens. Brass or stainless steel shims are the most common. A full circle of contact is common between cell, lens, and shim. Use of a semikinematic three-point contact set of shims requires the use of adhesives to lock the shims in place. Metal shims are easy to install, are stable, and provide good accuracy. Unfortunately, metal shims do not provide any isolation from expansion or contraction of the lens cell with respect to the lens. Metal shims are therefore restricted to relatively limited changes in temperature.

Larger temperature changes are accommodated through the use of semiflexible shims. A mylar shim offers stability and accuracy of centering while at the same time providing sufficient elasticity to isolate the lens from changes in cell size due to temperature fluctuations. Mylar shims are installed under slight compression. Sufficient compression is provided so that during expansion of the cell the shim remains in compression. Should the shim compression be removed due to a change in temperature, the shim would become loose and the lens could move with respect to the cell.

Like the metal shim, mylar shims normally are in full contact around the circumference of the lens. To simplify installation, the mylar shim is cut into three equal-length (120-deg) segments. Normal mylar shim thickness is 80 to 130 μm. It is suggested that the actual as-fabricated diameters of lens and cell bore be measured and the shim thickness selected to match.

Under mechanical shock and vibration, a semiflexible mylar shim may shift position around the edge of the lens causing the lens to move. Motion of the shim is eliminated by locking shim, lens, and cell together. This is accomplished by adhesive bonding. A radial hole is punched in the center of each shim

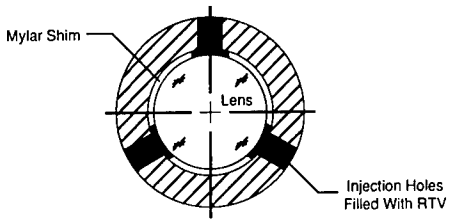


Fig. 3.9 Shimmed lens mount.

segment. The holes in the shim segments are aligned with radial adhesive injection holes in the cell. Injecting adhesive locks shim, cell, and lens together. Because the adhesive is used as a lock and is encapsulated, stability and outgassing of the adhesive are reduced concerns.[27]

Figure 3.9 shows a shimmed lens mount.

3.4.4 Seat and Retainer Mounts

Use of a mechanical seat and retainer is the classic means of lens mounting. This method offers good accuracy and stability at the expense of high stresses in the metal-to-glass interface.

Accuracy of Assembly. Accuracy of assembly using the seat and retainer lens mount is limited by the clearance required for free assembly between the lens outer diameter and cell bore. A minimum radial clearance of 25 μm is required for free assembly. Assuming a cell diameter tolerance of ± 25 μm, and a cell bore diameter tolerance of ± 25 μm, the maximum centering tolerance achieved using contact between lens and cell bore is 75 μm. Radial clearance is reduced by heating of the cell during assembly to obtain sufficient clearance. Any reduction of radial assembly clearance increases the chances of the lens binding in the cell during assembly. If the lens does bind, the chances of freeing it without damage are small. Producing a full spherical radius on the edge of the lens, with the radius coincident with the lens optical axis, can reduce the possibility of the lens binding during assembly. Producing such a spherical radius is expensive and may compromise the centering accuracy of the lens.

If either or both of the lens optical surfaces have a strong spherical radius of curvature, the centering accuracy of the lens in the cell bore may be improved through self-centering. A lens with spherical surfaces in contact with two coaxial rings tends to center under the axial clamping force of the two coaxial rings. This self-centering is exploited during mass production of lenses in a centering bell. The accuracy achieved by self-centering is determined by the coefficient of friction between the lens and cell and by the contact angle between optical surfaces and cell. As a rough rule of thumb, the optical radius of curvature necessary for self-centering to become effective is about 0.5 m. For low friction coefficients and strong surface curvatures, a lens may self-center in a cell to an accuracy of 10 μm.[28]

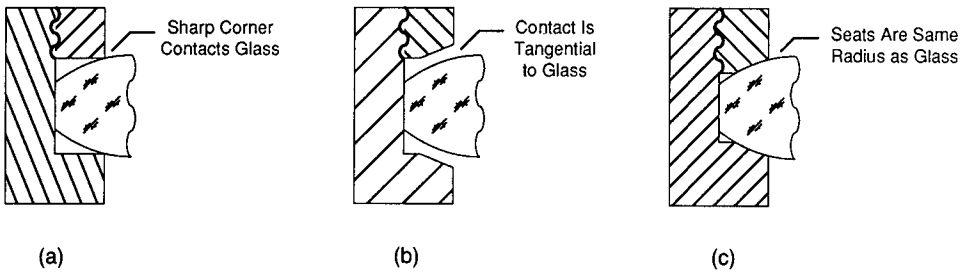


Fig. 3.10 Seat and retainer contact geometry: (a) sharp corner contact, (b) tangent contact, ad (c) spherical contact.

Seat and Retainer Geometry. The geometry of contact between a lens and cell determines the stress produced in the area of contact. Three types of contact are possible between the cell and lens:

1. *Sharp corner contact:* Contact between lens and metal is made at a sharp edge, usually with a 90-deg angle. This is an easy design to tolerance, but is a high-stress configuration. Production of a true sharp edge without burrs that reduce precision is difficult. This production difficulty is reduced at the cost of requiring more radial space outside the clear aperture by changing the included angle to 135 deg.
2. *Tangential contact:* Tangential contact between glass and metal is a lower stress design than the sharp corner, and is easier to produce. More radial clearance outside the clear aperture is required with this design in comparison with the sharp corner contact. Tangential contact is not possible with a concave surface, requiring the use of a separate mounting surface. Use of a separate mounting surface increases cost and lowers precision.
3. *Spherical contact:* A spherical seat with a radius of curvature matching the optical surface of the lens is the lowest contact stress design possible. Fabrication and tolerancing of the spherical contact are very difficult. More radial clearance outside the clear aperture of the lens is required with the spherical contact than with the sharp corner design. The use of spherical contact should be restricted to lens cells subjected to high accelerations.

Figure 3.10 shows the three types of contact geometries.

Retainer Design. Normally, the retainer is used to preload the lens and to ensure contact of the lens with the seat of the cell. The retainer should not be considered as a precision mechanical element. Retainer preload force must be large enough so that under any loading conditions the direction of the preload force does not change. For typical military aviation applications, a preload force of up to about 10 times the weight of the lens is required. A reduced preload is used for systems in a static laboratory environment. To allow for accelerations produced during shipping and handling, a minimum preload force of two to four times the weight of the lens is suggested.

For lens diameters up to 0.3 m, the retainer is provided with an outside thread matched to the thread of the cell. A loose fit between retainer and cell threads corresponding to a class 1 thread fit is very desirable. A loose thread

fit allows the assembly technician to determine when the lens first contacts the retainer and also allows the retainer to self-align against the lens to produce uniform contact around the circumference of the lens. The torque required to turn the retainer is related to the desired axial preload by[29]:

$$T \cong 0.2dF \ , \tag{3.50}$$

where F is the desired axial preload force, T is the retainer torque, and d is the retainer diameter. Very large retainer rings are provided with a bolt circle. Control of axial preload with a large number of bolts is extremely difficult.

Some axial compliance in the retainer may be required for lens cells exposed to large drops in temperature. The retainer may be faced with a silicone sponge sheet (normally about 12 durometers). Alternatively, an O-ring can be placed in the retainer to act as an elastic element. An O-ring of about 70 durometers is used and is compressed to 50 to 75% of nominal. Yet another approach is to incorporate a diaphragm flexure into the retainer. A diaphragm flexure is produced in a retainer by cutting a deep groove on the inside diameter of the retainer, almost to the outside of the retainer. The outer diameter of the groove produced should be greater than the mechanical diameter, or at least the contact diameter of the lens.

Mechanical Stress in the Lens and Cell. The maximum stress tolerable in a lens is estimated using the stress-optical coefficient, a lens material property. This property is also known as the stress birefringence of the lens material. The stress-optical coefficient determines the change in index of refraction of the material with applied stress. The optical path difference due to an applied stress on the lens is estimated by:

$$\text{OPD} = Kh\sigma \ , \tag{3.51}$$

where

OPD = optical path difference (or stress birefringence)
h = lens thickness
σ = applied stress
K = stress-optical coefficient.

If the retainer and seat do not contact the lens on a common, coaxial diameter, bending stress is induced in the lens due to the preload. The magnitude of this stress is estimated from[30]:

$$\sigma = \frac{3F}{2\pi(1/\nu)h^2}\left[\frac{1}{2}\left(\frac{1}{\nu} - 1\right) + \left(\frac{1}{\nu} + 1\right)\ln\left(\frac{d_1}{d_0}\right) - \left(\frac{1}{\nu} - 1\right)\frac{d_0^2}{2d_1^2}\right] \ , \tag{3.52}$$

where

σ = induced bending stress
F = preload force on the lens
d_0 = diameter of contact between seat and lens
d_1 = diameter of contact between retainer and lens
h = center thickness of the lens
ν = Poisson's ratio for the lens material.

Additional mechanical stress is produced in the area of contact between lens and seat or retainer. This contact stress is estimated using the Hertz contact stress theory. Contact stresses are highly localized and are usually more of a concern with respect to structural failure than the optical performance.

To use the Hertz contact stress theory to estimate contact stress due to a sharp corner in contact with the lens, it is necessary to assume that the sharp corner has some finite radius of curvature. For metallic retainers produced by conventional machining operations, the sharp edge can be assumed to have a radius of about 50 μm. The contact stress is then given by[31]:

$$\sigma = 0.798 \left\{ \frac{F/2\pi r[(d_G + d_m)/d_G d_m]}{(1 - \nu_G^2)/E_G + (1 - \nu_m^2)/E_m} \right\}^{1/2}, \tag{3.53}$$

where

σ = contact stress
F = retainer preload force on the lens
d_G = twice the lens optical radius of curvature
d_m = twice the radius of curvature of the sharp edge
ν_G = Poisson's ratio for the lens material
ν_m = Poisson's ratio for the cell material
E_G = the modulus of elasticity of the lens material
E_m = the modulus of elasticity of the cell material
r = the radial distance of the contact from the lens axis.

A very similar equation is used to find the contact stress due to a tangential contact with the lens:

$$\sigma = 0.798 \left[\frac{F/2\pi r}{(1 - \nu_G^2)/E_G + (1 - \nu_m^2)/E_m} \right]^{1/2}. \tag{3.54}$$

Thermal Stresses in the Lens Due to Mounting. Because of the difference in thermal coefficient of expansion between optical and structural (metallic) materials, stresses are induced in the lens due to changes in temperature. In particular, a drop in temperature may produce severe stresses in the lens. These stresses may affect the optical performance of the lens, and in the worst case, cause mechanical failure.

To avoid radial stress in the lens produced as the cell contracts due to a temperature change, there must be some initial clearance between the lens and cell bore. The minimum radial clearance required to ensure that the cell does not contact the lens in the radial direction during a drop in temperature is given by

$$h_r \cong \frac{1}{2} \Delta T d_G (\alpha_m - \alpha_G) = \frac{1}{2} \Delta T (d_m \alpha_m - d_G \alpha_G), \tag{3.55}$$

where

h_r = minimum radial clearance
ΔT = temperature drop

d_G = lens outer diameter
d_m = cell inner diameter
α_G = thermal coefficient of expansion of the lens material
α_m = thermal coefficient of expansion of the cell material.

If sufficient radial clearance is not provided between lens and cell, a drop in temperature induces a radial stress in the lens. This radial stress is estimated from[25]:

$$\sigma_r = \frac{K_r \Delta T(\alpha_m - \alpha_G)}{(1/E_G + r_G/E_m h_m)}, \tag{3.56}$$

$$K_r = 1 - \frac{h_r}{r_G \Delta T(\alpha_m - \alpha_G)}, \tag{3.57}$$

where

σ_r = radial stress induced in the lens
r_G = lens radius
h_r = initial radial clearance
h_m = cell wall thickness.

3.5 LENS BARRELS

The lens barrel provides environmental protection and maintains the correct spacing between lens elements.

3.5.1 Barrel Materials

Properties desirable in a lens barrel material are:

1. thermal coefficient of expansion match with the lens material
2. low density to reduce weight
3. high elastic modulus to reduce structural deformation
4. low material cost
5. low machining cost
6. dimensional stability
7. resistance to corrosion
8. easy to blacken, to reduce stray light.

Materials used for lens barrels include:

1. *Magnesium alloys:* primarily used to reduce weight and fabrication cost. They have a high thermal coefficient of expansion in comparison with most lens materials, and poor resistance to corrosion.
2. *Aluminum alloys:* used to reduce weight and fabrication cost. Aluminum casting alloys, such as 356, are used to reduce the cost of complex barrel structures. The best dimensional stability is attained in 6061, although 7075 is a good substitute. These have a high thermal coefficient of expansion in comparison with most lens materials. They may be blackened using an anodizing process, although anodizing may produce a change in physical size.

3. *Titanium alloys:* used to reduce weight and to obtain a good thermal coefficient of expansion match with lens materials. They are expensive to fabricate. The 6A1-4V alloy is most commonly encountered. These have a very good resistance to corrosion.
4. *Stainless steel:* used for corrosion resistance and to obtain a good thermal coefficient of expansion match with lens materials. It is more expensive to fabricate than aluminum, but cheaper than titanium. Type 416 is relatively easy to machine and may be chemically blackened without a physical change in size. Type 17-4 PH (precipitation hardening) has superior dimensional stability and resistance to corrosion. The 300 series are used for vacuum applications; unfortunately, these alloys have poor dimensional stability.
5. *Brass:* used for corrosion resistance and to obtain a good thermal coefficient of expansion match with lens materials. It is heavy and more expensive to fabricate than aluminum. It is not as corrosion resistant as stainless steel. "Naval brass" has superior dimensional stability. It is often used in combination with either aluminum or stainless steel to obtain good friction and wear resistance in moving parts.
6. *Beryllium:* used primarily for extreme weight reduction. It has an excellent stiffness-to-weight ratio and has a thermal coefficient of expansion about half that of aluminum. It is very expensive to fabricate, is brittle, and very vulnerable to corrosion.
7. *Composites:* Composite materials are used primarily to reduce weight and to provide a better thermal coefficient of expansion match with lens materials. They are very expensive to fabricate, and their properties may deteriorate in the presence of moisture.
8. *Plastics:* Polycarbonates, and in particular the glass-reinforced polycarbonates, are the most common plastics used for lens barrels. Plastics are used to reduce weight and fabrication cost. A secondary advantage is resistance to corrosion. They have poor dimensional stability.

Metallic and nonmetallic materials' properties are given in Tables 3.4 and 3.5, respectively.

3.5.2 Barrel Design

Common Bore Assembly. The simplest configuration of a barrel is a common bore assembly. In a common bore assembly, all components—lenses, retainer, and spacers—are made to the same outer diameter and then placed in a barrel with a single common bore. Spacing between elements is maintained by spacers and a single retainer provides the preload force for the entire assembly. Friction between the components and the barrel bore makes it difficult to ensure contact between all spacers and lenses. Since the retainer must keep all elements in contact, the preload force tends to be very high.

Figure 3.11 shows a common bore assembly.

Individual Seat Assembly. In many cases, the diameters of the individual lenses will vary. Typically, the lens diameters will uniformly increase or decrease along the optical axis in the direction toward or away from a stop. In this case, an individual seat and retainer can be provided for each element in the barrel. For maximum precision, the seat can be machined to match the

Table 3.4 Metallic Materials Properties

Material	Density ρ (kg/m^3) $\times 10^3$	Elastic Modulus E (GPa)	Thermal Expansion $\propto$ (10^{-6}/K)	Specific Heat C_ρ (J/kg K)	Thermal Conduction K (W/m K)	Microyield Stress MYS (MPa)	Specific Stiffness E/P m $\times 10^6$	Thermal Distortion $\propto/K$ (m/W) $\times 10^{-9}$	Diffusivity $D = k/c_\rho\rho$ (m^2/s) $\times 10^{-6}$
Aluminum (6061-T6)	2.71	69.0	23.0	960	171	65–200	2.60	136	66
Beryllium (I-70)	1.85	304	11.2	1.82×10^3	220	17	16.8	50.9	65
Copper (OFHC)	8.94	11.7	16.7	385	392	15	0.133	42.6	114
Molybdenum (TZM)	10.2	318	5.0	272	146	160	3.18	103	53
Titanium (6AL-4V)	4.43	110	8.8	565	7.2	482	2.53	1.22	2.9
Magnesium (AZ 31B)	1.85	44.8	25.2	1000	76	41	2.47	332	41
Beryllium copper (Cu-2Be)	8.25	131	17.5	233	130	64–77	1.62	135	68
Invar 36	8.03	145	0.54	502	13.8	69	1.84	39.1	3.4
Stainless steel (Type 310)	8.03	193	16.6	502	16.3	122	2.45	1018	4.0
Aluminum-silicon carbide (SXA 24-T6)	2.91	117	13	795	125	117	4.10	104	54

Table 3.5 Nonmetallic Materials Properties

Material	Density ρ (kg/m^3) $\times 10^3$	Elastic Modulus E (GPa)	Expansion Coefficient $\propto$ (10^{-6}/K)	Specific Heat C_p (J/kg K)	Thermal Condition K (W/m K)	Microyield Stress MYS (MPa)	Specific Stiffness E/P (m) $\times 10^6$	Thermal Distortion $\propto/K$ (m/W) $\times 10^{-6}$	Diffusivity $D = k/\rho c_p$ (m^2/s) $\times 10^{-6}$
Glass (BK-7)	2.53	80.7	7.1	879	1.12	UTS ~7	3.19	6.3	0.504
Pyrex	2.23	65.5	3.3	838	1.13	UTS ~7	2.94	2.9	0.604
Fused silica	2.20	73.2	0.56	741	1.37	UTS ~50	3.33	0.41	0.840
Fused silica (ULE)	2.20	67.7	0.03	766	1.31	UTS ~50	3.08	0.023	0.777
Zerodur	2.53	91	0.050	821	1.64	UTS ~57	3.60	0.031	0.800
Sapphire	3.98	345	8.4	419	27.2	UTS ~414	8.67	0.31	16.3
Germanium	5.33	83	6.0	310	5.86	72	1.56	1.0	3.55
ZnSe (CVD)	5.27	71.9	7.3	340	18.0	UTS ~52	1.36	0.41	10.0
MgF	3.18	98.8	18.7	853	12.6	UTS ~86	3.11	1.48	4.65
Graphite/epoxy (G70, Isotropic)	1.78	93	0.02	921	35	138	5.22	0.00057	21.3
Silicon	2.33	131	4.2	544	130	UTS ~34	5.66	0.032	102

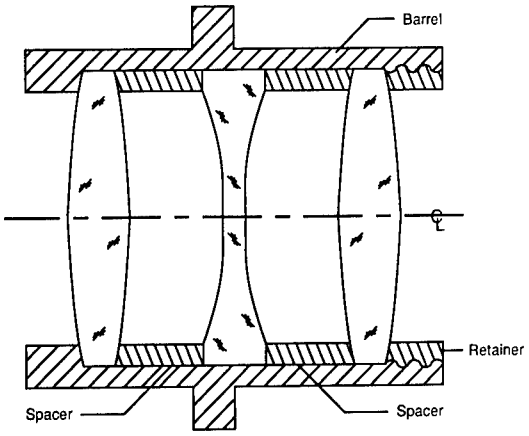


Fig. 3.11 Common bore assembly.

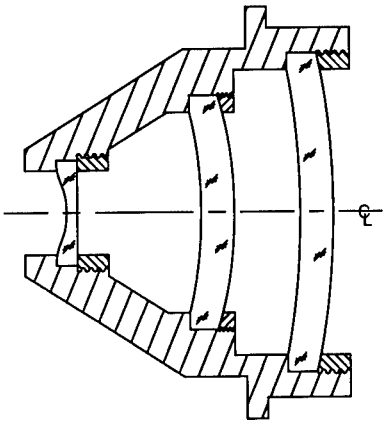


Fig. 3.12 Individual seat assembly.

measured diameter of the lens exactly. To guarantee centering of the individual lens elements, the critical portions of the barrel should be machined in a single setup.

Figure 3.12 shows an individual seat assembly.

Subcell Assembly. Maximum precision of the barrel assembly is obtained using subcell assembly. In subcell assembly, individual subcells each containing an element are interference fitted (press fitted) into a common ore. The centering error is then the sum of the centering error between lens and subcell, the out of roundness of the outside diameter of the subcell, and the out of roundness of the inside diameter of the barrel bore.[24] Typically, out of roundness is controlled during classic machining operations to about 2.5 μm. If the lens is centered with respect to the outer diameter of the subcell to 5 μm, a total centering error of 10 μm is obtained.

Two different techniques are used to center the individual lenses with respect to the subcells. In the first technique, the lens is spun into a metal cell and the cell outer diameter is machined concentric with the optical axis. The second technique centers the lens with respect to the subcell using mechanical and optical means, and then bonds the lens to the subcell. Semiflexible adhesive bonding is often used as a bonding technique to isolate the lens from expansion or contraction of the cell. A vent hole must be provided in each subcell to permit trapped air to escape as the subcell is pressed into the barrel.

Stress is induced in the lens when the subcell is pressed into the main barrel. If the lens is bonded to the subcell using semiflexible adhesive bonding, the stress induced in the lens is given by[32]:

$$H_C = 0.44 h_C \left[1 - \frac{1}{2}\left(\frac{L_C}{L_B}\right) + \frac{7}{4}\left(\frac{L_C}{L_B}\right)^2 \right] , \tag{3.58}$$

$$\varepsilon_r = \frac{1}{h_r}\left(\frac{\delta}{2}\right)\frac{h_B}{h_B + H_C} , \tag{3.59}$$

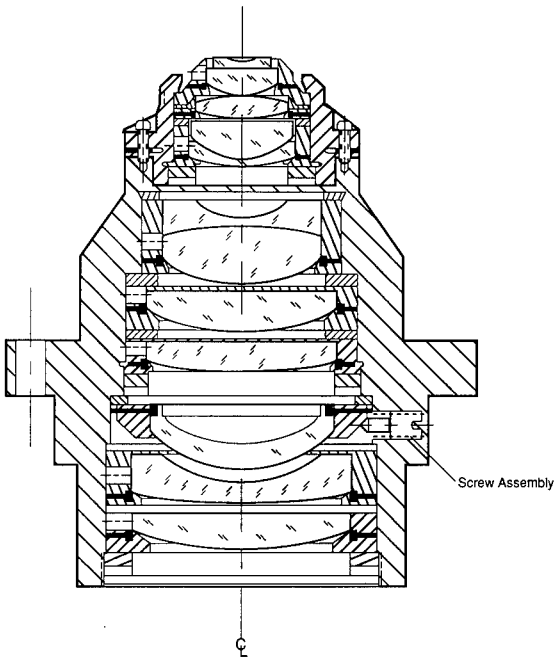


Fig. 3.13 Subcell assembly.

$$\sigma_G = \frac{E_r \varepsilon_r}{(1 + \nu_r)} \left[1 + \frac{\nu_r}{(1 - 2\nu_r)} \right], \tag{3.60}$$

where

σ_G = stress induced in the lens
E_r = elastic modulus of the adhesive
ν_r = Poisson's ratio for the adhesive
h_r = adhesive radial thickness
δ = radial interference between subcell and barrel
h_B = barrel radial thickness
h_C = subcell radial thickness
L_C = subcell axial length
L_B = effective axial length of the barrel (normally taken as 1.5 to 3 times the subcell axial length).

Figure 3.13 shows a subcell barrel.

3.5.3 Barrel Sealing

Barrels should be sealed to prevent the entrance of foreign matter. Dust and water vapor (which can cause fogging of the barrel interior with a temperature

drop) should be excluded. This requires hermetic sealing of the barrel. Hermetic sealing is obtained through the use of conventional O-ring seals in the lens barrel retainers. O-rings should not be placed in a seat unless precautions are taken to ensure full compression. If the O-ring is not fully compressed, the lens position is uncertain.

Semiflexible adhesive mounting of the lens provides an acceptable hermetic seal, because the adhesive extends completely around the circumference of the lens. If no pressure differential is expected between the barrel interior and the outside environment, a relatively simple flat gasket can be used.

If the barrel is exposed to pressure changes, it is desirable to provide vents to reduce pressure differentials. Pressure differentials across the outer (end) elements of the barrel can move the elements out of position and strain the elements out of shape. If a vent is provided, a dust filter and desiccant are very desirable to reduce the possibility of dust or water vapor entering the barrel interior.

Pressurizing the interior of the barrel with a dry gas is another technique for preventing the entry of foreign contaminants into the barrel. Typically dry nitrogen or dry helium is used. Both gases will diffuse through the barrel seals requiring periodic purging and refilling.

3.6 MIRROR DESIGN AND MOUNTING

Any surface error on a mirror is magnified by a factor of 2 on the reflected wave front. For this reason, mounting stress is very important in mirrors. Self-weight deflection of mirror surfaces is a problem in large mirrors. System constraints often require lightweight mirrors.

To minimize deflection due to self-weight or polishing pressure during fabrication, mirrors should be as thick as possible. The traditional rule for minimum thickness in right circular cylinder mirrors is a six-to-one diameter to thickness or aspect ratio. Much thinner mirrors can be made, but the fabrication cost increases rapidly with an increased aspect ratio. The sensitivity of mirror surfaces to deflection requires that support and preload forces be located so as to minimize bending.

3.6.1 Self-Weight Deflection of Mirrors

Self-weight deflection of mirror surfaces has two components: axial and radial. Axial deflection occurs when the mirror is on its back with the optical axis vertical. Radial deflection occurs when the mirror is on its edge, with its optical axis horizontal. At any arbitrary orientation, radial and axial deflection are combined to find the resulting surface deflection using the following equation[33]:

$$\delta_m = [(\delta_A \cos\theta)^2 + (\delta_R \sin\theta)^2]^{1/2} , \tag{3.61}$$

where

δ_m = self-weight deflection of the mirror
θ = angle between the mirror surface optical axis and local vertical
δ_A = axial component of the mirror surface deflection (surface deflection in the axis vertical position)

δ_R = radial component of the mirror surface deflection (surface deflection in the axis horizontal position).

Axial self-weight deflection is calculated using the following equation[34]:

$$\delta_A = C_S \frac{\rho}{E} \frac{d^4}{h^2}(1 - \nu^2) = C_S \frac{\rho h}{E} \frac{d^4}{12D} , \tag{3.62}$$

where

δ_A = axial self-weight mirror surface deflection
C_S = a support condition constant
ρ = mirror material density
E = modulus of elasticity of the mirror material
d = mirror diameter (or radius)
h = mirror thickness
ν = Poisson's ratio for the mirror material
D = mirror flexural rigidity.

Many axial mirror mounts consist of a number of support points in contact with the mirror back. The simplest multiple-point axial support uses three points on a common diameter. To produce minimum self-weight deflection, the support diameter should be 0.645 of the mirror diameter. In this case, the axial self-weight deflection of the mirror is given by[35]:

$$\delta_A = 0.318 \frac{\rho}{E} \frac{r^4}{h^2}(1 - \nu^2) = 0.0265 \frac{\rho h}{E} \frac{r^4}{D} , \tag{3.63}$$

where δ_A is the maximum axial surface deflection and r is the mirror mechanical radius.

It is sometimes more convenient to provide three support points equally spaced about the edge of the mirror. In this case, the axial self-weight deflection is given by:

$$\delta_A = 0.412 \frac{\rho}{E} \frac{r^4}{h^2}(1 - \nu^2) = 0.0344 \frac{\rho h}{E} \frac{r^4}{D} . \tag{3.64}$$

Equations (3.63) and (3.64) differ only in the value of the support condition constant and are variations of Eq. (3.62). A very similar approach is used to calculate the axial self-weight deflection of a rectangular mirror. A common three-point axial support for a rectangular mirror places one support point at the middle of one of the long edges, and the remaining two support points at the corners of the opposite long edge. If supported this way, the self-weight deflection of a rectangular mirror is given by[36]:

$$\delta_A = \frac{12}{C} \frac{\rho}{E} \frac{L^4}{h^2}(1 - \nu^2) , \tag{3.65}$$

$$C = \frac{7(L/b)}{\{1 + [0.461(L/b)]^{13}\}^{1/13}} , \tag{3.66}$$

where L is the length of the long edge of the mirror and b is the length of the short edge of the mirror.

Simple three-point axial mirror supports provide acceptable support for systems requiring "diffraction-limited" performance in the visible, if the mirror is small (below 0.2 m in diameter) and stiff (solid, with a six-to-one aspect ratio). Larger mirrors, or mirrors with low stiffness, require more complex support. Typically, a multipoint support scheme is used for large optics. The number of support points required is given by[37]:

$$N \cong \frac{1.5r^2}{h}\left(\frac{\rho}{E}\frac{1}{\delta}\right)^{1/2} , \tag{3.67}$$

where N is the number of support points required and δ is the required maximum self-weight deflection of the mirror surface.

Details on complex multiple-point axial support systems are found in the literature and are outside the scope of this chapter.[38] Normally, IR systems do not require mirror supports of this complexity. One multiple-point support scheme that is not difficult to implement uses six-point supports equally spaced on a common diameter that is 0.681 of the mirror diameter. Self-weight deflection of a mirror on a six-point support of this type is given by:

$$\delta_A = 0.0414\frac{\rho}{E}\frac{r^4}{h^2}(1 - \nu^2) = 0.00345\frac{\rho h}{E}\frac{r^4}{D} . \tag{3.68}$$

When the mirror is on its edge, with its optical axis horizontal, the radial component of the self-weight deflection of the mirror surface is calculated using an approach developed by Schwesinger[39]:

$$\gamma = \frac{r^2}{2hR} , \tag{3.69}$$

$$\delta_R = (a_0 + a_1\gamma + a_2\gamma^2)\frac{2\rho r^2}{E} , \tag{3.70}$$

where

δ_R = rms mirror surface deflection
R = mirror surface optical radius of curvature
r = mirror semidiameter
h = mirror axial thickness
γ = mirror thickness variation parameter,

and a_0, a_1, and a_2 are support condition constants.

Table 3.6 provides support condition constants for a variety of radial supports. Many common radial supports for mirrors use two support points located below the mirror. Radial surface deformation is at a minimum for a two-point support if the points are located 90 deg apart, and equally spaced with respect to local vertical.[40] In addition, for minimum surface deflection, friction between the two support points and the mirror edge should be as low as possible.

Table 3.6 Schwesinger Equation Constants

Support	a_0	a_1	a_2
One point at 0 deg	0.06654	0.7894	0.4825
Two points at 45 deg	0.05466	0.2786	0.1100
Two points at 60 deg	0.09342	0.7992	0.6875
180 deg sling	0.00074	0.1067	0.0308

Astigmatism is induced in the mirror optical surface if the two support points are not placed in the same plane as the mirror's center of gravity. For minimum radial self-weight deflection, one of the best supports is a simple sling or band wrapped around the lower 180 deg of the mirror.[41]

3.6.2 Lightweight Mirrors

Conventional solid-glass right circular cylinder mirrors are relatively heavy, with a weight given by[42]:

$$W = 246D^{2.92} , \tag{3.71}$$

where W is the mirror weight in kilograms and D is the mirror diameter in meters. For many applications, such as satellite systems, a mirror with a weight given by Eq. (3.71) is unacceptably heavy. Pointing and tracking systems are often relatively limited in drive torque, requiring the use of low moment of inertia and, hence, lightweight mirrors. The thermal inertia of solid mirrors may be excessive, requiring a lightweight mirror design. Lightweight mirrors are usually much more expensive to fabricate than conventional solid mirrors and are difficult to mount.

Conventional lightweight mirrors have about 30 to 40% of the weight of the same diameter solid six-to-one aspect ratio mirror. The weight of these mirrors is estimated by:

$$W = 120D^{2.82} , \tag{3.72}$$

where W is the mirror weight in kilograms and D is the mirror diameter in meters.

State-of-the-art lightweight mirrors have about 20% of the weight of the same diameter solid six-to-one aspect ratio mirror. The weight of these mirrors is estimated by:

$$W = 53D^{2.67} , \tag{3.73}$$

where W is the mirror weight in kilograms and D is the mirror diameter in meters. Figure 3.14 compares mirror weight as a function of diameter for lightweight and solid mirrors.

A useful parameter in evaluating performance of a lightweight mirror is the ratio of the self-weight deflection of the lightweight mirror to the self-weight deflection of a solid mirror of the same diameter. If this ratio exceeds

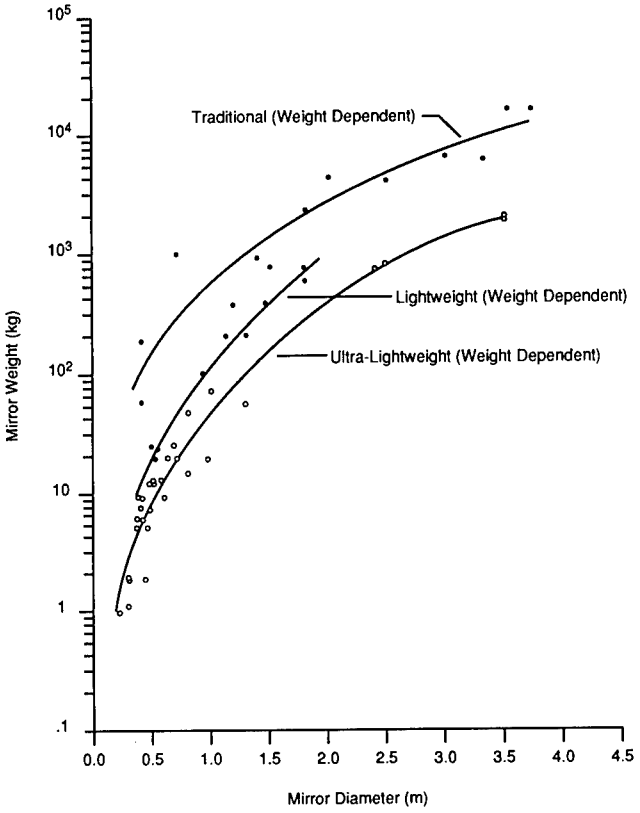


Fig. 3.14 Mirror weight as a function of diameter for traditional, lightweight, and ultra-lightweight mirrors.

unity, the lightweight mirror has less stiffness than the conventional solid mirror. A lightweight mirror that is not as stiff as a solid mirror of the same weight has very limited utility. More common, but less useful performance parameters for lightweight mirrors are the area density, and also weight in terms of percentage of an equivalent diameter six-to-one aspect ratio solid.[43]

Contoured Back Mirrors. Contouring the back of a lightweight mirror reduces the weight and improves the stiffness-to-weight ratio of the mirror. Three types of mirror back contour are used: the single arch, the double arch, and the symmetric shape. A contoured back mirror of optimum design can have a deflection ratio of 0.50 or less when compared to a solid mirror of the same weight.[44]

The simplest type of contoured back mirror is produced by tapering the back of the mirror from a thick hub to a thin edge. A straight taper is easiest to produce. Self-weight deflection for a single-tapered mirror is obtained by a parabolic taper on the back of the mirror, with the vertex of the parabola coincident with the back of the mirror. If a parabolic taper is used, this type of contoured back mirror is called a *single-arch mirror.* Deflection efficiency of the single arch is poor, with a deflection ratio of about 2 when compared with a solid mirror of the same weight. Weight of a single-arch mirror can be

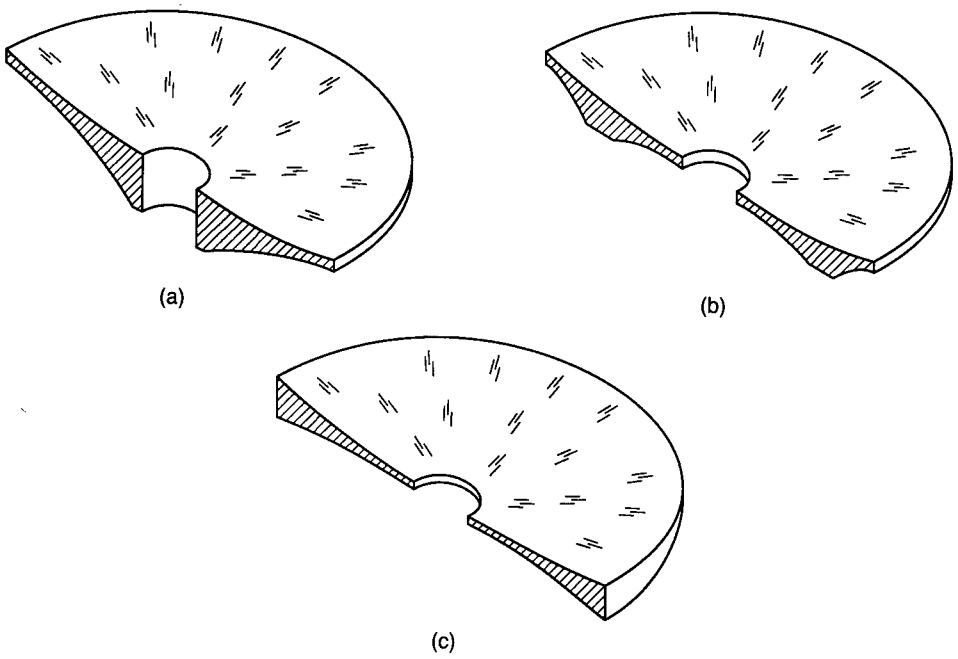


Fig. 3.15 Contoured back mirrors: (a) single arch, (b) double arch, and (c) double concave.

reduced to about 40% of the same diameter solid. Simplicity is the main virtue of the single-arch mirror. A single-arch mirror is normally mounted using a central hub mount. Since the weight of the mirror mount is very low, good efficiency is possible when the weight of the single arch and its mount is compared to the weight of a conventional solid mirror and mount.

A double parabolic taper is used in the double-arch mirror. The thickest portion of the mirror is a ring located at the 0.5 to 0.7 point of the mirror diameter. Mirror thickness is reduced using a parabolic taper from this ring toward a thin edge. Mirror thickness is also reduced using a parabolic taper from this ring toward the center, so that the center and edge of the mirror are both equally thin. Both parabolic tapers have vertices located at the back of the mirror. The self-weight deflection of a double-arch mirror is about 0.25 of a solid mirror of the same aspect ratio and about 0.16 that of a solid mirror of the same weight. Weight of double-arch mirrors can be reduced to about 30% of the same diameter solid mirror. Mounting of the double-arch mirror is more complex than that of a single-arch mirror, and requires the use of three or more support sockets in the back of the mirror.

If the mirror is intended for use in the axis horizontal position, a symmetric shape has the advantage of minimum self-weight deflection. Symmetric double convex or double concave mirrors have very small radial components of deflection in the one edge position. When used in the axis vertical position, the symmetric shape produces excessive self-weight deflection.

Contoured back mirror shapes are shown in Fig. 3.15.

Rib-Reinforced Lightweight Mirrors. A rib-reinforced lightweight mirror consists of a thin faceplate with a regular array of ribs attached to the back of the faceplate. The regular array of intersecting reinforcing ribs forms repeating pockets or cells in the back of the mirror. This type of mirror is called an *open-back mirror.* If a second plate is attached to the back of the array of ribs, a sandwich structure is developed. This type of mirror, with two plates connected by an array of ribs, is called a *sandwich mirror.*

Although complex in structure, the self-weight deflection of either open-back or sandwich mirrors can be calculated by reducing the stiffness of the mirror to that of an equivalent solid mirror. This equivalent stiffness is the equivalent flexural rigidity of the mirror. For an open-back mirror, the equivalent flexural rigidity is given by[45]:

$$\eta = \frac{(2B + t_c)t_c}{(B + t_c)^2} , \tag{3.74}$$

$$h_B^3 = \frac{\{[1 - (\eta/2)][h_F^4 - (\eta h_F^4/2)] + (h_F + h_c)^4(\eta/2)\}}{[h_F + (\eta 2h_c/2)]} , \tag{3.75}$$

$$D = \frac{Eh_B^3}{12(1-\nu^2)} , \tag{3.76}$$

where

D = equivalent flexural rigidity of the mirror
h_B = equivalent bending thickness
h_F = faceplate thickness
E = modulus of elasticity of the mirror material
ν = Poisson's ratio for the mirror material
B = cell inscribed circle diameter ($B = 0.5774a$ for a triangular cell, $B = a$ for a square cell, $B = 1.732a$ for a hexagonal cell)
a = length of the cell sides
η = rib solidity ratio
t_c = cell wall thickness
h_c = rib height.

The weight of an open-back mirror is given by:

$$W = \frac{\pi}{4}\rho d^2(h_F + \eta h_c) , \tag{3.77}$$

where

W = mirror weight
d = mirror diameter
ρ = mirror material density.

Similar methods are used to find the equivalent flexural rigidity of a sandwich mirror. A sandwich mirror employs a faceplate and a backplate, connected by an array of intersecting ribs forming a repeating pattern of cells inside the mirror. The equivalent flexural rigidity of a sandwich mirror is given by:

$$h_B^3 = (2h_F + h_c)^3 - \left(1 - \frac{\eta}{2}\right) h_c^3 \ , \tag{3.78}$$

$$D = \frac{Eh_B^2}{12(1 - \nu^2)} \ , \tag{3.79}$$

where h_F is the face plate thickness (the backplate is assumed to have the same thickness) and h_B is the equivalent thickness of the mirror.

The weight of a sandwich mirror is given by:

$$W = \frac{\pi}{4} \rho d^2 (2h_F + \eta h_c) \ . \tag{3.80}$$

There is an optimum relationship between faceplate thickness, rib thickness, and cell size for minimum self-weight deflection. Figures 3.16 and 3.17 relate these parameters for an open-back mirror and a sandwich mirror, respectively. Some caution is indicated in employing these figures, due to a fabrication issue involving mirror surface deformation corresponding to the structure of the ribs. During fabrication, the mirror faceplate will deflect between ribs due to polishing pressure. This deflection, called *quilting* (because the periodic pattern of deflection resembles the square of a quilt), is permanently polished into the mirror surface figure. Quilting causes a reduction in energy in the central maximum of the diffraction disk. Reduction of quilting is possible by increasing the faceplate thickness, decreasing the spacing between ribs, or by using lighter polishing pressures. Figure 3.18 gives the representative quilting for different types of cells.

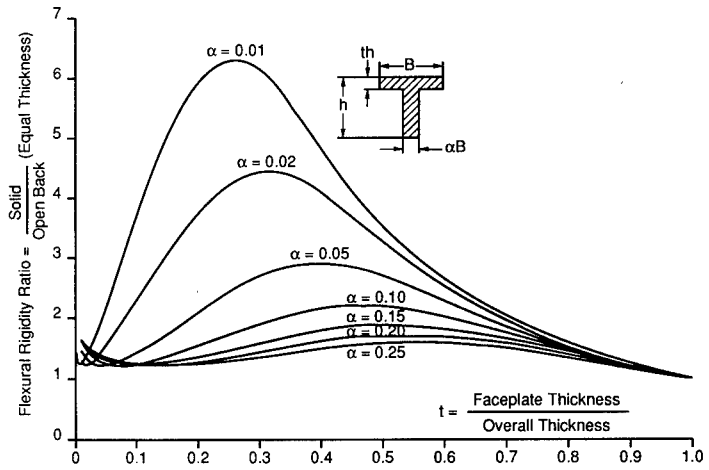


Fig. 3.16 Open back mirror flexural rigidity.

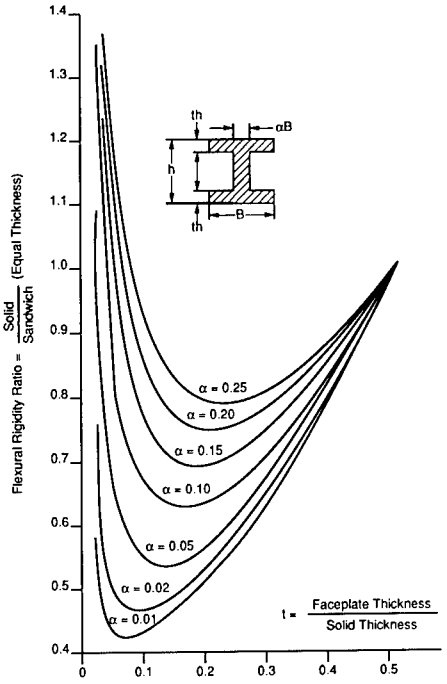


Fig. 3.17 Sandwich mirror flexural rigidity.

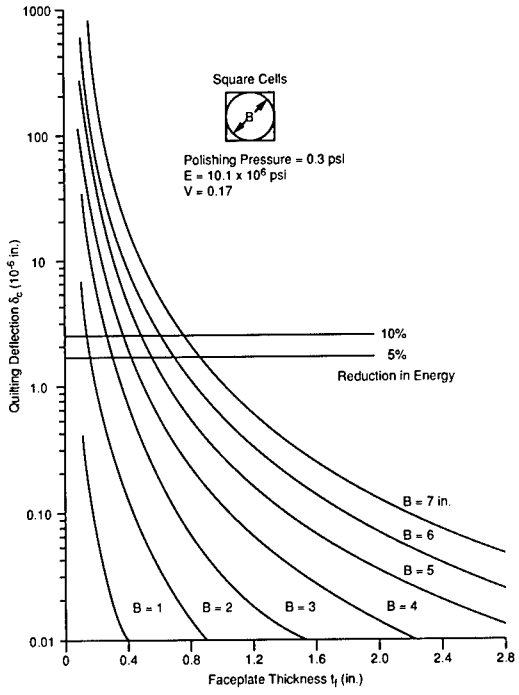


Fig. 3.18 Faceplate quilting under polishing pressure for lightweight mirors.

3.6.3 Metal Mirrors

Mirrors made of metal are a common feature of many IR systems.[46] Beryllium and aluminum are the most common metal mirror materials. Copper and molybdenum are used in high-energy laser mirror applications. Interest is growing in the metal matrix composites, most notably the aluminum/silicon carbide reinforced materials. Metal mirrors have the following advantages:

1. *Ease of fabrication:* Metal mirrors are fabricated using nontraditional methods such as casting, machining, and diamond turning. This ease of fabrication makes it possible to utilize surface shapes such as off-axis aspherics, which would be very difficult to fabricate in glass using traditional optical shop methods.
2. *Ease of mounting:* It is possible to produce threaded holes in metal mirrors and to use these holes for mounting. Integral mounts can be combined into some mirror configurations.
3. *Thermal stability:* The thermal conductivity of metals used in mirrors is much higher than that of glass mirror materials. Metal mirrors reach thermal equilibrium faster than glass mirrors and have good resistance to thermal distortion due to temperature gradients.
4. *Athermalization:* By making the mirrors and support structure out of a single metal, the system alignment and focus are maintained over a wide range of temperatures.
5. *Specific stiffness:* The stiffness-to-weight ratio of certain metals, notably beryllium, is much better than that of traditional optical materials. Use of high stiffness-to-weight ratio metals allows reduction in mirror weight.
6. *Nuclear hardness:* Certain metals, notably beryllium, are resistant to damage when exposed to high nuclear radiation fluxes.

Metal mirrors have significant disadvantages:

1. *Long-term stability:* Metals exhibit dimensional instability. The optical surface figure of metal mirrors changes with time.
2. *Surface scatter:* Scatter from a bare metal mirror surface is excessive for some applications.
3. *Bimetallic bending:* Surface scatter from a bare metal mirror surface is reduced by plating the mirror with electroless nickel. The thermal coefficient of expansion of electroless nickel is different from the thermal coefficients of expansion of either aluminum or beryllium (the two most common metal mirror materials). This difference in thermal coefficient of expansion can cause the metal mirror surface to deform as the temperature changes and induces stress in the mirror.
4. *Corrosion:* Metals are electrochemically active and are vulnerable to corrosion in certain service environments.

Dimensional stability is a major concern in metal mirrors. Changes in the optical figure of metal mirrors amounting to one wave per year (1 wave = 633 nm) are well documented.[47] Cyclic temperature changes or stress induced by mounting can also cause permanent deformations in the mirror surface.[48] The three basic causes of dimensional instability of metal mirrors are metallurgical instability, microcreep under load, and release of residual stress.

Metallurgical stability is a complex subject whose origins are outside the scope of this chapter.[49] Certain metals exhibit worse stability than others. For example, Invar has a stability of about one to two parts per million per year,[50] while the stainless steel 17-4 PH has a stability of about 1 part in 10 million per year.[51] Long-term stability is best assessed from experience.[52] As a rough rule of thumb, the higher the transition temperature or melting temperature of the material, the better the stability.

Andrade's beta law is often used to predict dimensional change in materials. Although no absolute proof of the applicability of this law to all materials exists, it has proven reliable for the metals commonly used in mirrors.[53] Andrade's beta law states that the change in physical size with time of a metal is proportional to the time raised to a power. This power is typically about 0.33, or:

$$\varepsilon = \beta t^m \ , \tag{3.81}$$

where

ε = creep strain at some time
β = a constant depending on material, stress, and temperature
t = time
m = time dependent constant (this is usually about 0.33, with a range of 0.26 to 0.39).

Determining the amount of microcreep that may occur under load is difficult. It is possible to determine the microyield properties of the material. Microyield is defined as a permanent plastic deformation in the metal that occurs at a stress below the yield stress of the material. *Microyield strength* is defined as the amount of stress required to produce a permanent plastic deformation in the material of one part per million. Microyield strength (MYS) is also known as the *precision elastic limit* (PEL) of the material. A common practice in designing metal optics is to limit all stresses below the microyield strength of the material. This practice attempts to limit optical surface deformation due to dimensional instability. For certain metals in the strain range of 10^{-7} to 10^{-4}, the following equation is used to predict microyield behavior[54]:

$$\sigma = K\varepsilon^n \ , \tag{3.82}$$

where

σ = applied stress
K = microstrain hardening coefficient
n = microstrain hardening exponent
ε = permanent strain.

Residual stress release in metal mirrors over time is another source of dimensional instability.[55] Residual stress release is reduced by using low residual stress fabrication processes and by heat treatment after fabrication. Standard simple stress relieving heat treatments are insufficient for optimum stability of metal mirrors. Special cyclic heat treatments are necessary for optimum stability.[56] Table 3.7 gives stabilizing heat treatments for aluminum, beryllium, and metal matrix composites.

Table 3.7 Some Suggested Stabilization Heat Treatments for Metal Mirrors

1. Aluminum (6061-T6)
   a. Heat parts in furnace to 375°F ± 10°F. Hold for 3 h. Furnace or still air cool to room temperature. Chill to −40°F for 0.5 h. Warm to room temperature. Temperature change not to exceed 15°F per minute.
   b. Repeat step (a).
   c. Heat to 325°F for 1 h. Still air cool to room temperature.
2. Beryllium
   a. Reduce temperature to 80 K ± 10 K. Hold for 0.5 h. Warm to room temperature. Temperature change not to exceed 100 K per hour.
   b. Hold at room temperature for 0.5 h.
   c. Raise temperature to 373 K. Hold for 0.5 h. Cool to room temperature.
   d. Hold at room temperature for 0.5 h.
   e. Repeat steps (a) through (d).
   Note: Mirror is maintained in a dry atmosphere to eliminate condensation during heat treatment.
3. Aluminum/silicon carbide reinforced metal matrix composite
   a. Cool to −75°F ± 10°F. Hold for 1 h. Temperature change not to exceed 5°F per minute. Do not quench mirror.
   b. Warm to 212°F. Hold for 1 h.
   c. Cool to room temperature.
   d. Repeat steps (a) through (c) for a total of five cycles.

The most common metal mirror material is aluminum. Type 6061-T6 aluminum alloy is the most common aluminum used for metal mirrors. Type 6061 aluminum is inexpensive, has well-established properties, and has good dimensional stability. Diamond turning can be used with 6061 aluminum to produce mirror surfaces. Metallurgical impurities in 6061 cause relatively rapid wear in diamond tools during the diamond-turning process. To avoid this wear, the 50 series aluminum alloys are sometimes used. The 50 series alloys are available in higher purity and give better tool life in diamond-turning operations. Unfortunately, the long-term stability of the 50 series alloys is less than that of the 60 series.[57]

Beryllium is an exotic metal that offers a stiffness-to-weight ratio about 6.5 times greater than that of aluminum. Beryllium has a thermal coefficient of expansion about half that of aluminum with comparable thermal conductivity. These very attractive properties are at least partially offset by controversial long-term dimensional stability and by the high cost of beryllium.

Beryllium is a highly anisotropic metal. This anisotropy is overcome by the use of powder metallurgy, by using such techniques as hot isotropic pressing to produce mirrors. Residual anisotropy of beryllium is at least partially responsible for the controversial performance of this material.[58]

Beryllium is a toxic material. Toxicity is an issue if beryllium dust is inhaled into the lungs. There is no hazard associated with handling or using beryllium metal, so long as no beryllium particulate is produced. Appropriate care during fabrication operations reduces risk due to the toxicity of beryllium to a very low level. Toxicity and the requisite handling precautions are partially responsible for the high cost of beryllium.

The metal matrix composites (MMC), notably the aluminum/silicon carbide reinforced materials, have considerable promise for use in IR systems.[59] Aluminum/silicon carbide composites can be tailored to have a specific stiffness

about 1.6 times greater than aluminum. The thermal coefficient of expansion of the MMC materials can be tailored to match exactly that of electroless nickel plating, therefore eliminating a cause of thermal distortion in metal mirrors. Initial studies of MMC materials indicate that these materials have higher microyield strengths than either beryllium or aluminum. The cost of the MMC materials is higher than aluminum, but may be competitive with beryllium.

3.6.4 Thermal Effects on Mirrors

Changes in temperature, or temperature gradients, can cause changes in the optical surface figures of mirrors. Since IR systems are often used over wide temperature ranges, evaluation of temperature effects on mirrors is an important topic.

General Effects of Temperature on Mirrors. If the mirror is exposed to an overall change in temperature, and to linear temperature gradients in each axis of a Cartesian coordinate system centered at the mirror optical surface vertex, with the temperature distribution of the form[60]:

$$T(x,y,z) = C_o + C_1(x) + C_2(y) + C_3(z) \ , \tag{3.83}$$

where $T(x,y,z)$ is the temperature distribution, C_o is the overall mirror temperature, and C_1, C_2, and C_3 are the linear temperature gradients, in each axis (x,y,z).

Mirror distortion caused by the preceding temperature distribution is given by:

$$\begin{aligned}
\delta(r,\theta) = {} & \frac{\alpha C_3}{8R^2} r^4 && \text{(spherical)} \\
& + \frac{\alpha C_1}{2R} r^3 \cos\theta + \frac{\alpha C_2}{2R} r^3 \sin\theta && \text{(coma)} \\
& + \left(\frac{\alpha C_3 h}{2R} - \frac{\alpha C_3}{2} + \frac{\alpha C_o}{2R} \right) r^2 && \text{(focus)} \\
& + (\alpha C_1 h - C_4) r \cos\theta + (\alpha C_2 h - C_5) r \sin\theta && \text{(tilt)} \\
& + \frac{\alpha C_3 h^2}{2} + \alpha C_o h - C_6 \ , && \text{(piston)}
\end{aligned} \tag{3.84}$$

where

$\delta(r,\theta)$ = mirror surface deformation
r = radial position on the mirror surface
θ = angular position on the mirror surface
α = thermal coefficient of expansion of the mirror material
R = optical surface radius of curvature
h = axial thickness of the mirror
C_4, C_5 = constants defining the tilt of the mirror
C_6 = a constant defining the axial position of the mirror surface.

The mirror surface radius of curvature changes if the mirror is exposed to a steady-state linear axial thermal gradient. Change in the radius of curvature is given by:

$$\frac{1}{R_0} - \frac{1}{R} = \frac{\alpha}{K} q \ , \tag{3.85}$$

where

R = original radius of curvature
R_0 = radius of curvature of the thermally distorted mirror
K = thermal conductivity of the mirror material
q = heat flux absorbed per unit area.

In Eq. (3.85) the material parameter determining mirror deformation is the ratio of thermal coefficient of expansion to thermal conductivity. This ratio is called the *thermal distortion parameter* of the material. The thermal distortion parameter is often used to select materials for use in applications where there is a thermal gradient. Metals have better thermal distortion parameters than ordinary glasses. For example, aluminum has a thermal distortion parameter of 136×10^{-9} m/W, while a borosilicate glass has a thermal distortion parameter of 2.9×10^{-6} m/W.

It is often of interest to determine how rapidly a mirror will reach thermal equilibrium following a temperature change. The time required for an optical element to reach thermal equilibrium is controlled by the material property of thermal diffusivity. The thermal diffusivity of a material is given by:

$$a = \frac{K}{\rho C_\rho} \ , \tag{3.86}$$

where

a = thermal diffusivity
ρ = material density
K = material thermal conductivity
C_ρ = material specific heat.

Heat transfer in mirrors is a complex subject that is beyond the scope of this discussion. A simple approximation is often useful to determine the rate at which the mirror changes temperature. If only one side of the mirror transfers heat, and if there is no transfer at the edge of the mirror, the interior temperature of the mirror after some time is given by:

$$T' = T\left[1 - \exp\left(\frac{\pi^2 ta}{h^2}\right)\right] \ , \tag{3.87}$$

where

T = initial mirror temperature
T' = mirror temperature following a sudden temperature change after time t
h = mirror thickness.

Effects of Spatial Variation of Thermal Coefficient of Expansion of Mirror Materials. Most mirror materials have a spatial variation of thermal coefficient of expansion. A spatial variation of thermal coefficient of expansion in the mirror material causes the mirror surface to distort as the temperature changes. For example, in a right circular cylinder mirror, a spatial variation of thermal coefficient of expansion from center to edge will cause a change in mirror radius of curvature as the temperature is changed. As a rule of thumb, mirror materials have a spatial variation in thermal coefficient of expansion of about 3%. This implies that low thermal coefficient of expansion materials are less affected by this variation than high thermal coefficient of expansion materials. Table 3.8 gives values of the spatial variation of thermal coefficient for materials.

If there is a linear variation in thermal coefficient of expansion through the axial thickness of the mirror, the change in mirror figure with temperature is given by:

$$\delta \cong \frac{r^2}{2h}\Delta\alpha\Delta T \ , \tag{3.88}$$

where

δ = change in mirror optical surface figure
$\Delta\alpha$ = total change in thermal coefficient of expansion, through the axial thickness of the mirror
ΔT = temperature change
h = mirror axial thickness
r = mirror mechanical radius.

Astigmatism is caused in the mirror optical surface figure if the thermal coefficient of expansion is different in one direction in the plane of the mirror.

Table 3.8 Spatial Variation in Thermal Coefficient of Expansion for Various Mirror Materials

	α Homogeneity	
Material	$\alpha_{(300\ K)}$ 10^{-6}/K	$\Delta\alpha_{(300\ K)}$ 10^{-9}/K
1. Aluminum (6061-T6)	23.0	60
2. Beryllium (VHP I-70A)	11.5	130
3. Beryllium (HIP I-70A)	11.5	30
4. Borosilicate (Schott Duran)	3.2	30
5. Borosilicate (Ohara E-6)	3.0	50
6. Fused quartz (Heraeus Amersil T08E)	0.50	5
7. Fused silica (Corning 7940)	0.56	2.0
8. ULE (Corning 7971)	0.03	4.0
9. Glass ceramic (Schott Zerodur)	0.05	40
10. Glass ceramic (Cervit C-101)	0.03	15

The amount of optical surface deflection caused by this spatial variation in thermal coefficient of expansion is given by:

$$\delta \cong \frac{r^2}{4h}(\alpha_x - \alpha_y)\Delta T \ , \tag{3.89}$$

where α_x is the thermal coefficient of expansion in the x direction of the mirror surface plane and α_y is the thermal coefficient of expansion in the y direction of the mirror surface plane.

Bimetallic Bending Effects in Metal Mirrors. Metal mirrors are plated with electroless nickel to reduce surface scatter and to enhance surface reflection. Electroless nickel has a different thermal coefficient of expansion, 13×10^{-6} m/m K, than either aluminum, 23×10^{-6} m/m K, or beryllium, 11.2×10^{-6} m/m K (see Ref. 61). The difference in thermal coefficient of expansion causes a bimetallic bending in the mirror surface when the plated mirror temperature is changed. If the mirror temperature is changed uniformly, so that there are no temperature gradients, and if both sides of the mirror are plated, the resulting optical surface deflection and mirror stresses are given by[61]:

$$\frac{1}{R_o} - \frac{1}{R} = \frac{6}{h}\frac{\psi_p}{\psi_m}(\alpha_p - \alpha_m)\Delta T \times \left(\frac{h_{p1} - h_{p2}}{h}\right)\left[1 - \left(\frac{h_{p1} - h_{p2}}{h}\right)\left(4\frac{\psi_p}{\psi_m} - 1\right)\right] \ , \tag{3.90}$$

$$\sigma_{rm} = \psi_p(\alpha_p - \alpha_m)\Delta T\left[3\frac{(h_{p2} - h_{p1})}{h} + \frac{(h_{p2} + h_{p1})}{h}\right] \ , \tag{3.91}$$

$$\sigma_{rp} = \frac{\psi_p}{\psi_m}\sigma_{rm} - \psi_p(\alpha_p - \alpha_m)\Delta T \ , \tag{3.92}$$

$$\psi_m = \frac{E_m}{1 - \nu_m} \ , \tag{3.93}$$

$$\psi_p = \frac{E_p}{1 - \nu_p} \ , \tag{3.94}$$

where

R_o = original mirror surface radius of curvature
R = mirror surface radius of curvature after the temperature change
h = mirror axial thickness
h_{p1}, h_{p2} = plating thicknesses on the front and back of the mirror
α_m = mirror material thermal coefficient of expansion
α_p = plating material thermal coefficient of expansion
E_p = plating material elastic modulus
E_m = mirror material elastic modulus

ν_m = Poisson's ratio for the mirror material
ν_p = Poisson's ratio for the plating material
ΔT = temperature change
σ_{rm} = temperature-induced radial stress in the mirror surface
σ_{rp} = temperature-induced radial stress in the mirror plating material
ψ_p = plating material parameter
ψ_m = mirror material parameter.

Surface deflection of the plated mirror is eliminated by making the plating thickness on the front and back of the mirror equal. Although equal plating thickness eliminates surface distortion as the mirror temperature is changed, stress is still induced in the mirror and plating. Radial stress in the mirror may cause a change in mirror optical surface figure if the stress exceeds the microyield strength of the mirror material. Mirror stress is reduced by increasing mirror thickness.

A similar set of equations is used to determine the change in the mirror optical surface figure when the plated mirror is exposed to a steady-state temperature gradient:

$$\frac{1}{R_o} - \frac{1}{R} \cong \frac{\alpha_m q}{K_m}\left[1 + \frac{3\psi_p}{\psi_m}\frac{(\alpha_p - \alpha_m)}{\alpha_m}\frac{h_p}{h}\right], \tag{3.95}$$

$$\sigma_{rm} = 0, \tag{3.96}$$

$$\sigma_{rp} = -\psi_p(\alpha_p - \alpha_m)\frac{qh}{K_m}, \tag{3.97}$$

$$\Psi_p = \frac{E_p}{1 - \nu_p}, \tag{3.98}$$

$$\psi_m = \frac{E_m}{1 - \nu_m}, \tag{3.99}$$

where

R = mirror surface radius of curvature after the mirror is exposed to an axial heat flux
K_m = mirror material thermal conductivity
q = axial heat flux per unit area absorbed by the mirror
h_p = plating thickness
h = mirror thickness
σ_{rm} = radial stress in the mirror induced by the axial heat flux
σ_{rp} = radial stress in the plating induced by the axial heat flux

3.6.5 Mirror Mounting

Kinematic Mirror Mounts. Mirror mounts are designed using kinematic principles. According to kinematic theory, any rigid body has $6 - N$ degrees of

freedom, where N is the number of points in contact with the body. Design of a mirror mount involves selecting the number and location of the N contact points. Any mirror that is contacted by more than six contact points is said to be overconstrained. An overconstrained mirror is likely to be distorted and uncertain in position.

Circular mirrors are normally not sensitive to rotation of the mirror about its optical axis. Since rotation of the mirror about the optical axis does not have to be constrained, a kinematic mount for a circular mirror will use five contact points. The classic form of a five-point kinematic mount for a mirror consists of three contact points against the back of the mirror and two contact points below the mirror. The three contact points against the back of the mirror should be equally spaced on a diameter that is 0.645 of the overall mirror diameter. The two support points below the mirror should be 90 deg apart, and equally spaced at 45 deg on either side of local vertical. This five-point configuration minimizes self-weight deflection of the mirror surface. Figure 3.19 shows the classic five-point kinematic mirror mount.

An alternative kinematic configuration uses three equally spaced points of attachment to the mirror circumference. Each point of attachment provides constraint in two directions: radial and tangential. Self-weight deflection of

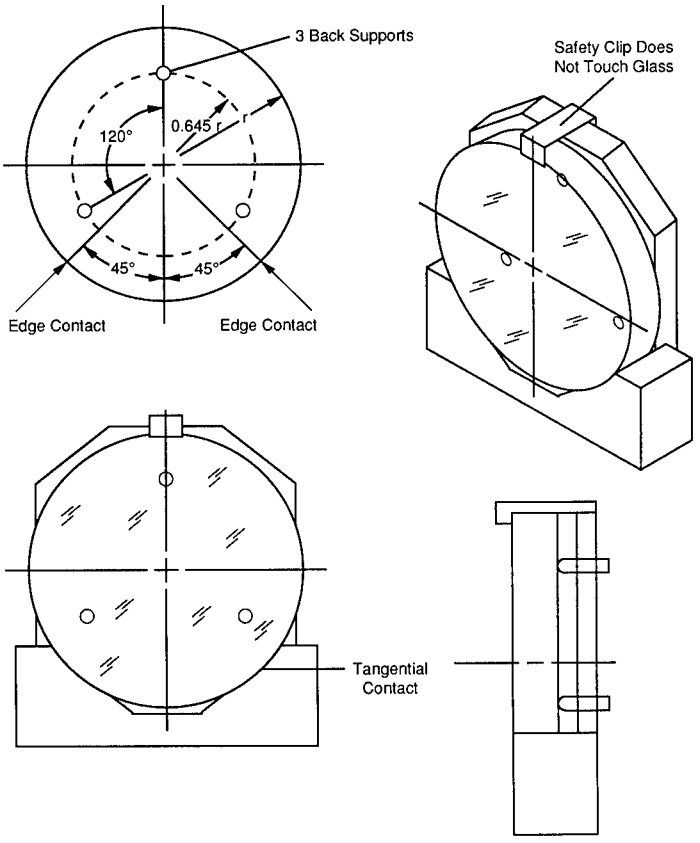


Fig. 3.19 Five-point kinematic mirror mount.

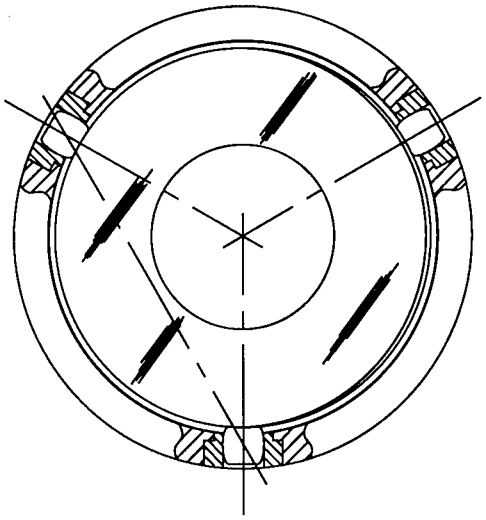


Fig. 3.20 Semikinematic mirror mount patented by Mesco.

this type of kinematic mount is greater than that of the classic five-point mount. The main advantages of the three-point edge mount are the ability to attach the mirror to the mount to a surrounding ring and the suitability for using the three mounting points to athermalize the mirror.

Rectangular mirrors are mounted using a variant of the classic five-point kinematic support. The back of the mirror is provided with three supports, and two more supports are located along one edge. One back support is located at the midpoint of one long edge. The other two back supports are located at the corners of the opposite edge. Both edge supports are located on the same long edge as the preceding two supports, and are located a distance 0.22 of the overall edge length from each corner. This five-point support minimizes self-weight deflection of a rectangular mirror. Figure 3.20 shows a five-point kinematic mount for a rectangular mirror.

Semikinematic Mirror Mounts. True point contact between the mirror and mount produces relatively large contact stresses in the point contacts. These large stresses may exceed the microyield strength of the materials used, resulting in instability of the mirror position. Deflection may be excessive in point contacts due to high stress. Contact stress is reduced if each contact point is replaced with a small contact area. Kinematic principles are used to locate the contacts. This concept of using kinematic located points with areas of contacts large enough to reduce stresses to acceptable levels is called *semikinematic design*. Semikinematic mounting is the most frequently used technique for mirror mounts.

The geometry of the contact areas used in semikinematic mounting is a major design problem. If the contact areas are not flat and coplanar, stress is induced in the mounted optic. Stress is reduced by improving the flatness and by reducing tilt between the contact areas. Semikinematic mounts are often used in diamond-turning applications where the usual tolerance on flatness and coplanarity of the contact area is the same as the mirror surface. For most

applications, this tolerance is too extreme and can be relaxed. A more realistic tolerance for flatness and tilts of the contact surface is a tolerance equivalent to the surface deformation in the contact area.

An alternative to very small tolerances in the contact areas is to allow one contact in each pair of contacts to rotate with respect to the other. The simplest arrangement for providing this rotation is a spherical surface in contact with a cone. This sphere and cone geometry is easily modified to that of a sphere sliding in a cylinder. Three such sphere cylinder contacts around the edge of a mirror provide a semikinematic mount that is also athermalized. A change in physical size of the mount with temperature due to a difference in thermal coefficient of expansion between mount and optic is removed by radial translation of the spheres in their respective cylindrical mounts. This type of mirror mount has been patented by Mesco[62] and is shown in Fig. 3.20.

Flexural Mirror Mounts. Flexural mirror mounts are similar in principle to semikinematic mounts.[63] Kinematic principles are used to determine the stiffness of each flexure used to mount the mirror. The classic flexural mirror mount uses three flexures attached to the edge of the mirror. Each flexure is stiff in the axial and tangential directions and is compliant in the radial direction. This combination of stiffness and compliance accurately locates the mirror, yet allows the mount to expand or contract radially with respect to the mirror. The mirror is isolated from stresses due to temperature changes and thermal coefficient of expansion difference between mirror and mount.[64]

The stiffness of a flexure-mounted mirror is given by:

$$K_R = \frac{3}{2}(K_{FR} + K_{FT}) \ , \tag{3.100}$$

$$K_A = 3K_{FA} \ , \tag{3.101}$$

where

K_R = radial stiffness of the mirror mount
K_{FR} = radial stiffness of one mirror mount flexure
K_{FT} = tangential stiffness of one mirror mount flexure
K_A = axial stiffness of the mirror mount
K_{FA} = axial stiffness of one mirror mount flexure.

Error in locating the flexures can induce bending forces into the mirror. This lack of assembly accuracy complicates the design of the mirror mount flexures. The effect of assembly inaccuracy is reduced by introducing additional directions of stiffness and compliance into each mirror mount flexure. In the classic three-flexure design, assembly inaccuracy effects are reduced by placing a set of flexures between the mounting flexures and the mirror. This additional set of flexures provides compliance in rotation about two orthogonal axes. Both rotation axes are parallel to the plane of the mirror surface. Further reduction in mirror surface deflection is possible by attaching the flexures to the mirror in the plane of the center of gravity.[65] An example of a flexure-mounted mirror is shown in Fig. 3.21.

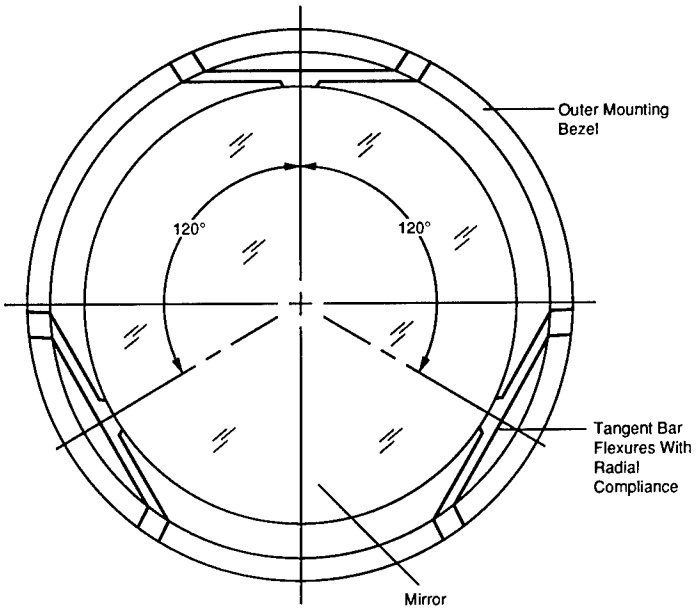


Fig. 3.21 Flexure mirror mount with tangent bars.

Bonded Mirror Mounts. Bonding of mirrors into mounts employs semikinematic principles. Bonded areas should be located using kinematic principles. The thickness of the bond between mirror and mount allows relaxation of the tolerances of the contact areas. Bonding has the disadvantages of limited stiffness and possible outgassing of the adhesive.

The simplest type of bonded mirror mount consists of a large circular bonded area on the back of the mirror. The bond area is determined using Eq. (3.35). Stability of the mirror position is improved by providing three contact points against the back of the mirror in addition to the bonded area. The bonded area then acts as a spring to keep the mirror in contact with the defining points.

Large areas of adhesive in contact with the mirror can produce severe stresses when the temperature is changed due to a difference in thermal coefficient of expansion between mirror and adhesive. Thermal stress is reduced by breaking up the large bond area into smaller areas. One possible semikinematic configuration uses three equally spaced bond areas on the back of the mirror. The three bond areas are located at the optimum 0.645 diameter.

Most adhesives shrink as they cure. Shrinkage of the adhesive used to mount a mirror can cause stress and surface deformation. Stress caused by shrinkage of the adhesive is reduced by isolating the area of bonding from the mirror surface.

Isolation of the bond area from the mirror optical surface is possible by using a modified mirror configuration. One such configuration uses a cylindrical "stalk" on the back of the mirror. This stalk is normally about one-third the mirror diameter and has a length equivalent to the mirror thickness. Adhesive is placed around the circumference of the stalk. This type of mirror mount is known as a *mushroom mount,* from the resemblance of the mirror to a mushroom.[66] Figure 3.22 shows a bonded mushroom mirror.

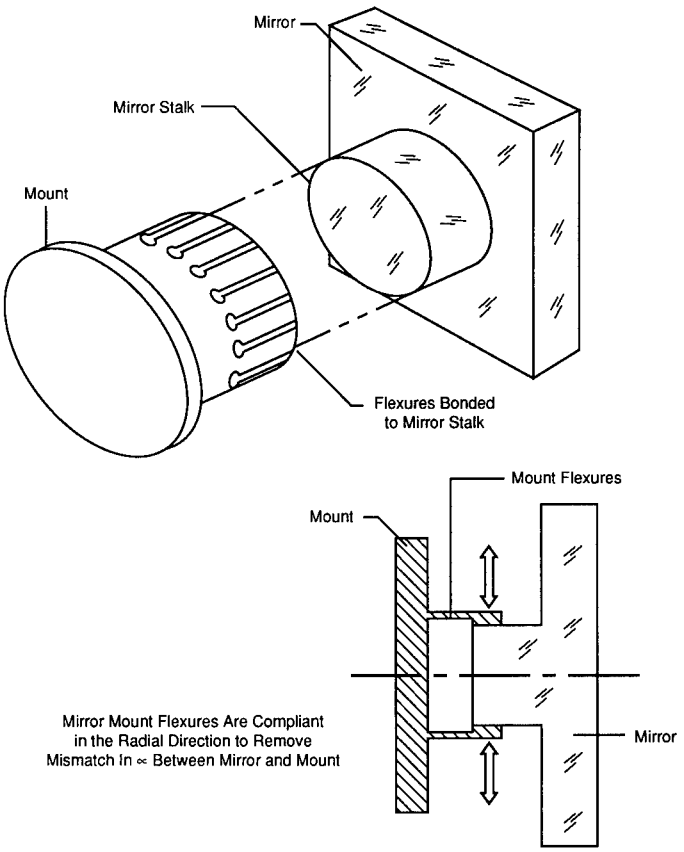


Fig. 3.22 Mushroom mirror mount.

The mushroom mount principle is turned inside out in the center hub mount. The center hub mount uses a hole bored in the center of the mirror. For a Cassegrain primary mirror, this hole would extend through the mirror. For other types of mirrors, the hole would extend close to the surface, but would not break through. The depth of the hole in the mirror back must be sufficient to reach the plane of the center of gravity of the mirror. Normally, the bond area is limited to a line around the circumference of the inside of the hole, and at the location of the center of gravity.

Another bonded mount uses three bosses around the circumference of the mirror. The mirror is bonded to its mount via adhesive placed on the three bosses. To isolate the mirror from expansion or contraction of the mirror mount, each boss is connected to the mirror mount by a flexure. The flexure is stiff in the axial and tangential directions, and is compliant in the radial direction.

3.7 ATHERMALIZATION

3.7.1 Structural Athermalization

Changes in temperature can cause the spacing between optical elements of a system to change. This change can create a change in focus, or disturb the

alignment of the system. The optical support structure is said to be *athermalized* if it compensates for these temperature effects.

Often overlooked in evaluating temperature effects is the effect of the atmosphere inside or near the system. The index of refraction of air varies with temperature. If there is a temperature change along the path, optical aberration is introduced into the beam. The index of refraction of air at a constant pressure of 760 mm Hg is given by[67]:

$$n_T - 1 = (n_{15} - 1)\frac{1.0549}{1 + 0.00366\,T_a}\,, \tag{3.102}$$

$$(n_{15} - 1) \times 10^8 = 8{,}342.1 + \frac{2{,}406{,}030}{130 - 1/\lambda^2} + \frac{15{,}996}{38.9 - 1/\lambda^2}\,, \tag{3.103}$$

where

n_T = refractive index of air at a temperature T
n_{15} = refractive index of air at 15°C
T_a = air temperature (degree centigrade)
λ = wavelength (micrometers).

If possible, the air inside and near an optical instrument should be maintained in a uniform temperature condition. Attempts to make the air more homogeneous by stirring with fans are usually not successful. Placing a close-fitting thermal-insulated cover around the optical beam path normally reduces air temperature variation. Heat sources should be located so that heat does not rise into the beam. Very sensitive optical instruments can be isolated from air-temperature variation effects by operating in a helium atmosphere.[68] In extreme cases, the optical system may need to be operated in a vacuum.[69]

3.7.2 Same Material Athermalization

When both structure and optics are made of the same material, an athermal structure is produced. A uniform change in temperature of a reflective optic changes the focal length and the physical size of the mirror. If the mirror is mounted using a material that is the same as the mirror, the same change in temperature produces a corresponding change in physical size in the mirror mount. This corresponding size change in the mirror mount eliminates mounting stress due to the temperature change.

Change in focus with temperature is proportional to the thermal coefficient of expansion of the mirror. If the focus is located using a structure made of the same type of material as the mirror, the length of the structure changes with temperature in an amount corresponding to the change in focus. Focus is maintained if both mirror and structure are made of the same type of material.

This same-material athermalization principle is employed in all metal optical systems. Common metal materials used for same-material athermalization include aluminum and beryllium. An all-glass Cassegrain telescope can employ the same principle. In this case, the secondary mirror is mounted on a window made of the same type of glass as the primary mirror.

Same-material athermalized systems are commonly employed in IR applications. Performance of this type of athermalized system is limited by vul-

nerability to temperature gradients. A gradient along or perpendicular to the optical axis will disturb the performance of a same-material athermalized system.

3.7.3 Athermal Compensators

The index of refraction of transmissive optical materials changes with temperature. The change in refractive index of a material with temperature is called the *thermo-optic coefficient* of the material. This change in refractive index of material with temperature causes the focal length of a lens to change with temperature. Lens power is defined as the reciprocal of the lens focal length. The change in lens power with temperature is then given by[70]:

$$\frac{dK}{dT} = K(-\beta) = \alpha_G - \frac{1}{(n_G - n_{\text{AIR}})}\left(\frac{dn_G}{dT} - n_G\frac{dn_{\text{AIR}}}{dT}\right), \tag{3.104}$$

where

dk/dT = change in lens power with temperature
K = lens power
β = optothermal expansion coefficient
α_G = thermal coefficient of the lens material
dn_G/dT = thermo-optic coefficient of the lens material
n_G = refractive index of the lens material
dn_{AIR}/dT = change in the refractive index of air with temperature.

Neglecting air temperature effects, an approximate version of Eq. (3.104) is given by:

$$\beta \cong \alpha_G - \frac{1}{(n_G - 1)}\frac{dn_G}{dT}. \tag{3.105}$$

Table 3.9 gives the optothermal expansion coefficient of some typical lens materials.

Table 3.9 Thermo-Optic Coefficients for Optical Materials

Material	n	λ (nm)	dn/dt (10^{-6}/K)
PK51	1.53019	546.1	−8.5
FK3	1.46619	546.1	−0.1
BK7	1.51872	546.1	3.0
LaK10	1.72340	546.1	5.0
SF5	1.67764	546.1	5.8
SF6	1.81265	546.1	11.6
Fused silica	1.45850	587.6	8.1
CVD ZnSe	2.473	1150	59.7
CVD ZnS	2.279	1150	49.8
Silicon	2.38	10 μm	162.0
KRS-5	2.37	10 μm	−235.0
Germanium	4.003	10 μm	396.0

For a system of thin lenses, the change in system power with temperature is given by:

$$\frac{dK_s}{dT} = K_s(-\beta_s) = K_s\left(-\sum_{i=1}^{n}\frac{K_i}{K_s}\beta_i\right) , \tag{3.106}$$

where

dK_s/dT = change in system power with temperature
β_s = optothermal coefficient of expansion of the system
K_s = system power
K_i = individual element power
β_i = optothermal coefficient of expansion of the individual element
n = number of elements.

For passive maintenance of focus with temperature, the thermal coefficient of expansion of the structure must equal the optothermal coefficient of expansion. This requirement is difficult to meet with a single type of material. Most optical systems do not have optothermal coefficients of expansion that match the thermal coefficients of expansion of common structural materials. In some cases, the optothermal expansion coefficient of the system is negative.

Virtually any thermal coefficient of expansion, including negative values, can be achieved in a structure by using a pair of materials. A combination of a high and low thermal coefficient of expansion material is used in series to produce the desired thermal coefficient of expansion. By adjusting the lengths of the two materials, the thermal coefficient of expansion of the structure is adjusted. For passive maintenance of focus with temperature, the following equations are used to calculate the material lengths:

$$\beta_s f = \alpha_1 L_1 + \alpha_2 L_2 , \tag{3.107}$$

$$L_2 = \frac{f(\beta_s/\alpha_1 - 1)}{(\alpha_2/\alpha_1 - 1)} , \tag{3.108}$$

$$L_1 = f - L_2 , \tag{3.109}$$

where

f = system focal length
L_2 = length of the high thermal coefficient of expansion material
α_2 = thermal coefficient of expansion of the high thermal coefficient of expansion material
L_1 = length of the low thermal coefficient of expansion material
α_1 = thermal coefficient of expansion of the low thermal coefficient of expansion material.

Table 3.10 gives representative combinations of materials to match the optothermal coefficients of expansion of some optical materials. Figure 3.23 shows two sample bimetallic compensator configurations.

Table 3.10 Some Bimetallic Compensators for Single Elements

Glass Type	β_s (m/m K $\times 10^{-6}$)	Material 1	α_1 (m/m K $\times 10^{-6}$)	$L_1 \times f$	Material 2	α_2 (m/m K $\times 10^{-6}$)	$L_2 \times f$
TiF6	20.94	Aluminum	23	0.678	Stainless steel	16.6	0.322
BK1	3.28	Invar	0.54	0.829	Stainless steel	16.6	0.171
LaKN9	0.32	Invar	0.54	1.01	Aluminum	23	−0.01
BAK4	−0.23	Invar	0.54	1.034	Aluminum	23	−0.034
KzFS1	−2.89	Invar	0.54	1.153	Aluminum	23	−0.153
ZnSe	−28.24	Stainless steel	16.6	1.233	Plastic (ABS) (polyurethane)	2099	−0.233
Silicon	−64.10	Stainless steel	16.6	1.419	Plastic (ABS) (polyurethane)	209	−0.419
Germanium	−85.19	Stainless steel	16.6	1.358	Plastic (polyethylene)	301	−0.358

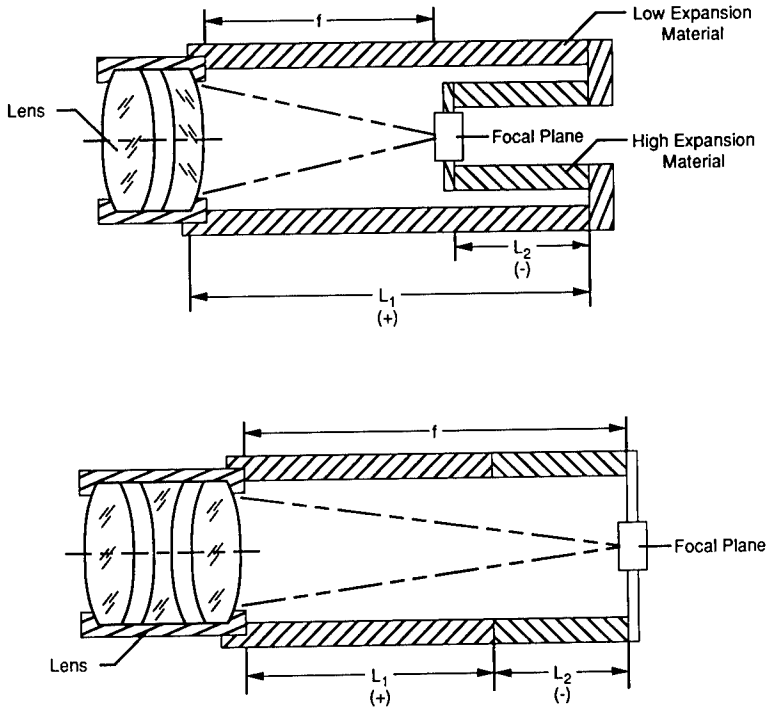


Fig. 3.23 Athermal bimetallic compensators.

Mirror systems, and systems with small optothermal coefficients of expansion, require structures made of low thermal coefficient of expansion materials for effective athermalization. Typical low thermal coefficient of expansion materials used for this purpose include Invar, Super Invar, Corning Code 7971 ULE™, Schott Zerodur™, and the graphite epoxy composites. All of these materials present special problems to the designer.

Invar and Super Invar are heavy, expensive, difficult to machine, and have controversial long-term stability. These disadvantages limit all-Invar structures to relatively small sizes. The disadvantages of Invar materials are avoided by using this type of material in a metering structure. In a metering structure, the optical elements are supported by a structure made of conventional material. The optical mounts are attached to the structure via a compliant connection. Motion of each mount along the optical axis of the system is possible.[71] All mounts are tied together by Invar rods. The Invar rods maintain correct spacing and, therefore, preserve system focus as the support structure expands or contracts. The same approach is used with fused silica or zero thermal coefficient of expansion glass ceramic rods. A metering rod structure is shown in Fig. 3.24.

Graphite epoxy composites have considerable promise for athermalizing structures.[72] The thermal coefficient of expansion of a graphite epoxy composite material can be tailored to virtually any value, at least in one direction. Graphite composite structures have exceptional stiffness-to-weight ratios. A serious

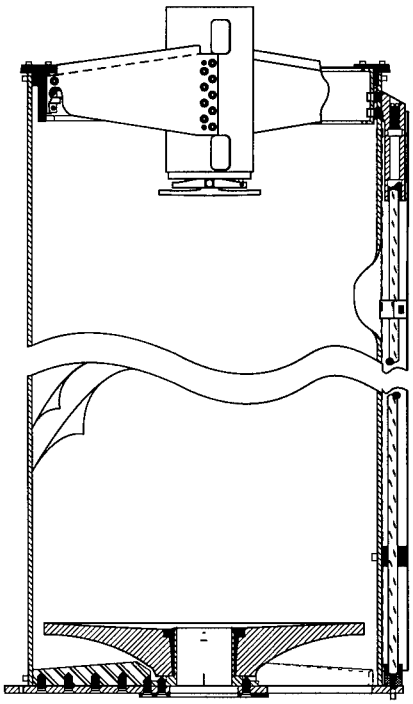


Fig. 3.24 Metering rods maintain focus in an athermalized Cassegrain telescope at the Space Dynamics Laboratory, Utah State University.

disadvantage of the graphite epoxy composite materials is dimensional instability caused by the absorption of water. A change in humidity can cause a graphite epoxy structure to change in size enough to affect the performance of the optical system.[73] Humidity absorption in graphite epoxy structures is reduced by cladding the structure with a flexible hermetically sealed membrane. Damage to the membrane in service and residual water inside the membrane are two problems of this approach.

3.8 EXAMPLE PROBLEMS

3.8.1 Window Design

A 500-mm-diam circular ZnSe window is used on an aircraft flying at an altitude of 12 km. Instrument bay pressure and temperature are 27 KPa and 300 K, respectively. Ambient conditions are a pressure of 19 KPa and a temperature of 217 K. A window failure probability of 10^{-4} with a safety factor of 4 is acceptable. Calculate the safe window thickness and the deflections due to pressure and temperature differentials.

Solution. Failure strength is given using Weibull statistics. Solving Eq. (3.19) for stress, and using the lowest values for ZnSe in Table 3.7:

$$\sigma_a = \sigma_0[-\ln(1 - P_F)]^{1/m}$$
$$= 23 \times 10^6[-\ln(1 - 10^{-4})]^{1/6}$$
$$= 7.27 \times 10^6 \text{ Pa} .$$

Using conservative practice, assuming the window to be simply supported, Eq. (3.6) gives the minimum window thickness:

$$\frac{d}{h} = 2\left[\frac{8\sigma_F}{3\Delta P(3 + \nu)\text{SF}}\right]^{1/2}$$
$$= 2\left[\frac{8(7.27 \times 10^6)}{3(27 \times 10^3 - 19 \times 10^3)(3 + 0.28)4}\right]^{1/2}$$
$$= 27.2 ,$$
$$h = \frac{0.5}{27.2} = 0.018 \text{ m} .$$

From the properties given in Tables 3.5 and 3.9, and using Eq. (3.2), the thermal deformation of the window due to the axial temperature gradient is:

$$q = \frac{\Delta TK}{h} = \frac{(300 - 217)(18)}{0.018} = 83 \times 10^3 \text{ W/m}^2 ,$$
$$\frac{1}{f} = \frac{n - 1}{n}\left(\frac{\alpha}{K}\right)^2 (0.018)(83 \times 10^3)^2 = 12.4 \times 10^{-6} \text{ m}^{-1} .$$

The pressure-induced distortion is found using Eq. (3.4):

$$\text{OPD} = 8.89 \times 10^{-3}\frac{(n - 1)\Delta P^2 d^6}{E^2 h^5}$$
$$= \frac{(8.89 \times 10^{-3})(2.473 - 1)(8 \times 10^3)^2(-5)^6}{(71.9 \times 10^9)^2(0.018)^5}$$
$$= 1.37 \times 10^{-9} \text{ m} .$$

This example shows that window deformation due to pressure and temperature differentials is usually not a problem in most common service environments.

3.8.2 Dome Design

A 50-mm-radius ZnS dome is used as a streamlined window on a mortar shell. The dome is 2.4 mm thick. Firing acceleration is $11 \times 10^3 g$. What are the stress in the dome and the failure probability at that stress?

Solution. Because the ratio of dome radius to thickness exceeds 10(50/2.4 = 20.8), Eqs. (3.12) and (3.13) are used to find the stress. The equivalent external pressure is found using Eq. (3.14):

$$\Delta P = \frac{a}{g}\rho h = (11 \times 10^3)(4.1 \times 10^3)(9.81)(0.0024) = 1.06 \times 10^6 \text{ Pa} .$$

The maximum stresses are:

$$\sigma_m = \Delta P\left(\frac{R}{h}\right) = (1.06 \times 10^6)\left(\frac{50 \times 10^{-3}}{2.4 \times 10^{-3}}\right) = 22.1 \times 10^6 \text{ Pa} .$$

Using Eq (3.19), the failure probability is:

$$P_F = 1 - \exp\left[-\left(\frac{\sigma_a}{\sigma_0}\right)^m\right] = 1 - \exp\left[-\left(\frac{22.1 \times 10^6}{29 \times 10^6}\right)^9\right] = 0.08 .$$

3.8.3 Prism Mounting

A fused silica right triangular prism, 25 mm thick and 25 mm on each face, is adhesive bonded to a mount. Prism acceleration is 10 *g*. Using an adhesive with a bond strength of 14 MPa, and a safety factor of 4, what bond area is required?

Solution. Prism weight is given by:

$$W = \rho hA = \frac{1}{2}\rho hbL$$

$$= \frac{1}{2}(2.20 \times 10^3)(0.025)(0.035)(0.018) = 17.3 \times 10^3 \text{ kg} .$$

Using Eq. (3.35),

$$S = \frac{aWSF}{\sigma} = \frac{(10)(17.3 \times 10^{-3})4(9.81)}{(9.81)(14 \times 10^6)} = 49.42 \times 10^{-6} \text{ m}^2 .$$

If a circular bond area is used, its diameter will be:

$$D = \left(\frac{4S}{\pi}\right)^{1/2} = \left[\frac{4(49.42 \times 10^{-6})}{\pi}\right]^{1/2} = 7.932 \times 10^{-3} \text{ m} .$$

This diameter is approximately one-third the face length.

3.8.4 Bonded Lens Cell

A double convex BK7 lens is bonded into an aluminum lens cell. The lens is 50 mm in diameter, with a 20-mm central thickness. Radii of curvature are

0.15 and 0.10 m. Using an elastomer adhesive with properties given below, determine the optimum adhesive thickness required to athermalize the lens and the self-weight-induced radial deflection of the lens.

Solution. First find the lens edge thickness using Eq. (3.41):

$$h_e = h_T - h_{s1}h_{s2}$$

$$= h_t - \left[R_1 - \left(R_1^2 - \frac{D^2}{4}\right)^{1/2}\right] - \left[R_2 - \left(R_2^2 - \frac{D^2}{4}\right)^{1/2}\right]$$

$$= 0.020 - \{(0.15)^2 - [(0.15)^2 - (0.025)^2]^{1/2}\} - \{(0.10)^2 - [(0.025)^2]^{1/2}\}$$

$$= 0.020 - 0.0021 - 0.0032 = 0.0147 \text{ m} .$$

Then find the lens weight using Eqs. (3.37) and (3.42):

$$W = \frac{\pi}{4}\rho h_e D^2 + \pi\rho h_{s1}^2\left(R_1 - \frac{h_{s1}}{3}\right) + \pi\rho h_{s2}^2\left(R - 2 - \frac{h_{s2}}{3}\right)$$

$$= \frac{\pi}{4}(2.53 \times 10^3)(0.014)(0.050)^2$$

$$+ \pi(2.53 \times 10^3)(0.0021)^2\left(0.15 - \frac{0.0021}{3}\right)$$

$$+ \pi(2.53 \times 10^3)(0.0032)^2\left(0.10 - \frac{0.0032}{3}\right)$$

$$= 86.31 \times 10^{-3} \text{ kg} .$$

The adhesive properties are:

$\alpha = 200 \times 10^6$ m/m K
$E = 160$ MPa
$G = 56$ MPa
$\nu = 0.43$.

The optimum bond thickness is given by Eq. (3.48):

$$h_r = \frac{D_g(\alpha_m - \alpha_g)}{2(\alpha_r - \alpha_m)}$$

$$= \frac{0.05(23 \times 10^{-6} - 7.1 \times 10^{-6})}{2(200 \times 10^{-6} - 23 \times 10^{-6})} = 2.246 \times 10^{-3} \text{ m} .$$

The self-weight deflection is given by Eq. (3.49):

$$\delta_L = \frac{W_L}{(\pi/2)d_G(h_G/h_r)[E_r/(1 - \nu_r^2) + G_r]}$$

$$= \frac{(86.31 \times 10^{-3})(9.81)}{(\pi/2)(0.05)(0.0147/0.00246)\{(160 \times 10^6)/[1 - (-43)^2] + 56 \times 10^6\}}$$

$$= 7.151 \times 10^{-9} \text{ m} .$$

This example shows that the adhesive thickness required to athermalize the lens is relatively large. However, this adhesive bond is relatively stiff.

3.8.5 Mechanical Lens Mount

The BK7 lens from the example in Sec. 3.8.4 is mounted in a mechanical seat and retainer cell. The lens is to be used at a low temperature of 220 K; assembly is at 300 K. The lens cell is aluminum, with a 5-mm radial wall thickness. A sharp-edged seat and retainer are used. Initial assembly preload is equal to 10 times the lens weight. Initial radial clearance between lens and cell is 2.5 μm. Find the preload and temperature stresses on the lens. Retainer and seat have a 0.04-m common diameter and a corner radius of 50 μm.

Solution. According to Eq. (3.53), the maximum stress is independent of lens radius if the lens radius is much greater than the contact radius, so a single-stress calculation will suffice:

$$\sigma = 0.798\left\{\frac{(F/2\pi r)[(d_G + d_m)/d_G d_m]}{(1 - \nu_G^2)/E_G + (1 - \nu_m^2)/E_m}\right\}^{1/2}$$

$$= 0.798\left(\frac{[(847 \times 10^{-3})/2\pi(0.02)]\{(0.10 + 0.01)/[(0.10)(0.01)]\}}{[1 - (0.2)^2]/(80.7 \times 10^9) + [1 - (0.3)^2]/(69 \times 10^9)}\right)^{1/2}$$

$$= 4.34 \times 10^6 \text{ Pa} .$$

The radial stress caused by the temperature drop is given by Eqs. (3.56) and (3.57):

$$K_R = 1 - \frac{h_r}{r_G \Delta T(\alpha_m - \alpha_g)}$$

$$= 1 - \frac{2.5 \times 10^6}{(0.02)(80)(23 \times 10^{-6} - 7.1 \times 10^{-6})} = 0.902 ,$$

$$\sigma_r = \frac{K_r \Delta T(\alpha_m - \alpha_g)}{1/E_G + r_G/E_m h_m} - \frac{(0.902)(980)(23 \times 10^{-6} - 7.1 \times 10^{-6})}{1/(80.7 \times 10^9) + 0.02/[(69 \times 10^9)(0.005)]}$$

$$= 16.3 \times 10^6 \text{ Pa} .$$

This example illustrates the very high stresses created by assembly preload forces and thermal effects.

3.8.6 Small Circular Mirror Mount

A 0.3-m-diam *f*/5 aluminum mirror is mounted at a 45-deg angle with respect to vertical. The mirror is a 50-mm thick right circular cylinder. An optimum

five-point (three back, two edge) kinematic mount is used. What is the self-weight deflection of the mirror?

Solution. From Eq. (3.61), the self-weight deflection is the sum of the radial and axial deflection components. The optimum location for the three-axial supports is at 0.645(0.3) = 0.194 m. Equation (3.63) gives the axial deflection:

$$\delta_A = 0.318\frac{\rho}{E}\frac{r^4}{h^2}(1 - \nu) = 0.318\,\frac{(2.71 \times 10^3)(9.81)}{69 \times 10^9}\,\frac{(0.15)^4}{(0.05)^2}$$

$$\times [1 - (0.3)^2] = 22.6 \times 10^{-9}\ \text{m} .$$

The optimum location for the radial supports is at two points 90 deg apart. Equations (3.69) and (3.70), with data from Table 3.6, are used to find the deflection. The mirror axial thickness is given by:

$$h = h_e - [R - (R^2 - r^2)^{1/2}] = 0.05 - 3 + [(3)^2 - (0.15)^2]^{1/2} = 0.046\ \text{m} ,$$

$$\gamma = \frac{r^2}{2hR} = \frac{(0.15)^2}{2(0.046)(3)} = 0.081 ,$$

$$\delta_r = (a_0 + a_1\gamma + a_2\gamma^2)\frac{2\rho r^2}{E}$$

$$+ [0.05466 + (0.2786)(0.081) + (0.1100)(0.081)^2]$$

$$\times \frac{2(2.71 \times 10^3)(9.81)(0.15)^2}{69 \times 10^9} = 1.35 \times 10^{-9}\ \text{m} .$$

Then using Eq. (3.61),

$$\delta_m = [(\delta_A \cos\theta)^2 + (\delta_r \sin\theta)^2]^{1/2} + \{[(922.6 \times 10^{-9})\cos 45\ \text{deg}]^2$$

$$+ [(1.35 \times 10^{-9})\sin 45\ \text{deg}]^2\}^{1/2} + 16.0 \times 10^{-9}\ \text{m} .$$

This example shows that the radial deflection component of a small mirror is very small, and can be neglected in many applications.

3.8.7 Lightweight Mirror Design

The aluminum mirror in Sec. 3.8.6 is replaced with a beryllium sandwich mirror. The beryllium sandwich mirror faceplate thickness is 3 mm. Square cells are used in the core, with a cell length of 50 mm and a cell-wall thickness of 1.5 mm. The beryllium mirror is 50 mm thick. Compare the weight and stiffness of the beryllium and aluminum mirrors.

Solution. The weight of a lightweight sandwich mirror is given by Eq. (3.80). First calculate the rib solidity ratio using Eq. (3.74):

$$\eta = \frac{(2\beta + t_c)t_c}{(\beta + t_c)^2} = \frac{[2(0.05) + 0.0015]0.0015}{(0.05 + 0.0015)^2} = 0.0574 \ .$$

The core height is

$$h_c = h - 2h_f = 0.05 - 2(0.003) = 0.044 \text{ m} \ .$$

Then the weight is:

$$W = \frac{\pi}{4}\rho d^2(2h_f + \eta h_c)$$

$$= \frac{\pi}{4}(1.85 \times 10^3)(0.3)^2[2(0.003) + (0.0574)(0.044)] = 1.11 \text{ kg} \ .$$

The weight of the aluminum mirror is:

$$W = \frac{\rho r}{4}\rho h d^2 = \frac{\pi}{4}(2.71 \times 10^3)(0.005)(0.3)^2 = 9.58 \text{ kg} \ .$$

The ratio of the weight is:

$$\frac{W_{\text{BE}}}{W_{\text{AL}}} = \frac{1.11}{9.58} = 0.12 \ .$$

From Eq. (3.62), self-weight bending (stiffness to weight) is controlled by the ratio of the mirror area density to flexural rigidity. Area density is given by dividing mirror weight by mirror area:

$$A_W = \frac{w}{(\pi/4)d^2} = \frac{1.11}{(\pi/4)(0.3)^2} = 15.8 \text{ kg/m}^2 \ .$$

Flexural rigidity of a sandwich mirror is given by Eqs. (3.78) and (3.79):

$$h_B^3 = (2h_f + h_c)^3 - \left(1 - \frac{\eta}{2}\right)h_c^3 = [2(0.003) + 0.044]^3$$

$$- \left[1 - \frac{0.0574}{2}\right](0.044)^3 = 42.26 \times 10^{-6} \text{ m}^3 \ .$$

$$D = \frac{Eh_B^3}{12(1 - \nu^2)} = \frac{(304 \times 10^9)(42.26 \times 10^{-6})}{12[1 - (0.017)^2]} = 1.071 \times 10^6 \text{ Pa m}^3 \ .$$

Comparing the aluminum and beryllium mirrors:

$$\frac{[(\rho/E)(1/h^2)]_{AL}}{(A_W/D)_{BE}} = \frac{[(2.71 \times 10^3)(9.81)]/[(69 \times 10^9)(0.05)^2]}{15.8/(1.071 \times 10^6)}$$

$$= \frac{154 \times 10^{-6}}{14.8 \times 10^{-6}} = 10.4 \ .$$

So the aluminum mirror deflects 10.4 times more than the beryllium mirror.

3.8.8 Thermal Effects on Mirrors

A right circular cylinder (10-to-1 diameter-to-thickness ratio) 0.5-m-diam *f*/2 concave mirror is used in a cryogenic IR telescope. The mirror is cooled from 300 to 100 K. There is a temperature gradient of 2 K/m along the optical axis, and gradient of 1 K/m in the plane of the mirror. What are the thermal deformations of the mirror?

Solution. From the problem statement and Tables 3.5 and 3.8:

$$R = 2fD = 2(2)(0.5) = 2 \text{ m} \ ,$$

$$r = \frac{D}{2} = \frac{0.5}{2} = 0.25 \text{ m} \ ,$$

$$h = \frac{d}{10} = \frac{0.5}{10} = 0.50 \text{ m} \ ,$$

where

$$\alpha = 560 \times 10^{-9} \text{ m/m K}$$
$$\Delta\alpha = 2 \times 10^{-9} \text{ (m K)}^{-1}$$
$$C_0 = 300 \text{ K} - 100 \text{ K} = 200 \text{ K}$$
$$C_2 = 1 \text{ K/m}$$
$$C_3 = 2 \text{ K/m} \ .$$

Using Eq. (3.84), the following aberration terms are computed:

$$\text{Spherical: } \delta = \frac{\alpha C_3}{8R^2} r^4 = \frac{(560 \times 10^{-9})(2)(0.25)^4}{8(2)^2} = 137 \times 10^{-12} \text{ m} \ ,$$

$$\text{Coma: } \quad \delta = \frac{\alpha C_1}{2R} r^3 \cos\theta + \frac{\alpha C_2}{2R} r^3 \sin\theta$$

$$= \frac{(560 \times 10^{-9})(1)(0.25)^3}{(2)(2)} = 2.19 \times 10^{-9} \text{ m} \ ,$$

$$\text{Focus: } \quad \delta = \left[\frac{\alpha C_3 h}{2R} - \frac{\alpha C_3}{2} = \frac{\alpha C_o}{2R}\right] r^2$$

$$= (560 \times 10^{-9})(0.25)^2 \left[\frac{2(0.05)}{2(2)} - \frac{2}{2} + \frac{200}{2C_2}\right]$$

$$= 1.72 \times 10^{-6} \text{ m} \ .$$

Using Eq. (3.88), the aberration due to an axial variation in thermal coefficient of expansion is computed:

$$\delta \simeq \frac{r^2}{2h}\Delta\alpha\Delta T = \frac{(0.25)^2(2 \times 10^{-9})(0.05)(200)}{2(0.05)} = 12.5 \times 10^{-9}\ \text{m} .$$

Using Eq. (3.89), the aberration due to transverse variation in thermal coefficient of expansion is computed:

$$\delta \simeq \frac{r^2}{4h}(\alpha_x - \alpha_y)\Delta T$$

$$= \frac{(0.25)^2}{4}(0.05)(2 \times 10^{-9})(0.5)(200) = 62.5 \times 10^{-9}\ \text{m} .$$

This example shows that the greatest effect of a temperature change on a mirror is a shift in focus. In this example, spatial variation in thermal coefficient of expansion has a stronger effect than the thermal gradients. This is of interest since fused silica has a low thermal coefficient of expansion and good uniformity of thermal coefficient of expansion.

3.8.9 Thermal Bimetallic Bending of a Metal Mirror

The mirror in Sec. 3.8.8 is plated with electroless nickel. The plating thickness on the mirror's optical surface is 75 μm, and plating thickness on the back of the mirror is 125 μm. What is the change in radius and mirror stress due to bimetallic bending when the mirror temperature is reduced from 300 to 100 K?

Solution. The change in mirror radius caused by bimetallic bending is given by Eqs. (3.93), (3.94), and (3.90):

$$\psi_m = \frac{E_m}{1 - \nu_m} = \frac{69 \times 10^{-9}}{1 - 0.3} = 98.6 \times 10^9\ \text{Pa} ,$$

$$\psi_p = 196 \times 10^9\ \text{(data from Ref. 1)} ,$$

$$\frac{1}{R_o} - \frac{1}{R} = \frac{6}{h}\frac{\psi_p}{\psi_m}(\alpha_p - \alpha_m)\Delta T\left(\frac{h_{p1} - h_{p2}}{h}\right)\left[1 - \left(\frac{h_{p1} - h_{p2}}{h}\right)\left(4\frac{\psi_p}{\psi_m} - 1\right)\right] ,$$

$$\frac{1}{R} = \frac{1}{2} - \frac{6}{0.05}\frac{196 \times 10^9}{98.6 \times 10^9}(13 \times 0^{-6} - 23 \times 10^{-6})(100)$$

$$\times \left(\frac{75 \times 10^{-6} - 125 \times 10^{-6}}{0.05}\right)$$

$$\times \left[1 - \left(\frac{75 \times 10^{-6} - 125 \times 10^{-6}}{0.05}\right)\left(4\frac{196 \times 10^9}{98.6 \times 10^9} - 1\right)\right] ,$$

$$= 0.5 - 240 \times 10^{-6} = 499.7598 \times 10^{-3} ,$$

$$\Delta R = R_o - R = 2 - (499.7598 \times 10^{-3})^{-1} = 961 \times 10^{-6}\ \text{m} .$$

The mirror stress is given by Eq. (3.91):

$$\sigma_{rm} = \psi_p(\alpha_p - \alpha_m)\Delta T\left[3\frac{(h_{p2} - h_{p1})}{h} + \frac{(h_{p2} + h_{p1})}{h}\right]$$

$$= 196 \times 10^9(13 \times 10^{-6} - 23 \times 10^{-6})(100)$$

$$\times \left[3\frac{(125 \times 10^{-6} - 75 \times 10^{-6})}{0.05} + \frac{(125 \times 10^{-6} + 75 \times 10^{-6})}{0.05}\right]$$

$$= -196 \times 10^3 \text{ Pa} .$$

This example shows that the change in radius due to bimetallic bending is significant, at about 1 mm. The bimetallic bending stress is well below the 65-MPa microyield of aluminum.

References

1. W. P. Barnes, Jr., "Some effects of aerospace thermal environments on high-acuity optical systems," *Applied Optics* **5**(5), 701–711 (1966).
2. C. A. Klein, B. diBenedetto, and J. Pappis, "ZnS, ZnSe, and ZnS/ZnSe windows: their impact on FLIR system performance," *Optical Engineering* **25**(4), 519–531 (1986).
3. J. D. Stachiw and W. Loucks, "Design parameters for germanium windows under uniform pressure loading," *Proceedings of the SPIE* **131,** 57–72 (1978).
4. G. M. Dunn and J. D. Stachiw, "Acrylic windows for underwater structures," *Proceedings of the SPIE* **7,** D-XX-1–D-XX-2 (Oct. 1966).
5. J. Speare and A. Belloli, "Structural mechanics of a mortar launched IR dome," *Proceedings of the SPIE* **450,** 182–187 (1983).
6. R. H. Doremus, "Fracture statistics: a comparison of the normal, Weibull, and type 1 extreme value distributions," *Journal of Applied Physics* **54,** 193–198 (1983).
7. J. E. Ritter, Jr., "Assessment of reliability of ceramic materials," in *Fracture Mechanics of Ceramics,* Vol. 5, R. C. Bradt et al., Eds., Plenum Press, NY (1983).
8. E. B. Shand, "Fracture velocity and fracture energy of glass in the fatigue range," *Journal of the American Ceramic Society* **44,** 21–26 (1961).
9. S. M. Wiederhorn, *Reliability, Life Prediction and Proof Testing of Ceramics,* NBSIR 74-486, National Bureau of Standards (1974).
10. G. Sines, "Rationalized crack growth and time-to-fracture of brittle materials," *Journal of the American Ceramic Society* **59,** 370–371 (1976).
11. S. M. Wiederhorn, A. G. Evans, and D. E. Roberts, "A fracture mechanics study of the Skylab windows," in *Fracture Mechanics of Ceramics,* Vol. 2, R. C. Bradt, D. P. H. Hasselman, and F. F. Lange, Eds., Plenum Press, NY (1974).
12. A. G. Evans and E. R. Fuller, "Crack propagation in ceramic materials under cyclic loading conditions," *Metallurgical Transactions* **5,** 27–33 (1974).
13. R. D. Blevins, *Formulas for Natural Frequency and Mode Shape,* Van Nostrand Reinhold, NY (1979).
14. P. R. Yoder, Jr., "Design guidelines for bonding prisms to mounts," *Proceedings of the SPIE* **1013,** 112–117 (1989).
15. D. S. L. Durie, "Stability of optical mounts," *Machine Design* **40,** 184–190 (1968).
16. M. L. Lipshutz, "Optomechanical considerations for optical beam splitters," *Applied Optics* **7,** 2326–2328 (1968).
17. P. R. Yoder, Jr., *Opto-Mechanical Systems Design,* Marcel Dekker, NY (1986).
18. P. Kuttner, "Trends and limits in the production of optical elements and optical systems," *Proceedings of the SPIE* **163,** 24–30 (1979).
19. R. E. Parks, "Optical component specifications," *Proceedings of the SPIE* **237,** 455–463 (1980).

20. A. Zaltz and D. Christo, "Methods for the control of centering error in the fabrication and assembly of optical elements," *Proceedings of the SPIE* **330,** 39–48 (1982).
21. K. W. Hildebrand, "ISO standards for optical drawings," *Proceedings of the SPIE* **1015,** 151–160 (1988).
22. F. Schubert, "Center of gravity of lenses," *Machine Design,* pp. 111–112 (April 4, 1974).
23. K. H. Spring, "Bonding optical components to metal mounts," *British Journal of Applied Physics* **9,** 242–246 (1958).
24. K. H. Carnell, M. J. Kidger, A. J. Overill, R. W. Reader, F. C. Reavell, W. T. Welford, and C. G. Wynne, "Some experiments on precision lens centering and mounting," *Optica Acta* **21,** 615–629 (1974).
25. M. Bayar, "Lens barrel optomechanical design principles," *Optical Engineering* **20,** 181–186 (1981).
26. T. M. Valente and R. M. Richard, "Analysis of elastomer lens mountings," *Proceedings of the SPIE* **1533,** 21–26 (1991).
27. B. J. Kowalskie, *A User's Guide to Designing and Mounting Lenses and Mirrors,* UCRL-52411, Lawrence Livermore Laboratory (1978).
28. W. Zschommler, *Precision Optical Glassworking,* SPIE, Optical Engineering Press, Bellingham, WA (1984).
29. J. E. Shigley, *Mechanical Engineering Design,* 3rd ed., McGraw-Hill, NY (1977).
30. R. J. Roark and W. C. Young, *Formulas for Stress and Strain,* 5th ed., McGraw-Hill, NY (1975).
31. R. F. Delgado and M. Hallinan, "Mounting of lens elements," *Optical Engineering* **14**(1), S-11 (1975).
32. T. M. Valente and R. M. Richard, "Interference fit equations for lens cell design," *Proceedings of the SPIE* **1533,** 12–20 (1991).
33. J. W. Pepi, "Analytical predictions for lightweight optics in a gravitational and thermal environment," *Proceedings of the SPIE* **748,** 172–179 (1987).
34. R. Williams and H. F. Brinson, "Circular plate on multipoint supports," *Journal of the Franklin Institute* **297**(6), 429–447 (1974).
35. J. E. Nelson, J. Lubliner, and T. S. Mast, "Telescope mirror supports: plate deflections on point supports," *Proceedings of the SPIE* **332,** 212–228 (1982).
36. W. J. Nowak, "A parametric approach to mirror natural frequency calculations," *Proceedings of the SPIE* **450,** 164–167 (1983).
37. H. D. Hall, "Problems in adapting small mirror fabrication techniques to large mirrors," in "Optical telescope technology workshop (April, 1969)," NASA Report SP-233 (1970).
38. D. L. Crawford, A. B. Meinel, and M. W. Stockton, Eds., *A Symposium on Support and Testing of Large Astronomical Mirrors,* Kitt Peak National Observatory, Tucson (1968).
39. G. Schwesinger, "Optical effect of flexure in vertically mounted precision mirrors," *Journal of the Optical Society of America* **44,** 417 (1954).
40. A. J. Malvick, "Theoretical elastic deformations of the Steward Observatory 230-cm and the Optical Sciences Center 154-cm mirrors," *Applied Optics* **11,** 575–585 (1972).
41. D. Vukobratovich and R. M. Richard, "Roller chain supports for large optics," *Proceedings of the SPIE* **1396,** 522–534 (1990).
42. T. M. Valente, "Scaling laws for light-weight optics," *Proceedings of the SPIE* **1340,** 47–66 (1990).
43. T. M. Valente and D. Vukobratovich, "Comparison of the merits of open-back, symmetric sandwich, and contoured back mirrors as lightweight optics," *Proceedings of the SPIE* **1167,** 20–36 (1989).
44. M. K. Cho, R. M. Richard, and D. Vukobratovich, "Optimum mirror shapes and supports for lightweight mirrors subjected to self weight," *Proceedings of the SPIE* **1167,** 2–19 (1989).
45. P. K. Mehta, "Flexural rigidity characteristics of lightweighted mirrors," *Proceedings of the SPIE* **748,** 158–171 (1987).
46. W. P. Barnes, Jr., "Basic properties of metal optics," *Optical Engineering* **16**(4), 320–323 (1977).
47. E. Benn and W. W. Walker, "Effect of microstructure on the dimensional stability of cast aluminum substrates," *Applied Optics* **12,** 976–978 (1973).
48. L. Noethe, F. Franza, P. Giordano, and R. Wilson, "Optical wavefront analysis of thermally

cycled 500 mm metallic mirrors," *Proceedings of the IAU Colloquium No. 79: Very Large Telescopes, Their Instrumentation and Programs,* Garching (April 9–12, 1984).

49. C. W. Marshall and R. E. Maringer, *Dimensional Instability,* Pergamon Press, Elmsford, NY (1977).
50. *Physics and Applications of Invar Alloys,* Honda Memorial Series on Material Sciences, Maruzen Co. Ltd., Tokyo (1978).
51. M. R. Meyerson, P. M. Giles, and P. F. Newfield, "Dimensional stability of gage block materials," *Journal of Materials,* JMSLA **3,** 727–743 (Dec. 1968).
52. F. C. Holden, *A Review of Dimensional Instability in Metals,* NTIS AD602379, Battelle Memorial Institute, Columbus, OH (1964).
53. C. W. Marshall, R. E. Maringer, and F. J. Cepollina, "Dimensional stability and micromechanical properties of materials for use in an orbiting astronomical observatory," AIAA Paper No. 72-325 (1972).
54. C. W. Marshall, M. E. Hoskins, and R. E. Maringer, "Continuation of a study of stability of structural materials for spacecraft applications for the orbiting astronomical observatory project," NASA NAS5-11195, Battelle Memorial Institute, Columbus, OH (1969).
55. C. W. Marshall and R. E. Maringer, "Stress relaxation as a source of dimensional instability," *Journal of Materials,* JMSLA **6,** 374–387 (1971).
56. M. C. Gerchman, "Specifications and manufacturing considerations of diamond machined optical components," *Proceedings of the SPIE* **607,** 36–45 (1986).
57. A. A. Ogloza, D. L. Decker, P. C. Archibald, D. A. O'Connor, and E. R. Bueltmann, "Optical properties and thermal stability of single-point diamond-machined aluminum alloys," *Proceedings of the SPIE* **966,** 228–251 (1988).
58. R. A. Paquin, "Hot isostatic pressed beryllium for large optics," *Proceedings of the SPIE* **571,** 259–266 (1985).
59. W. R. Mohn and D. Vukobratovich, "Recent applications of metal matrix composites in precision instruments and optical systems," *Optical Engineering* **27**(2), 90–98 (1988).
60. E. Pearson and L. Stepp, "Response of large optical mirrors to thermal distributions," *Proceedings of the SPIE* **748,** 215–228 (1987).
61. D. L. Hibbard, "Dimensional stability of electroless nickel coatings," *Proceedings of the SPIE* **1335,** 180–185 (1990).
62. B. Mesco, "Kinematic mount," U.S. Patent No. 4,268,123 (1981).
63. D. Vukobratovich and R. M. Richard, "Flexure mounts for high-resolution optical elements," *Proceedings of the SPIE* **959,** 18–36 (1988).
64. E. Hog, "A kinematic mounting," *Astronomy and Astrophysics* **41,** 107–109 (1975).
65. J. Espiard, J. Paseri, G. Cerutti-Maori, and C. Singer, "Lightweight cold mirror and fixation," *Proceedings of the SPIE* **589,** 187–193 (1985).
66. W. Horn, "The Aurora Project square mirror arrays," *Proceedings of the SPIE* **817,** 218–223 (1987).
67. R. J. Penndorf, "Tables of the refractive index for standard air and the Rayleigh scattering coefficient for the spectral region between 0.2 and 20.0 microns and their application to atmospheric optics," *Journal of the Optical Society of America* **47,** 176–182 (1957).
68. J. Rosch, "Filling optical instruments with helium," *Applied Optics* **4,** 1672–1673 (1965).
69. O. Engvold et al., "Tests of vacuum vs helium in a solar telescope," *Applied Optics* **22,** 10–12 (1983).
70. T. H. Jamieson, "Thermal effects in optical systems," *Optical Engineering* **20**(2), 156–160 (1981).
71. V. Povey, "Athermalization techniques in infra red systems," *Proceedings of the SPIE* **655,** 142–153 (1986).
72. J. E. Stumm, G. E. Pynchon, and G. C. Krumweide, "Graphite/epoxy material characteristics and design techniques for airborne instrument application," *Proceedings of the SPIE* **309,** 188–198 (1981).
73. C. Blair and J. Zakrzewski, "Coefficient of thermal and moisture expansion and moisture absorption for dimensionally stable quasi-isotropic high modulus graphite fiber/epoxy composites," *Proceedings of the SPIE* **1303,** 524–535 (1990).

CHAPTER 4

Infrared Imaging System Testing

Gerald C. Holst
Martin Marietta
Orlando, Florida

CONTENTS

4.1 INTRODUCTION

Complete characterization of infrared imaging systems includes measuring the signal transfer function (SiTF), resolution, slit response function (SRF), noise equivalent differential temperature (NEDT), modulation transfer function (MTF), and the minimum resolvable temperature (MRT) difference. The MTF can be calculated from the contrast transfer function (CTF). All these metrics can be considered as output/input transformations. The output is typically a voltage level change on the analog video or the change in monitor luminance. Depending on the test, the input may be temperature difference, target angular subtence, or target spatial frequency. Table 4.1 gives the nomenclature used in this chapter.

Table 4.1 Symbols and Nomenclature

A_d	Detector area
CTF	Contrast transfer function
d	Target size (one bar)
d_s	Size of extended source
D_{col}	Collimator clear aperture
D_{sys}	Imaging system clear aperture
f_{col}	Collimator effective focal length
f_{sys}	Infrared system effective focal length
f_0	Spatial frequency corresponding to one-half the inverse of the detector IFOV
f_s	Sampling frequency
$f/\#$	f number equal to $f_{\text{sys}}/D_{\text{sys}}$
FPN	Fixed pattern noise
FWHM	Full width at half maximum
G	Electronic amplifier gain
$\mathcal{I}(n)$	Imaginary part of the complex transfer function
IFOV	Instantaneous field of view
LSF	Line spread function
$M_e(\lambda,T)$	Planck's blackbody radiation law
MRT	Minimum resolvable temperature
MTF	Modulation transfer function
NEDT	Noise equivalent differential temperature
NU	Nonuniformity
$R(\lambda)$	Detector spectral responsivity
$\mathcal{R}(n)$	Real part of the complex transfer function
SiTF	Signal transfer function
SRF	Slit response function
$T_{\text{atm}}(\lambda)$	Atmospheric spectral transmittance
$T_{\text{col}}(\lambda)$	Collimator spectral transmittance
$T_{\text{sys}}(\lambda)$	System spectral transmittance
T_B	Background temperature
T_T	Target temperature
TV	Active time for one video line

The responsivity function is the output/input transformation in which the target size is fixed and the target intensity is varied. It is typically S-shaped. For dc coupled systems, the dark current (or noise floor) limits the minimum detectable signal and saturation limits the maximum detectable signal. For many systems, the electronics have a limited dynamic range compared to the detector and then the output is centered about some average value. Saturation in the positive and negative directions about this average value is typically limited electronically (Fig. 4.1) by the dynamic range of an amplifier or analog-to-digital (A/D) converter. The linear portion (slope) of the responsivity function is the SiTF. The SiTF must be measured to calculate the NEDT. The SiTF is not typically reported, but the dynamic range often is. Dynamic range is defined as the maximum measurable signal divided by the minimum measurable signal. For infrared imaging systems, the minimum value is usually taken as the noise level or NEDT. For example, an infrared system whose output is linear over an input range of $\Delta T = 25°C$ and has an NEDT of 0.1°C has a dynamic range of 250:1 or 48 dB. For digital non-noisy systems, the minimum value is the least significant bit, so that an 8-bit A/D converter has a dynamic range of 256:1.

Resolution can be defined by a variety of methods; the two most popular are the instantaneous field of view (IFOV) and the imaging resolution. The (geometric) IFOV is the detector active size divided by the infrared imaging system's effective focal length. The imaging resolution is that slit angular subtence at which the SRF is 0.5 (Fig. 4.2). The SRF is the output/input transformation in which the target intensity is fixed and the target angular subtence is varied. For an ideal system, the imaging resolution is one-half of

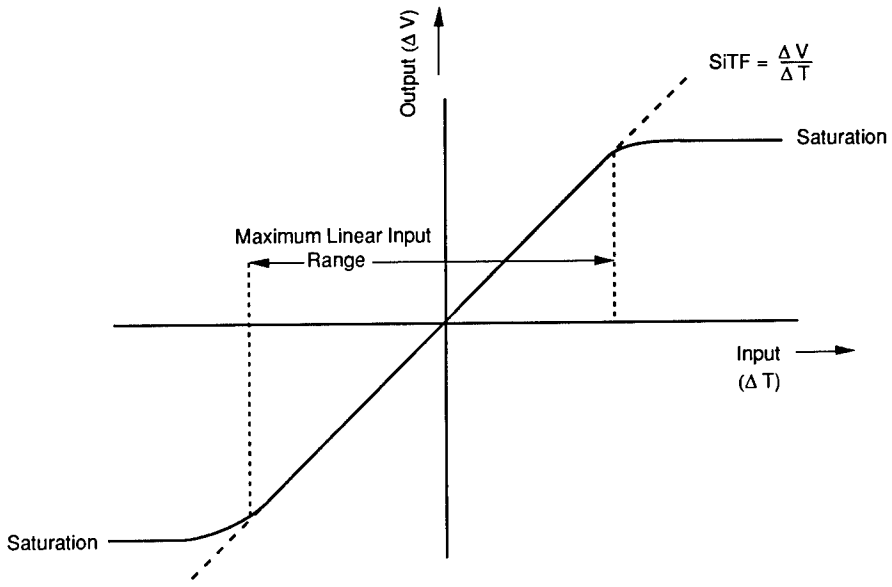


Fig. 4.1 Typical responsivity function illustrating the SiTF. The dynamic range is the maximum linear input range divided by the NEDT.

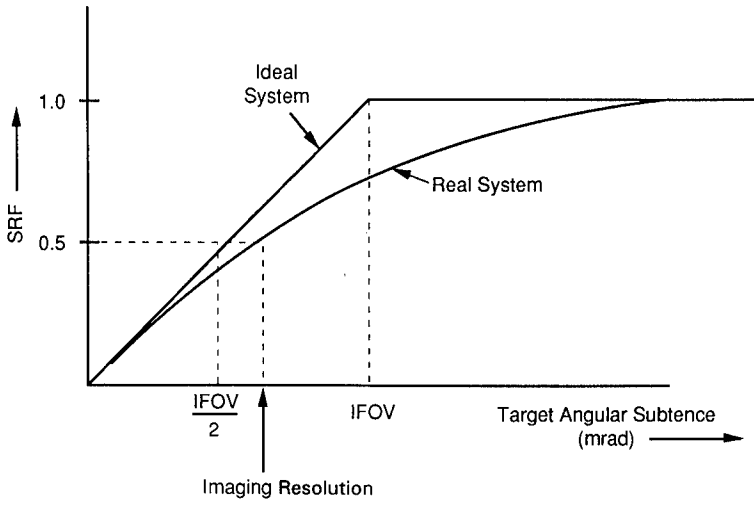


Fig. 4.2 Slit response function. For an ideal system, the imaging IFOV is 1/2 the geometric IFOV.

the IFOV. Commercial vendors tend to use the imaging resolution when specifying system resolution. The imaging resolution includes both the system's optical and electronic response and may be more representative of actual system response than the calculated geometric IFOV.

Noise is defined in the broadest sense as any unwanted signal components that arise from a variety of sources. The rms noise voltage can be referred to the input as that input that produces an SNR of unity. By dividing the rms noise voltage by the SiTF, the NEDT is obtained. The NEDT is a measure of system sensitivity only. It is an excellent diagnostic tool for production testing to verify performance. It is a poor system-to-system performance parameter and should be used cautiously when comparing systems built to different designs. Depending on the detector configuration and the scan pattern, spatial and temporal noise can occur in a variety of ways. Because of the many ways that noise can appear, it is evaluated in terms of a three-dimensional noise model. This noise model incorporates the NEDT, fixed pattern noise (FPN), and nonuniformity.

The MTF is a fundamental parameter used for system design, analysis, and specifications. The 3-D noise model parameters and MTF uniquely define system performance. Theoretically, the SRF and CTF can be calculated from the MTF. If the eye's detection threshold can be accurately modeled, the MRT can also be calculated. However, it is often easier to measure these parameters rather than calculate them.

Methods to measure the MTF depend on both optical and electronic signal considerations. It is the system's response to spatial sinusoidal signals. MTF is a measure of how well the system faithfully reproduces the scene. As the MTF decreases, scene detail associated with those specific spatial frequencies are reproduced with lower contrast. The limit, where the MTF approaches zero, is the system cutoff. The system can detect signals whose spatial frequencies are above cutoff but cannot faithfully reproduce them. For example, a four-bar

pattern appears as a blob. In undersampled systems such as staring arrays, the MTF is defined only up to the Nyquist frequency. Spatial frequencies above Nyquist are aliased to lower spatial frequencies. In this case, a four-bar pattern may appear as a distorted three-bar pattern. From a design point of view, the MTF should be "high" over the spatial frequencies of interest. The spatial frequencies of interest are application specific.

There are two general approaches for determining the MTF: the direct method, which consists of measuring the response to sinusoidal targets, and the indirect method in which a one-dimensional MTF is obtained from the Fourier transform of the line spread function (LSF). There are benefits and shortcomings to both methods. Sinusoidal patterns are available in the visible but not easy to fabricate for the infrared. One could use square (bar) targets to obtain the CTF and mathematically convert to the sinusoidal response (MTF) using a series approximation.

Modulation is defined by

$$\text{modulation} = \frac{B_{\max} - B_{\min}}{B_{\max} + B_{\min}} , \tag{4.1}$$

where $B_{\max}$ and $B_{\min}$ are the maximum and minimum intensity levels defined in Fig. 4.3. The modulation transfer function is given by

$$\text{MTF} = \frac{\text{image modulation}}{\text{object modulation}} . \tag{4.2}$$

It is plotted as a function of spatial frequency, as shown in Fig. 4.4. The CTF is the system's response to square waves. Using Fig. 4.4 as a reference, if the sinusoids are replaced with square waves, the CTF is obtained.

MRT is a subjective measure of image quality. It is a laboratory summary measure that combines visual sensitivity and resolution (Fig. 4.5). MRT is a measure of an observer's ability to resolve a four-bar pattern embedded in noise. Visual sensitivity is the measure of the lowest SNR that is acceptable to the eye. The resolution limit is related to the smallest detail that can be

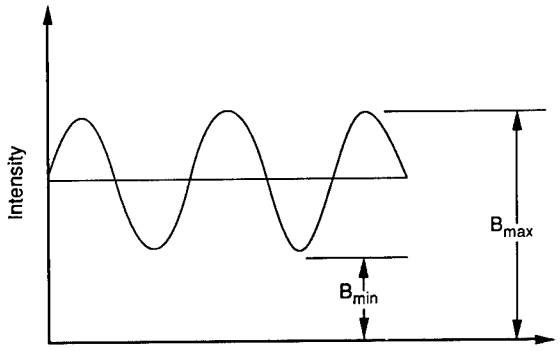


Fig. 4.3 Definition of modulation.

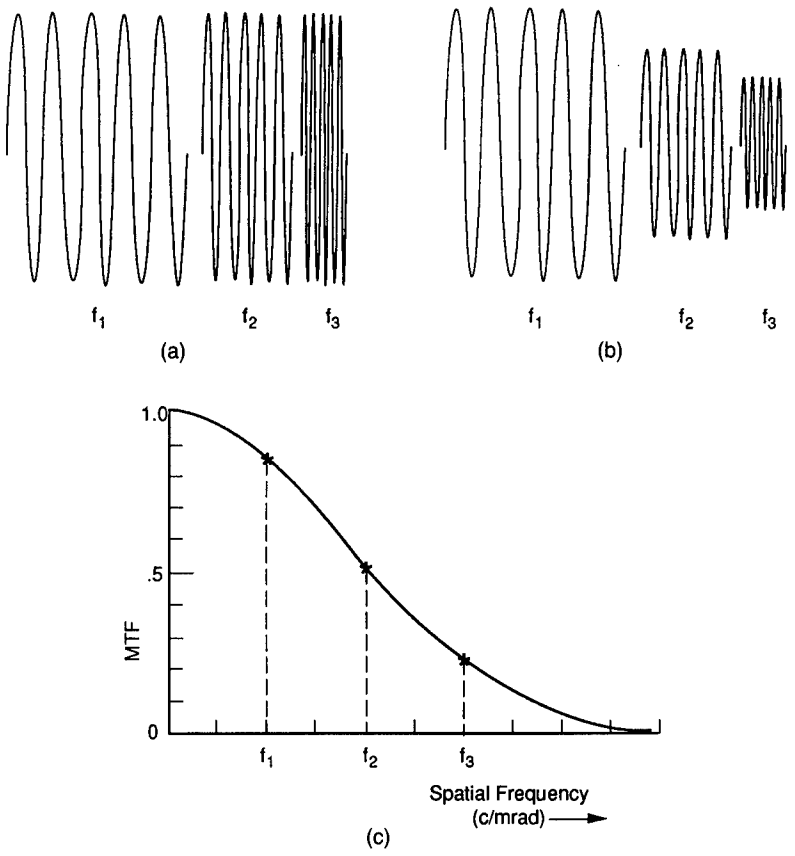


Fig. 4.4 Definition of modulation transfer for three spatial frequencies: (a) input (object) modulation, (b) output (image) modulation, and (c) the MTF.

faithfully reproduced. Because MRT is inversely related to the MTF, as the MTF $\rightarrow$ 0, the MRT $\rightarrow \infty$. Because of the eye's incredible ability to spatially and temporally integrate signals, the low-frequency MRT asymptote depends on the amount of low-frequency noise (nonuniformity) present. With minimal low-frequency noise, the MRT asymptotes to zero.[1] With modest noise, the MRT asymptotes[2,3] somewhere between 0.3 and 0.7 times the NEDT. The MRT results depend on decisions made by an observer. The results vary with training, motivation, and visual capacity. Because of the larger inter- and intraobserver variability, several observers are required.

Object space spatial frequency is used to characterize the MTF, CTF, and MRT. The angle subtended by one cycle (one bar and one space) is given by $2d/R$, where $2d$ is the spatial extent of one cycle and R is the distance from the infrared imaging system entrance aperture to the target (Fig. 4.6). If a collimator is used, R is replaced by the focal length of the collimator f_{COL} so that targets placed in the collimator's focal plane can be described in object

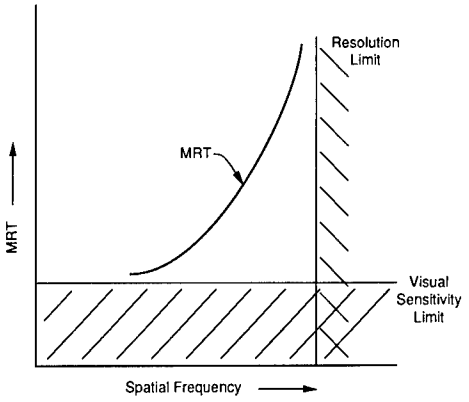


Fig. 4.5 MRT as a function of spatial frequency.

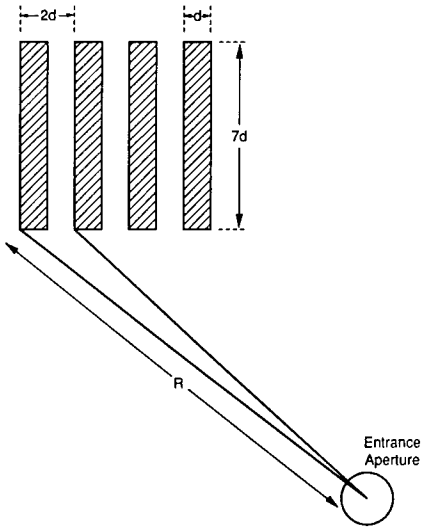


Fig. 4.6 Definition of spatial frequency. MRT targets have a 7:1 aspect ratio.

space. The object space spatial frequency f_x is the inverse of the target angular subtense and is usually measured in cycles per milliradian:

$$f_x = \frac{1}{1000}\left(\frac{R}{2d}\right) \quad \text{or} \quad f_x = \frac{1}{1000}\left(\frac{f_{\text{COL}}}{2d}\right) . \tag{4.3}$$

Although this chapter focuses on the characterization of infrared imaging systems, the test equipment, test methodology, and data analysis is generic to all electro-optical imaging systems. Aside from using different sources (e.g., a CIE C source for the visible and a blackbody for the infrared), the only major difference is the nomenclature. For example, the linear portion of the responsivity function is the responsivity for systems operating in the visible and the SiTF for infrared imaging systems. The rms noise voltage when referred to the input produces the noise equivalent background input for visible systems and the NEDT for infrared imaging systems.

4.2 TEST PHILOSOPHY

Before undertaking a test program, it is necessary to establish a test philosophy and write a thorough test plan. The test plan states the objectives, requirements, data analysis methodology, and the success criteria of the test. The test configuration and measuring conditions must be clearly stated.

Design engineers and test engineers must anticipate all foreseeable combinations of circumstances. Test procedures can only be developed when the operation of the system is fully understood and the test objectives are fully understood. An intimate understanding of how the system operates and the expected results is a fundamental necessity. The specific tests selected depend on the infrared system application.

Because of the complexities of modern infrared systems, it is essential to employ state-of-the-art measuring equipment and the latest test concepts and to ensure proper data analysis is used. However, finding adequate test equipment becomes a challenge as infrared system technology advances. With the emerging requirements for more sensitive and more accurate infrared systems, the measurement techniques become more stringent. In some areas, test equipment technology has not kept pace with the advances in infrared system technology.

Because the data set tends to be small, appropriate statistical measures must be considered. A variety of statistics are appropriate for imaging systems.[4] General data analyses can be found in texts such as Dixon and Massey[5] and Bendat and Piersol.[6] Formal error analysis techniques can be found in a variety of texts (see for example, Ref. 7).

At the conclusion of the test, fully document any test abnormality and all results. The recorded data should include as a minimum the raw data, all calculated results, ambient temperature, and any other pertinent measuring data. The data should be shown both graphically and in tabular form and, if appropriate, the calculated least-squares regression line should overlay the data points. Data should be in a format useful to multiple users and multiple analysts. With the widespread use of personal computers, it is prudent to store all data in a format consistent with that used by the more common spreadsheets and database programs.

4.3 RADIOMETRY AND THE ΔT CONCEPT

A linear system produces an output voltage difference that is proportional to the radiant exitance difference between a target and its background. Assuming a Lambertian source and a circular unobscured aperture, the voltage difference is

$$\Delta V = G\int_{\lambda_1}^{\lambda_2} R(\lambda)\frac{\Delta M_e}{4(f/\#)^2}T_{\text{atm}}(\lambda)T_{\text{col}}(\lambda)T_{\text{sys}}(\lambda)A_d\, d\lambda \ , \tag{4.4}$$

where ΔM_e is the spectral radiant exitance difference between the target and its background equal to $M_e(\lambda,T_T) - M_e(\lambda,T_B)$. For small signals and averaged transmittances, Eq. (4.4) can be approximated by

$$\Delta V = G\frac{\overline{T_{\text{col}}}\ \overline{T_{\text{atm}}}\ \overline{T_{\text{sys}}}\ \Delta T_{\text{source}}A_d}{4(f/\#)^2}\int_{\lambda_1}^{\lambda_2} R(\lambda)\frac{\partial M_e(\lambda,T_B)}{\partial T}\, d\lambda \ , \tag{4.5}$$

where the thermal derivative (derivative of Planck's blackbody law with respect to temperature) is evaluated at the background temperature T_B.

The responsivity function is the output/input transformation, or $\Delta V/\Delta M_e$. For systems operating in the visible, the source intensity is often varied by inserting neutral density filters. For infrared imaging systems, it is convenient to express a small radiant exitance difference by an effective temperature difference, or ΔT_{source}. Equation (4.5) can then be represented by

$$\Delta V = \text{SiTF}\,\Delta T \ , \tag{4.6}$$

where

$$\Delta T = \overline{T_{\text{col}}}\ \overline{T_{\text{atm}}}\ \Delta T_{\text{source}} \ , \tag{4.7}$$

where ΔT_{source} is proportional to the thermal derivative, which in turn is a function of wavelength and the background temperature (Fig. 4.7). Most infrared imaging systems do not measure temperature but rather respond to radiance differences, and radiance differences are nonlinear with scene temperature. As a result, all measurements that use ΔT as an input are affected by the background temperature. The ΔT concept is a matter of convenience. The radiant exitance difference between 280 and 281 K is different than that between 300 and 301 K even though the thermometric difference is 1 deg. In the 8- to 12-μm spectral region, a temperature differential of 1.2° at 280°C produces the same radiant exitance differential as 1° at 300°C (Fig. 4.8). Equiv-

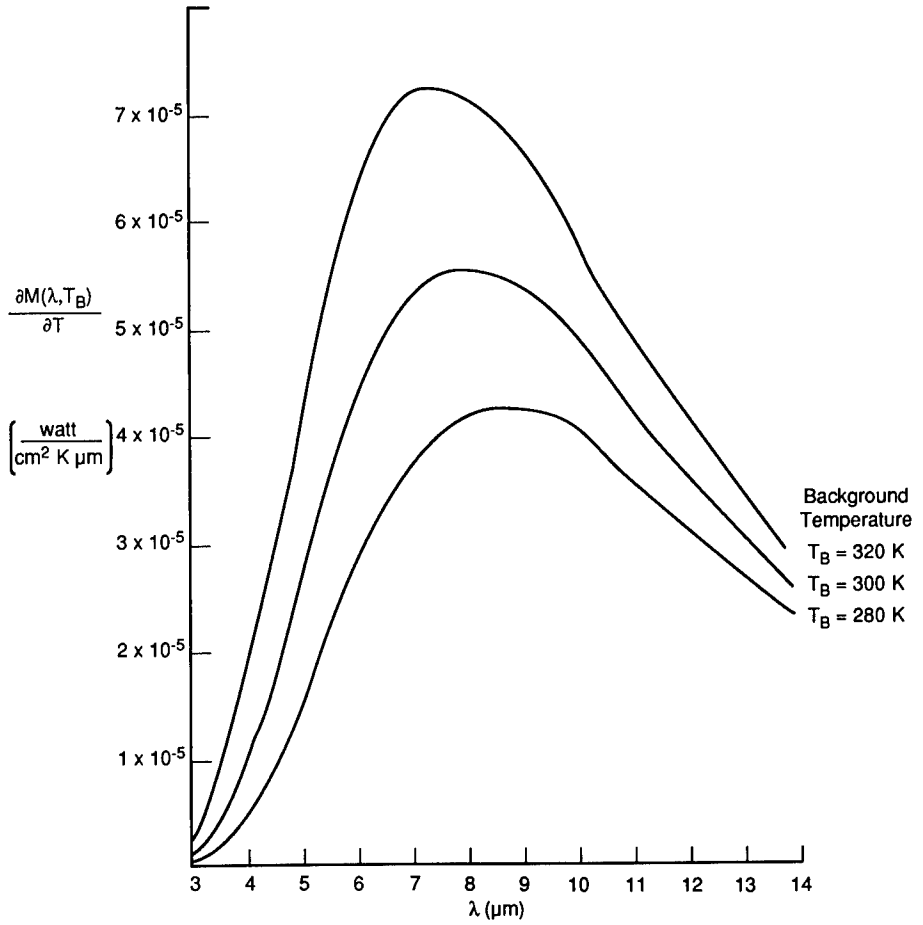


Fig. 4.7 Thermal derivative of Planck's blackbody law for three background temperatures.

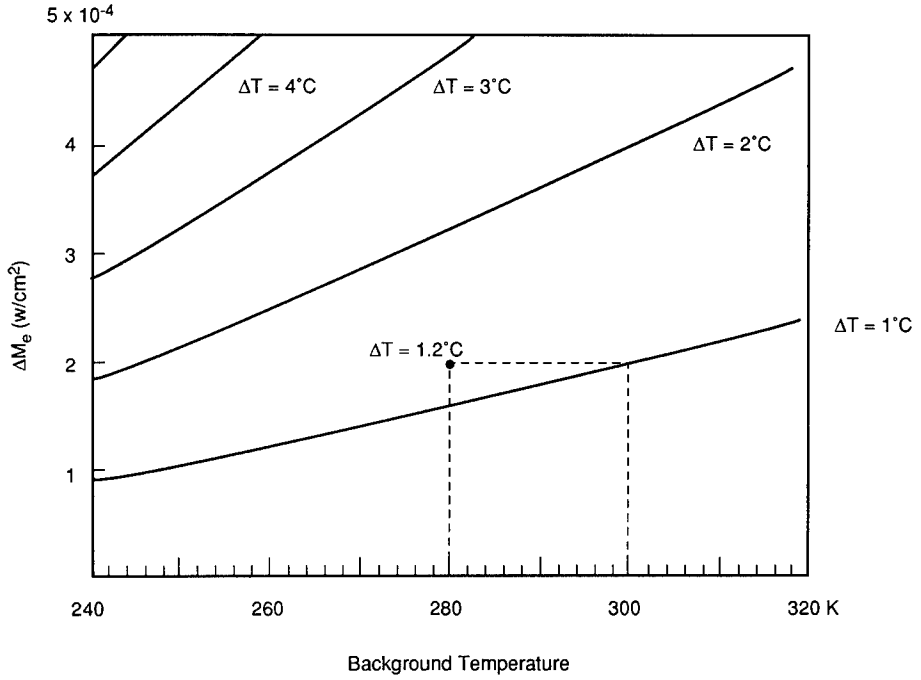


Fig. 4.8 Representative radiant exitance differences for a fixed thermometric difference in the 8- to 12-μm region. The precise difference depends on the system's spectral response.

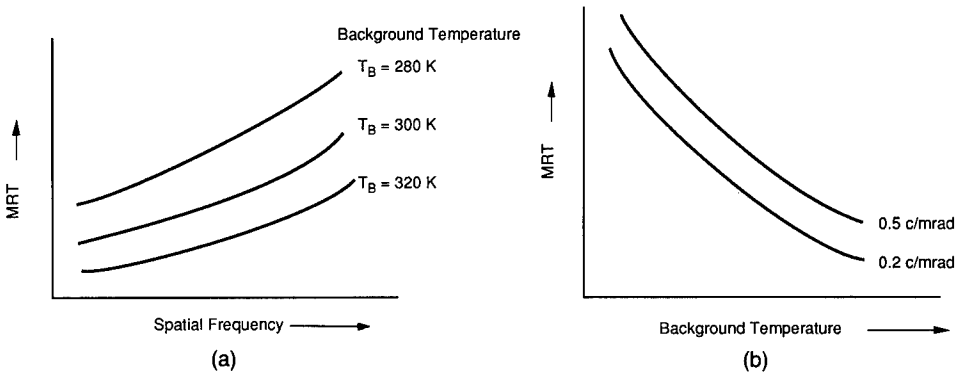


Fig. 4.9 MRT variation as a function of (a) spatial frequency and (b) background temperature.

alently, as the background temperature decreases, a larger thermometric difference is required to produce a constant SNR. A drift in ambient of 1 K manifests itself as a ΔT change of 0.02 K. Background temperature changes affect[8,9,10] the MRT (Fig. 4.9) and the NEDT (Fig. 4.10). The 3- to 5-μm region is more sensitive to background changes than the 8- to 12-μm region. By allowing the ambient temperature to drift from 300 to 305 K, the NEDT has approximately 14% variation in the 3- to 5-μm spectral region and approxi-

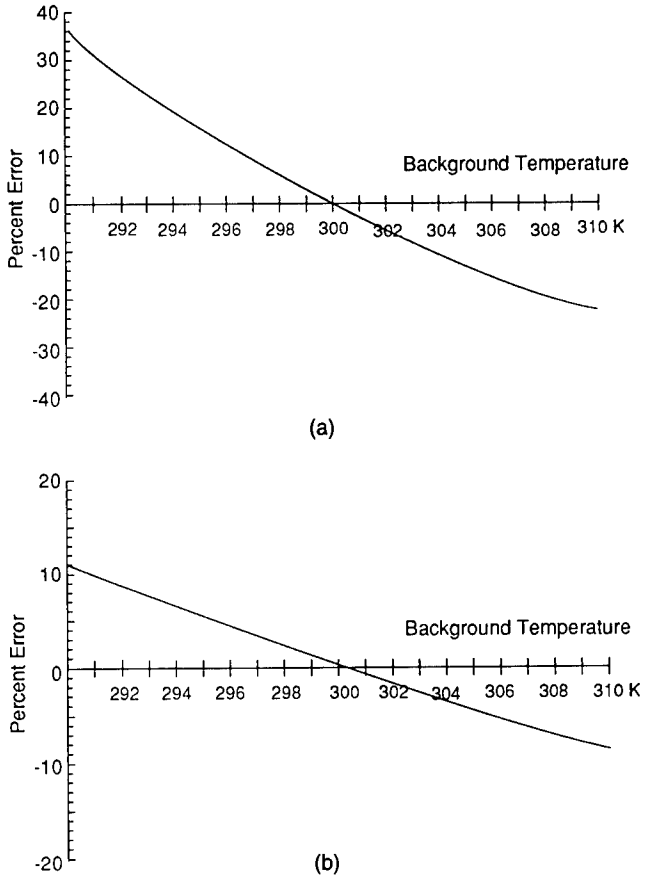


Fig. 4.10 Representative percentage error (relative to 300 K) in the NEDT for different background temperatures for (a) 3- to 5-μm and (b) 8- to 12-μm spectral region. The precise error depends on the system's spectral response.

mately 5% in the 8- to 12-μm spectral region. This clearly illustrates that although the ΔT concept may be useful for thermometers, it does not uniquely specify infrared imaging system performance unless both the spectral response of the system and the ambient (background) temperature are specified. Figures 4.8 through 4.10 should be taken as representative. The specific variations with ambient temperature depend on the spectral response of the system. The linear approximation given by Eq. (4.6) is valid only for small excursions about the assumed background temperature.

4.4 TEST CONFIGURATIONS

The purpose of the source, target, and collimator is to present standardized targets of known intensity to the infrared imaging system. The collimator aberrations should be significantly less than the aberrations of the system under test. Because aberrations[11] are inversely proportional to $f/\#$ or to powers

of $f/\#$, the $f/\#$ of the collimator should be greater than the $f/\#$ of the system under test. Equivalently, the focal length of the collimator should be much longer than the focal length of the system under test. As a general guideline, the collimator focal length should be at least five times that of the system under test. Collimators are lens or mirror assemblies that optically place targets at infinity. Collimators may contain either refractive or reflective elements. Either can be used in the visible spectral region, whereas reflective collimators are generally used in the infrared spectral region because infrared refractive optics tend to be dispersive and reflective mirrors do not have chromatic aberrations. Parabolic mirrors do not have spherical aberrations, and therefore parabolic mirrors tend to be a better choice. Off-axis parabolic collimators allow maximum throughput in the sense that the source does not physically interfere with the collimated beam.

Although a point source produces parallel light when placed in the focal plane of a collimator, an extended source of size d_s can be fully seen only in a well-defined region (Fig. 4.11) that extends out to a distance L from the collimator lens or mirror:

$$L = \frac{f_{\text{col}}}{d_s} D_{\text{col}} \ . \tag{4.8}$$

The infrared imaging system must be placed inside the cone defined by the collimator clear aperture and L. The maximum distance R that the system can be placed is determined by the system's aperture:

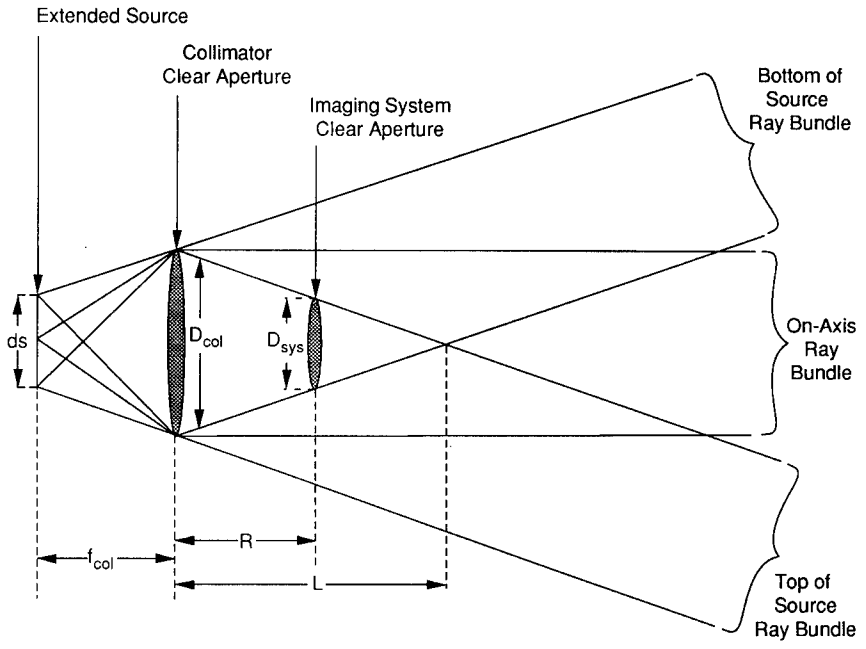


Fig. 4.11 Available working distance using a collimator and extended source.

$$R = \frac{f_{\text{col}}}{d_s}(D_{\text{col}} - D_{\text{sys}}) \ . \tag{4.9}$$

The value of d_s used for these calculations should be the largest anticipated size of the target and its background. The collimator aperture should be sufficiently large so that no extraordinary care is required to center the imaging system onto the collimator.

Example 1. What is the maximum distance that an infrared imaging system can be placed from a collimator whose focal length is 40 in. and clear aperture is 6 in.? The imaging system's entrance aperture is 5 in. The largest linear angular dimension (target plus background) to be viewed subtends 20 mrad.

$$\text{target size} = d_s = f_{\text{col}}\theta = (40)(20 \times 10^{-3}) = 0.8 \text{ in.} \ , \tag{4.10}$$

$$R \leq \frac{f_{\text{col}}}{d_s}(D_{\text{col}} - D_{\text{sys}}) = \frac{40}{0.8}(6 - 5) = 50 \text{ in.} \ . \tag{4.11}$$

At this distance, the infrared imaging system must be precisely on-axis with the collimator. The imaging system should be placed at a shorter distance to alleviate alignment difficulties. □

The source intensity reaching the infrared imaging system is modified by the collimator transmittance and the atmospheric transmittance [Eq. (4.7)]. For short pathlengths, the atmospheric transmittance approaches unity in the usual imaging bands. For longer pathlengths, the atmospheric transmittance must be considered in the measurements. For systems whose spectral response lies in those regions where the atmospheric transmittance affects the results, it is necessary to perform a detailed analysis to appropriately account for the atmospheric transmittance.

Figure 4.12 illustrates the various methods by which the output can be measured. The analog video can be captured with a frame grabber (image capture board), transient recorder, or an oscilloscope. Because of phasing effects, just satisfying the Nyquist criterion (two samples per highest frequency) is insufficient when amplitude and/or pulse fidelity are characteristics of interest. The measurement equipment bandwidth must be greater than the infrared imaging system bandwidth to avoid phasing problems. It is prudent to verify system bandwidth prior to selecting measuring equipment. The output video may be in a standard format in terms of timing (e.g., RS 170) but may have a much wider bandwidth than the standard video format. The sampling rate of the frame grabber is usually fixed and may be too slow to match the pixel rate of the infrared imaging system. In those cases, a frame grabber is appropriate for only those measurements that deal with low-frequency response, such as the SiTF.

The monitor luminance can be measured with a scanning microphotometer, stationary solid-state camera (CCD camera), or scanning fiber optic bundle. This method is appropriate for those systems that have an integral monitor (direct view devices) or are manufactured with specialized (unique) monitors. Monitor performance is not consistent,[12] and each monitor should be calibrated

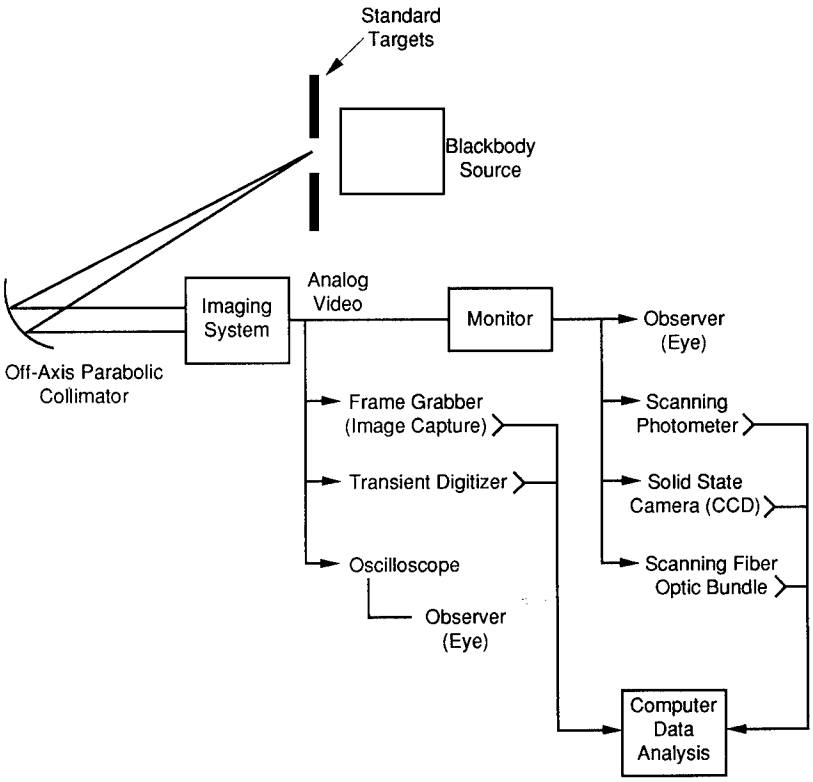


Fig. 4.12 Various methods of capturing the system output.

before starting any test. The advantage of measuring the monitor luminance is its greater adaptability. Neither the test equipment nor the measurement techniques need to be changed as different infrared imaging systems with different video standards are tested.

4.5 SIGNAL TRANSFER FUNCTION

The test configuration for measuring the responsivity function is shown in Fig. 4.13. Ideally, the background should fill the entire field of view so that there are no competing effects from the surroundings. In a well-designed system, the measured SiTF should approach the calculated value [Eqs. (4.5) and (4.6)]. The SiTF can vary from system to system if the spectral response of each system is different. SiTF by itself is not a very good metric for comparing different systems because it can be changed simply by changing the linear gain. If the SiTF is specified, it usually implies that the system is operating at maximum gain.

With ac coupled scanning systems, the signal is coupled in the horizontal direction but not in the vertical direction. As a result, there is an overall voltage shift (signal and background) in the scan direction, which is not present

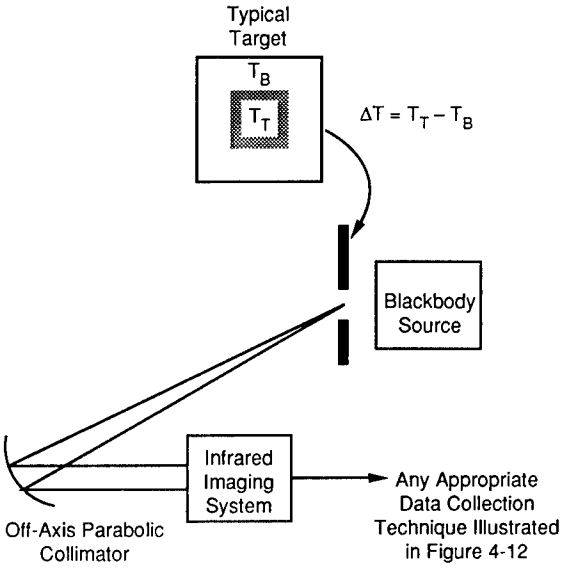


Fig. 4.13 Responsivity function test configuration.

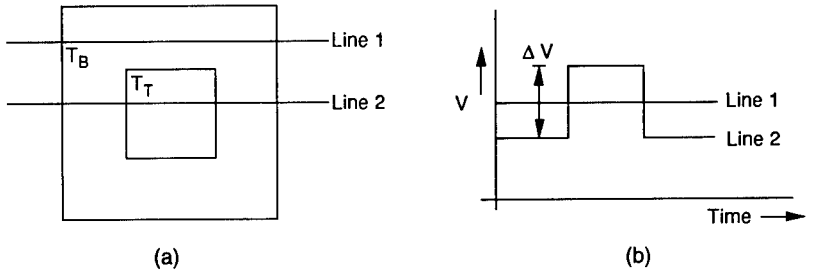


Fig. 4.14 The ac coupling changes the absolute value of the output when viewing a target: (a) two lines across the field of view and (b) output voltages for the two lines.

in the vertical direction (Fig. 4.14). The voltage difference between the signal and the background must be measured, not just the peak values of the signal. The ac coupling extends the dynamic range in that the available differential voltage (between the target and its background) can be equal to the full available output voltage if the target fills 50% of the field of view. Small-area targets reach saturation before large-area targets do and small-area targets have a smaller dynamic range. The target size must be specified if saturation or dynamic range is specified.

The SiTF is the linear slope of the responsivity curve (Fig. 4.1). The best estimate of the slope can be determined by performing a least-squares fit to N data point pairs $(\Delta V_i, \Delta T_i)$:

$$\text{SiTF} = \frac{N\sum_{i=1}^{N}\Delta V_i\Delta T_i - \sum_{i=1}^{N}\Delta V_i\sum_{i=1}^{N}\Delta T_i}{N\sum_{i=1}^{N}(\Delta T_i)^2 - \left(\sum_{i=1}^{N}\Delta T_i\right)^2} . \tag{4.12}$$

There may be an offset in the responsivity function, which may result from the inability to measure the background temperature accurately and/or from different emittances between the target and its background. It is this offset that should be noted at different laboratories when comparing test techniques. Unless it is known that no offset exists, the SiTF cannot be accurately determined by making a measurement at one input value only. The offset is

$$V_{\text{offset}} = \overline{\Delta V} - \text{SiTF}\,\overline{\Delta T} , \tag{4.13}$$

where

$$\overline{\Delta T} = \frac{\sum_{i=1}^{N}\Delta T_i}{N} , \tag{4.14}$$

and

$$\overline{\Delta V} = \frac{\sum_{i=1}^{N}\Delta V_i}{N} . \tag{4.15}$$

The SiTF test procedure is as follows:

1. Measure the output voltage difference between the target and its background for a variety of source intensities (ΔT_i).
2. Multiply all ΔT_i's by the average collimator transmittance and average atmospheric transmittance to obtain the ΔT at the infrared imaging system entrance window.
3. Plot all data points as well as the least-squares fit to the linear portion of the responsivity function.
4. Record the SiTF and the ambient temperature.

It is extremely important to have unambiguous testable specifications. Since ΔT is a function of the background temperature, the SiTF specification must clearly state the measurement conditions (Table 4.2).

Table 4.2 Typical SiTF Specification

The SiTF shall be greater than 0.2 *V*/°C when the background temperature is 20°C. If the measurement is performed at another background temperature, a correction factor must be applied (system specific).

Table 4.3 SiTF Data Set

ΔT_{source} (measured)	ΔV (mV) (measured)	ΔT (at entrance aperture)	ΔV	$\Delta T \Delta V$	$(\Delta T)^2$
−5	−330				
−4	−320				
−3	−300	−2.7	−300	810	7.29
−2	−185	−1.8	−185	333	3.24
−1	−80	−0.9	−80	72	0.81
1	140	0.9	140	126	0.81
2	220	1.8	220	396	3.24
3	310	2.7	310	837	7.29
4	350				
5	355				
		$\Sigma = 0$	$\Sigma = 105$	$\Sigma = 2574$	$\Sigma = 22.68$

Example 2: SiTF Calculation. An output of an infrared imaging system was measured for 10 intensity inputs (Table 4.3). The collimator transmittance is 0.90 and the atmospheric transmittance is assumed to be unity. What is the SiTF?

The input ΔT of -5, -4, 4, and 5 produced a nonlinear response. As a result, these four values were eliminated from the least-squares analysis.

$$\text{SiTF} = \frac{(6)(2574) - 0}{(6)(22.68) - 0} = 113.5 \tag{4.16}$$

The average voltage is $V = 105/6 = 17.5$ and $V_{\text{offset}} = 17.5 - 0 = 17.5$. The system output is

$$\Delta V_{\text{out}} = (113.5)\Delta T + 17.5 \ . \tag{4.17}$$

The data and least-squares fit are shown in Fig. 4.15. □

4.6 THREE-DIMENSIONAL NOISE MODEL

The noise analysis approach taken follows that of D'Agostino and Webb.[13] The noise is divided into a set of eight components that are related to the temporal and spatial dimensions of a three-dimensional coordinate system (Fig. 4.16). Analyzing the noise in this manner has the advantage that it simplifies the understanding of a complex phenomenon by breaking it down into a manageable set of components. Furthermore, it provides the system designer insight concerning possible hardware and software factors that may be responsible for the noise. For the system performance modeler, the method simplifies the incorporation of complex noise factors into model formulations. The 3-D noise model incorporates the NEDT, FPN, and nonuniformity.

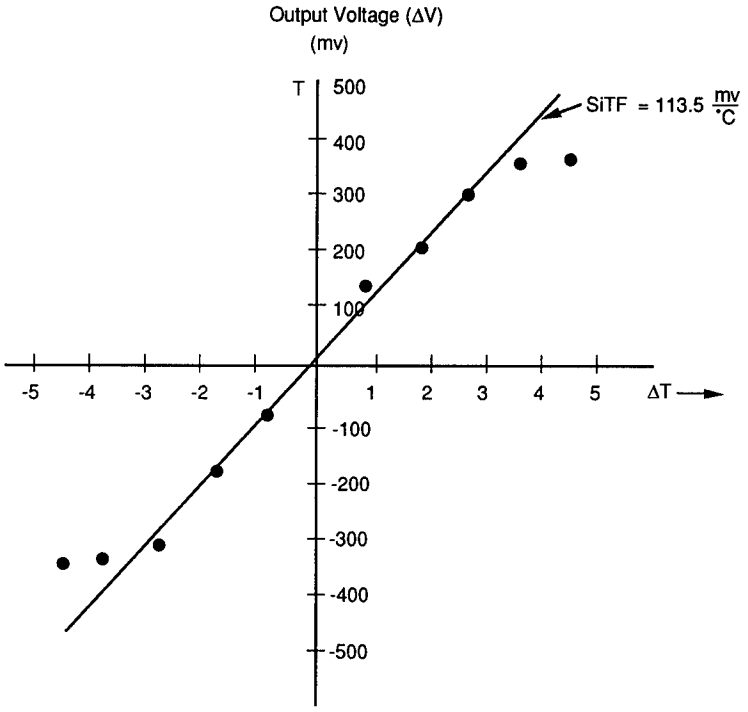


Fig. 4.15 SiTF from Example 4.2.

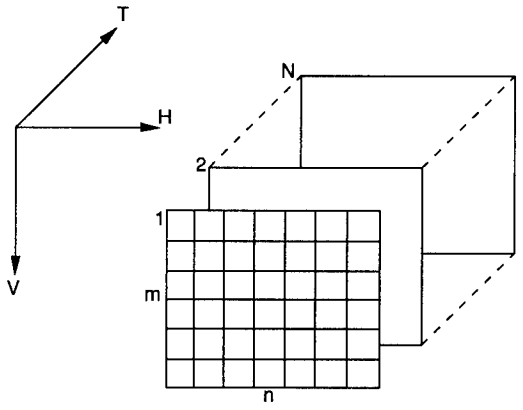


Fig. 4.16 Three-dimensional noise coordinate system.

The third dimension is the temporal dimension representing the framing sequence. The other two dimensions provide spatial information. However, depending on the infrared imaging system design, the horizontal dimension may represent time for a scanning system or may represent space for a staring system. For a staring array, m and n indicate detector locations. For pushbroom scanning systems, m indicates detector locations and n is the digitized analog

signal. For serial scanning systems, m is the number of raster lines and n is the digitized analog signal.

Noise may manifest itself in a variety of different ways, such as line-to-line nonuniformity, rain, moving bands, and flashing channels. These effects are difficult to quantify because of the transitory nature of the noise. Some noise may be very easy to perceive but difficult to measure. For example, the eye is very sensitive to intensity variations that change from frame to frame (flicker). The flicker may be low in intensity and not easily measured in a single frame but may be very noticeable in active video.

Table 4.4 lists seven noise components and some possible contributors to the components for serial scanning, parallel scanning, and staring array imaging systems. For mathematical completeness, the noise model has eight components with the eighth being the global average value S. Depending on the system design and operation, any one of these noise components could dominate. The origin of these components are significantly different and the

Table 4.4 Seven Noise Components of the 3-D Noise Model

3-D Noise Component	Description	Serial Scan	Parallel Scan	Staring Array
N_{TVH}	Random 3-D noise	Random and $1/f$ noise	Random and $1/f$ noise	Random and $1/f$ noise
N_{VH}	Spatial noise that does not change from frame to frame	Nonuniformity	Nonuniformity	FPN and nonuniformity
N_{TH}	Variations in column averages that change from frame to frame (rain)	—	EMI*	Readout electronics
N_{TV}	Variations in row averages that change from frame to frame (streaking)	EMI*	Transients (flashing detectors)	Readout electronics
N_V	Variations in row averages that are fixed in time (horizontal lines or bands)	Line-to-line interpolation	Detector response variations, line-to-line interpolation	Readout electronics, line-to-line interpolation
N_H	Variations in column averages that are fixed in time (vertical lines)	—	—	Readout electronics
N_T	Frame-to-frame intensity variations (flicker)	Frame processing	Frame processing	Frame processing

*Electromagnetic interference

Table 4.5 Three-Dimensional Noise Descriptors

	Temporal Noise Component	Spatial Noise Component
Pixel variations	N_{TVH}	N_{VH}
Row variations	N_{TV}	N_V
Column variations	N_{TH}	N_H
Frame variations	N_T	S

Table 4.6 Number of Elements in Each Noise Component

3-D Noise Component	Number of Elements, N_e
N_{TVH}	$m \times n \times N$
N_{VH}	$m \times n$
N_{TH}	$n \times N$
N_{TV}	$m \times N$
N_V	m
N_H	n
N_T	N
S	1

existence and manifestation depends on the specific design of the infrared system. Not all of the components may be present in every infrared system. Certain noise sources, such as microphonics, are more difficult to describe because they may appear in a wide variety of forms. This is illustrated by listing *readout electronics* as a catch-all phrase for staring array artifacts. The 3-D noise model provides the basic framework for analyzing the various noise sources. Depending on the system design and operation, the same noise source may appear in different noise components. The spatial and temporal components are shown in Table 4.5. The number of data elements for each component is given in Table 4.6 and illustrated in Fig. 4.17. The complete data set N_{TVH} was illustrated in Fig. 4.16.

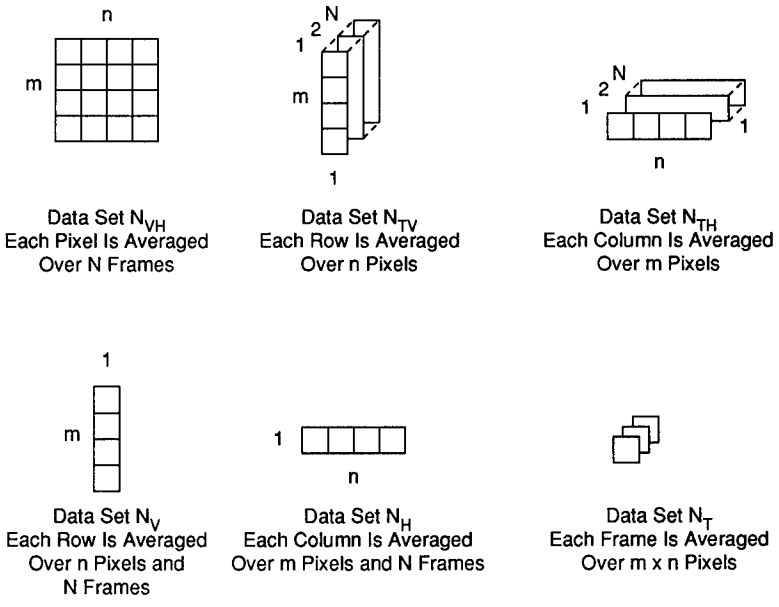


Fig. 4.17 Three-dimensional model data sets.

Table 4.7 NEDT, FPN, and Nonuniformity

3-D Noise Component	Frequency Component	Serial Scan	Parallel Scan	Staring Array
N_{TVH}	High	NEDT	NEDT	NEDT
	Low	$1/f$	$1/f$	$1/f$
N_{VH}	High	—	—	FPN
	Low	Nonuniformity	Nonuniformity	Nonuniformity

The temporal and spatial noise can be further divided into high- and low-frequency components (Table 4.7). The NEDT is defined for high-frequency temporal noise, and low-frequency temporal noise is $1/f$ noise. High-frequency spatial noise is the fixed pattern noise and low-frequency spatial noise is nonuniformity. The observer sees the effects of both spatial and temporal noise simultaneously. Low-frequency components appear as streaks or variations in luminance on the monitor, which affect the low-spatial-frequency MRT results. Unless the system is designed to reproduce large low-contrast targets, the low-frequency components tend to be annoying, but generally do not affect the intended use of the system. The high-frequency components can significantly affect the ability to resolve small detail.

Low-frequency noise is defined as any noise that has frequency components less than 150 kHz and high-frequency noise as noise that has components above 150 kHz, when referred to the standard RS 170 video format.[14] The 150-kHz bandpass filter can either be implemented in hardware or simulated in software. Because RS 170 has a bandwidth of approximately 5 MHz, the low-frequency component represents about 3% of the noise bandwidth. For other video formats, the low-frequency cutoff should be 3% of the bandwidth. The removal of low-frequency components from the high-frequency signals is called *trend removal.*[15] The high-pass filter removes any dc offset or trend in the data.

The test configuration to measure all noise sources is shown in Fig. 4.18. The detectors should be uniformly illuminated (flood illumination) either by placing a nonreflective opaque cloth over the infrared imaging system or hav-

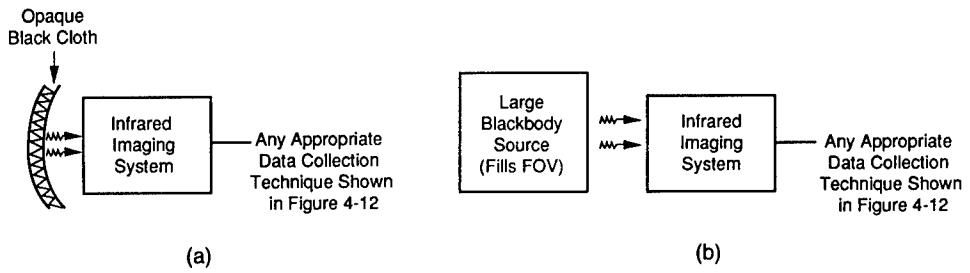


Fig. 4.18 Noise measurement test configuration: (a) flood illumination created by a single background (ambient) temperature using a black cloth and (b) viewing a large blackbody to create various background temperatures.

ing the system view a large, uniform blackbody source. The source or cloth must cover the entire area that the detector can sense. To ensure that any nonuniformity seen is not a result of the source (or the cloth), the source should be moved around to verify that the image does not move (e.g., any nonuniformity seen is produced by the system and not the source).

Historically,[16] NEDT was defined at the output of the postamplifiers. For this classical measurement, a simple single-pole filter is added whose 3-dB break frequency is equal to the reciprocal of twice the detector dwell time (Fig. 4.19). For system measurements, the NEDT is defined at the analog video or the output of the monitor. When measuring the system NEDT, the external 3-dB filter should *not* be used.

It is the data collection and analysis techniques that allow separation of the various temporal and spatial components. The high-frequency temporal components can be separated from the spatial components via a frame subtraction technique (Fig. 4.20). The inclusion of a high-pass filter provides the bandpass necessary to remove trends. The temporal components are reduced by the square root of the number of frames averaged, and the fixed pattern components are not. This averaged frame is then subtracted from a single frame to leave only the temporal variations. This subtraction method produces one frame (high-frequency component) of the data set N_{TVH} from which the NEDT can be calculated. The complete data set contains N frames.

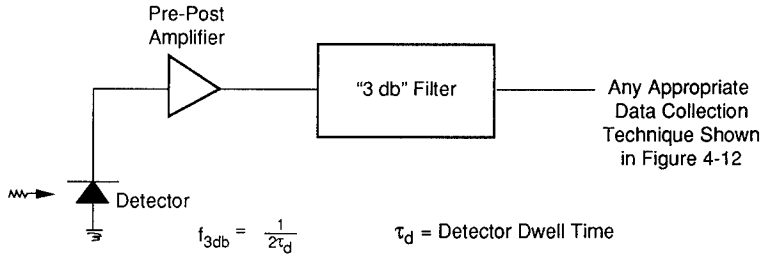


Fig. 4.19 The 3-dB filter. Used only when measuring the NEDT at the output of the postamplifiers.

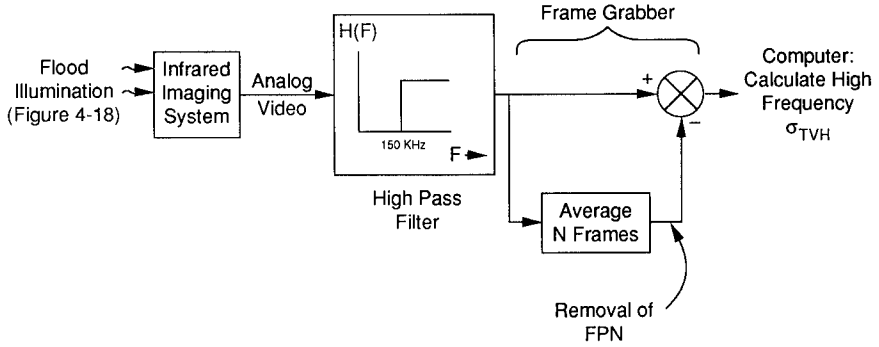


Fig. 4.20 Using a frame grabber to collect the high-frequency component of data set N_{TVH}. The NEDT is calculated from this data set.

The standard deviation σ_i of the data set N_i is the rms noise value for that data set. The noise components can be considered bidirectional if the calculated variance is independent of the direction chosen to perform the calculation. That is, the variances are essentially the same whether measured on a line, column, globally, or frame to frame. Systems exhibiting this behavior are said to be *ergodic*.[17] If the system does not exhibit any fixed pattern noise and the noise is considered truly bidirectional ($N_{VH} = N_{TH} = N_{TV} = N_V = N_H = N_T = 0$), then the averaging of many frames is not necessary, and a transient digitizer can be used to examine only one line (Fig. 4.21). If there are minimal low-frequency components, the high-pass filter can be eliminated.

FPN can be measured after averaging many frames together and then passing the data through a high-pass filter (Fig. 4.22). FPN is the standard deviation of the resultant data set. The amount of fixed pattern noise can be expressed as a percentage of the average value or the NEDT. When describing dc-coupled systems or detector outputs, FPN is usually normalized to the average value. But, for most systems, the NEDT reference is appropriate. By using the NEDT reference, the percentage FPN is independent of system gain and level settings.

Because of variations in detectors, variations in hardware and truncation errors imposed by the correction circuitry, the fixed pattern noise is not totally removed. As a result, the FPN is a function of the correction technique. At the calibration points, the FPN is minimized. At other background inputs, the FPN can be significant[18] (Fig. 4.23). The amount of FPN present is system specific. A large blackbody source is required to measure the FPN variation

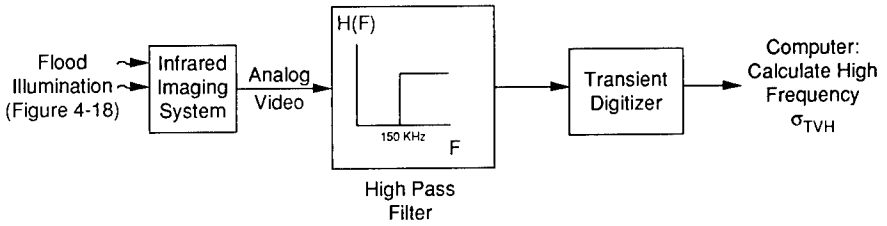


Fig. 4.21 Using a transient digitizer to collect the high-frequency component of data set N_{TVH}. This approach is valid if $N_{VH} = N_{TH} = N_{TV} = N_V = N_H = N_T = 0$. The NEDT is calculated from this data set.

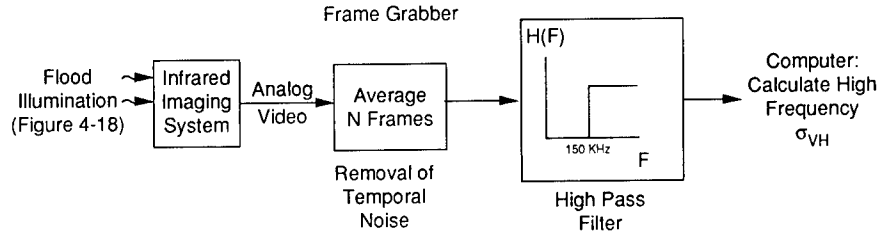


Fig. 4.22 Collecting data set N_{VH} with low-frequency and temporal noise removed. The FPN is calculated from this data set.

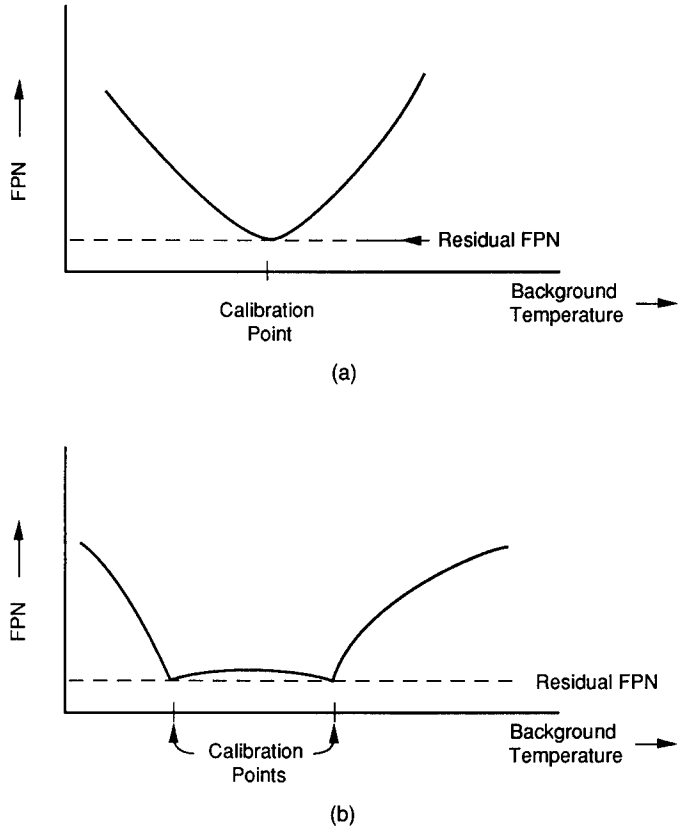


Fig. 4.23 FPN as a function of background temperature for systems with (a) single-point correction and (b) two-point correction.

with background temperature. A black cloth provides only one background temperature (ambient).

Nonuniformity and $1/f$ noise can be obtained with the same configurations shown in Figs. 4.20 to 4.22 except that the high-bandpass filter is replaced with a low-bandpass filter whose cutoff frequency is 150 kHz. As with FPN, nonuniformity is defined as the standard deviation of the resultant frame data. The amount of nonuniformity can be expressed as a percentage of the average value, or the NEDT. By using the NEDT reference, the nonuniformity percentage is independent of the system gain and level setting.

It is not possible to perform exhaustive testing, rather an inference is made about the population mean and variance based on a finite number of data points. As a result of the widespread usage of Gaussian statistics, it is almost always assumed that the data set is Gaussian distributed:

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{x - x_0}{\sigma_0}\right)^2 \right] . \tag{4.18}$$

The measured variance σ_i^2, an estimate of the true variance σ_0^2, is

$$\sigma_i^2 = \frac{N_e \sum_{i=1}^{N_e} x_i^2 - \left(\sum_{i=1}^{N_e} x_i \right)^2}{N_e(N_e - 1)} , \tag{4.19}$$

where N_e is given in Table 4.6. The measured mean x_i, an estimate of the true mean x_0, is

$$\overline{x_i} = \frac{\sum_{i=1}^{N_e} x_i}{N_e} . \tag{4.20}$$

From a statistical point of view, if the variance is measured K times, each measurement will provide a slightly different result because only a finite data set was analyzed each time. The best estimate of the variance is given by

$$\overline{\sigma^2} = \frac{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_K^2}{K} . \tag{4.21}$$

This assumes that the same number of data points were used to calculate each σ_i. Using this statistical approach, the best estimate of NEDT is

$$\text{NEDT} = \left. \frac{\overline{\sigma_{\text{TVH}}}}{\text{SiTF}} \right|_{f \geq 150 \text{ kHz}} . \tag{4.22}$$

The best estimate of the percentage FPN is

$$\text{FPN} = 100\% \left. \frac{(\overline{\sigma_{\text{VH}}}/\text{SiTF})}{\text{NEDT}} \right|_{f \geq 150 \text{ kHz}} . \tag{4.23}$$

The best estimate of the percentage nonuniformity is

$$\text{NU} = 100\% \left. \frac{(\overline{\sigma_{\text{VH}}}/\text{SiTF})}{\text{NEDT}} \right|_{f \leq 150 \text{ kHz}} . \tag{4.24}$$

Gaussian (or normal) distributions appear as straight lines when plotted on normal-probability graph paper.[19] The chi-squared goodness-of-fit methodology can also be used to determine if a sample data set follows a Gaussian distribution.[20] Although the normal-probability graph or chi-squared test provides a rigorous test for normality, a "quick look" may be adequate and can be obtained by creating a histogram of the data values and fitting a Gaussian curve to the histogram. In Fig. 4.24, the probability density function of a typical noise trace is plotted as a histogram. On the data is also plotted the assumed Gaussian distribution where the mean and standard deviation is calculated from Eqs. (4.19) and (4.20). If the distribution is not Gaussian, further de-

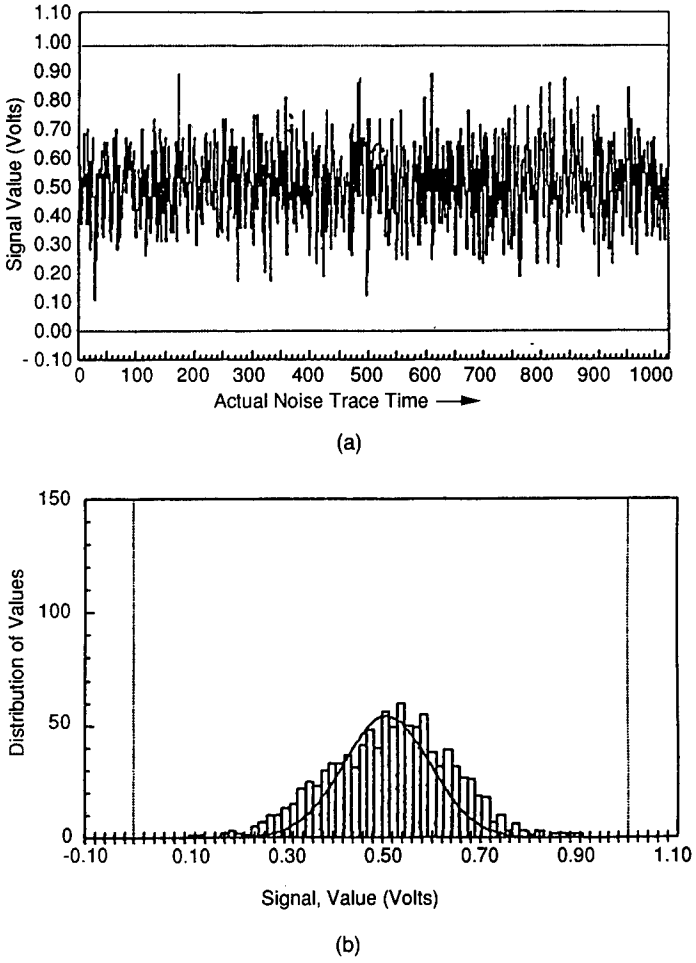


Fig. 4.24 Measured noise: (a) single noise trace and (b) histogram illustrating a Gaussian probability distribution.

scription is necessary for a complete characterization of the noise. Furthermore, if the data is not Gaussian distributed, the three-dimensional noise model must be reevaluated for the specific noise statistics present.

Because all the noise components are specified as an equivalent temperature differential, the measured value is dependent on the background temperature. Therefore, all specifications must include the measurement conditions (Table 4.8).

4.7 SLIT RESPONSE FUNCTION

The SRF is the transformation of input target angular subtence to output voltage for a fixed target intensity. The SRF is used to determine the imaging resolution. The test configuration is shown in Fig. 4.25. Because this test is a function of target angular subtence only, the source intensity is relatively

Table 4.8 Typical Specifications

NEDT	The NEDT shall not be greater than 0.2°C when the background is 23°C.
FPN	The FPN shall not be greater than 1% of the NEDT when the background is 23°C. The FPN shall not be greater than 5% of the NEDT over the entire environmental operational limits (−20 to +40°C).
Nonuniformity	Nonuniformity shall not be greater than 5% of the NEDT when the background is 23°C.
If the measurement is performed at another background temperature, a correction factor must be applied (system specific).	

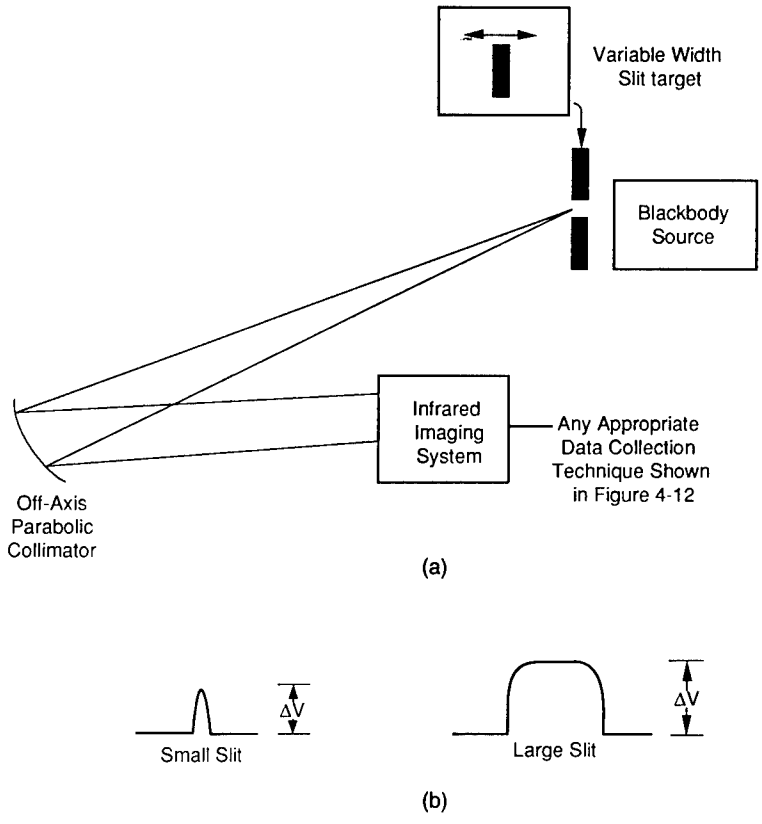


Fig. 4.25 Slit response function: (a) test configuration and (b) measured signals.

unimportant. It should be sufficiently high to provide a good SNR without entering a nonlinear region. It is best to measure the responsivity function prior to measuring the SRF to determine the maximum level allowable intensity without entering nonlinearity or saturation. Recommended target sizes range from 0.1 IFOV to 5 IFOV. The output is normalized to the maximum output that is achieved with a very large target so that the maximum value of the SRF is unity (Fig. 4.2). This value should be the same value as that obtained from the responsivity function for that input intensity. The test procedure is as follows:

1. Select a source intensity that maximizes the SNR without entering a nonlinear region or saturation. This can be determined by examining the responsivity function.
2. Critically align the target onto a detector to achieve maximum output.
3. Record the voltage difference between the target and its background.
4. Repeat steps 2 and 3 for each target size.
5. Divide all outputs by the maximum output. Maximum output should be obtained with the largest target used and should be the same output as that obtained with the SiTF test target when the source is at the same temperature.
6. Fully document any test abnormality and all results. The SRF should be presented both graphically and in tabular form.
7. The imaging resolution is that target angular subtence at which SRF = 0.5.

4.8 MODULATION TRANSFER FUNCTION

Methods to measure the MTF depend on both optical and electronic signal considerations. For convenience, the MTF is measured in two orthogonal axes (usually coincident with the array axes) to obtain two one-dimensional MTFs. The one-dimensional MTF is the Fourier transform of the LSF, where the LSF is the resultant waveform produced by the infrared imaging system when it is viewing an ideal line. Because an ideal line can be obtained by differentiating a step, the MTF can also be obtained from a step response (also known as the edge response and knife-edge response). The test configuration is illustrated in Fig. 4.26.

The ideal line does not have any finite extent and the slit is a practical implementation of the idealized line. As the slit becomes narrower, the flux passing through it diminishes and a point is reached where the slit intensity can no longer be sensed. Therefore a finite slit is required. The angular subtense of the slit should be smaller than the IFOV with a value less than 0.1 IFOV recommended. A thin straight heated nichrome wire can be used as an approximation to a slit.

The advantages in using a knife-edge target are that the target is simpler to construct than a narrow slit and there is no MTF correction as required when employing a slit. The edge is differentiated to obtain the LSF and then

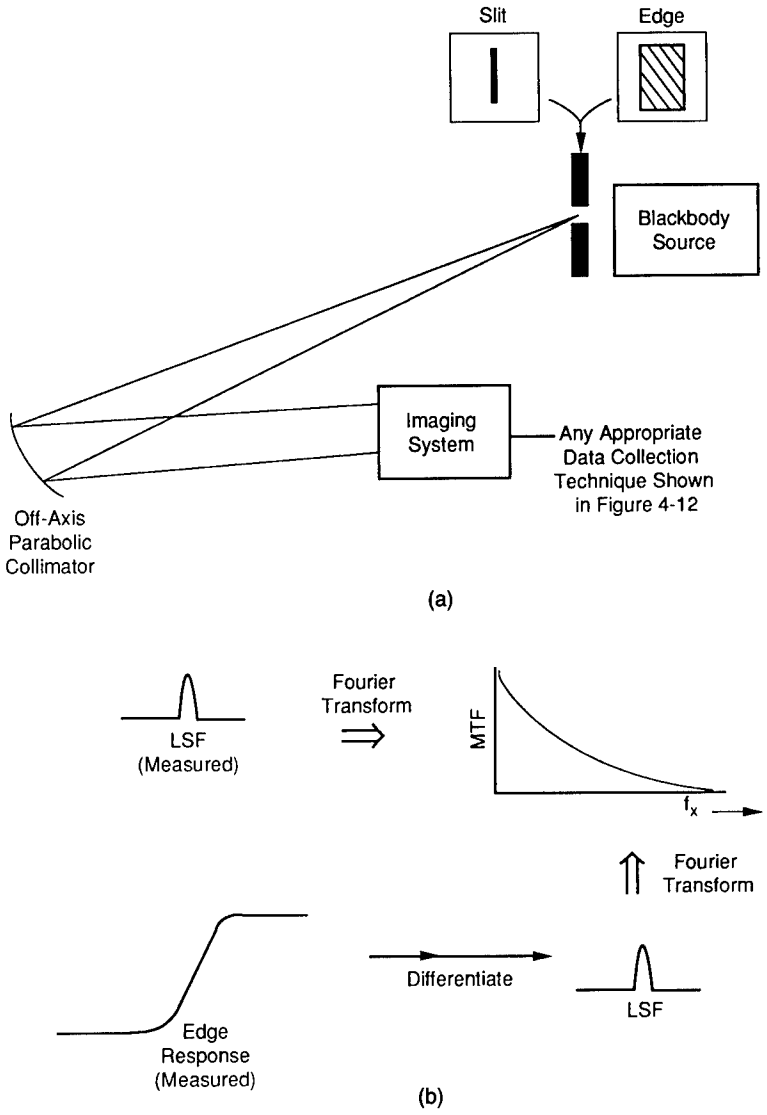


Fig. 4.26 Modulation transfer function: (a) test configuration and (b) methodology to obtain the MTF.

Fourier transformed. However, for noisy systems, differentiation accentuates the noise and this corrupts the resultant MTF.

MTF measurements, although they appear straightforward, may be difficult to perform as a result of digitization, phasing, background removal, jitter, noise, and improper normalization. The magnitude of each of these effects depends on the infrared imaging system design.

4.8.1 Digitization and Phasing Effects

The effects of sampling on the MTF can be dramatic.[21,22] Sampling is an inherent feature of all electronic imaging systems. The scene is sampled in either one or both directions, depending on the scanning mechanism and discrete nature of the detector elements. MTF theory is based on isoplanatism and electronic linearity. Sampled data systems are neither isoplanatic nor completely linear, and the MTF can have a range of values, depending on the phase between the internal sampler and the location of the target. It is appropriate to adjust the location of the target such that the highest possible MTF is obtained. This provides the best performance of the system.

The analog signal (containing the LSF) is digitized prior to performing the Fourier transform. The number of samples across the LSF should be sufficiently high so that phasing effects do not introduce any additional artifacts. To minimize these artifacts, at least four samples across the LSF at full width at half maximum (FWHM) (50%) amplitude are required. This corresponds approximately to 10 samples across the entire LSF. This is consistent with the six samples across the edge spread function (full width of the LSF) recommended by Granger[23] and the eight samples FWHM recommended by Dainty and Shaw.[24]

For scanning systems, a frame grabber is not appropriate for horizontal MTF measurements unless it can capture four samples per FWHM of the line spread function. For staring systems, the frame grabber may be synchronized with the array so that the frame grabber collects data from every available pixel. In the vertical direction, a frame grabber typically collects all the data. For example, RS 170 consists of 480 lines and frame grabbers are usually designed to capture all of the lines.

Because staring arrays are undersampled, special precautions must be used.[25,26] The LSF can be constructed by two methods. The first is to move a slit in subpixel increments across the detector and measure the response at each location. The second is to create a periodic array of point sources whose locations are at noninteger locations with respect to the detector locations. This noninteger arrangement is equivalent to varying the phase between the point sources and the detectors. To perform this test, each point source size and relative location must be known precisely. Each response must be normalized according to its point source intensity, which is proportional to the point source size. The complete LSF can be reconstructed by overlaying the various responses according to each location[27] (Fig. 4.27). A one-dimensional response can be obtained using slits (Fig. 4.28). The slit width must be known accurately to account for the differences in radiance, which results in different amplitudes. The slits should be sufficiently separated to ensure that the individual LSFs do not overlap. Care must be exercised in defining the location of the data points in time so that the Fourier transform is done properly. Because most Fourier transform algorithms require equispaced data points, it may be necessary to interpolate between the data points.

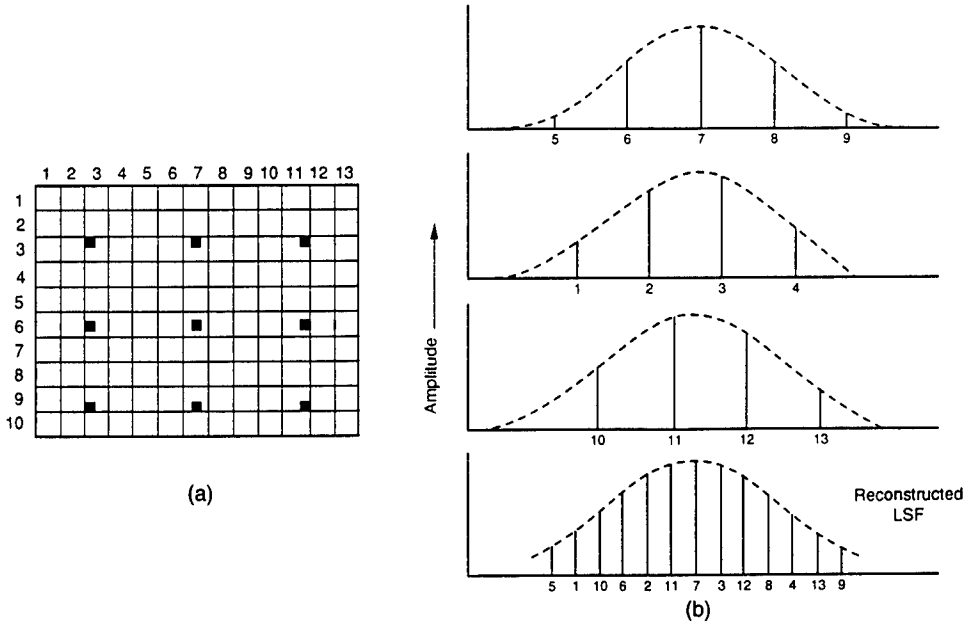


Fig. 4.27 Reconstruction of the LSF from a periodic array of point sources: (a) point sources imaged onto a detector array and (b) overlaid responses using data obtained on row 6.

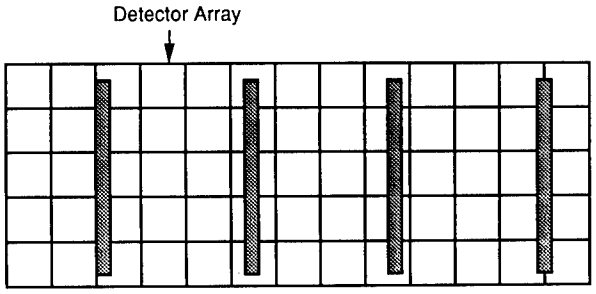


Fig. 4.28 The LSF can be reconstructed from a periodic array of slits. The data are overlaid as illustrated in Fig. 4.27(b).

4.8.2 Background Removal

Before calculating the Fourier transform, the background must be removed. This is done either by making a separate measurement of the background or by assuming the background is uniform. If improperly removed, the calculated MTF may be too high (too much background removed) or too low (insufficient background removed) (Fig. 4.29).

Background trends resulting from target inhomogeneities, shading, nonuniformity, and $1/f$ noise must be removed prior to performing the Fourier transform. Very low frequency $1/f$ noise can manifest itself as drift (slow variations in background level). Medium $1/f$ noise and shading appear as a trend

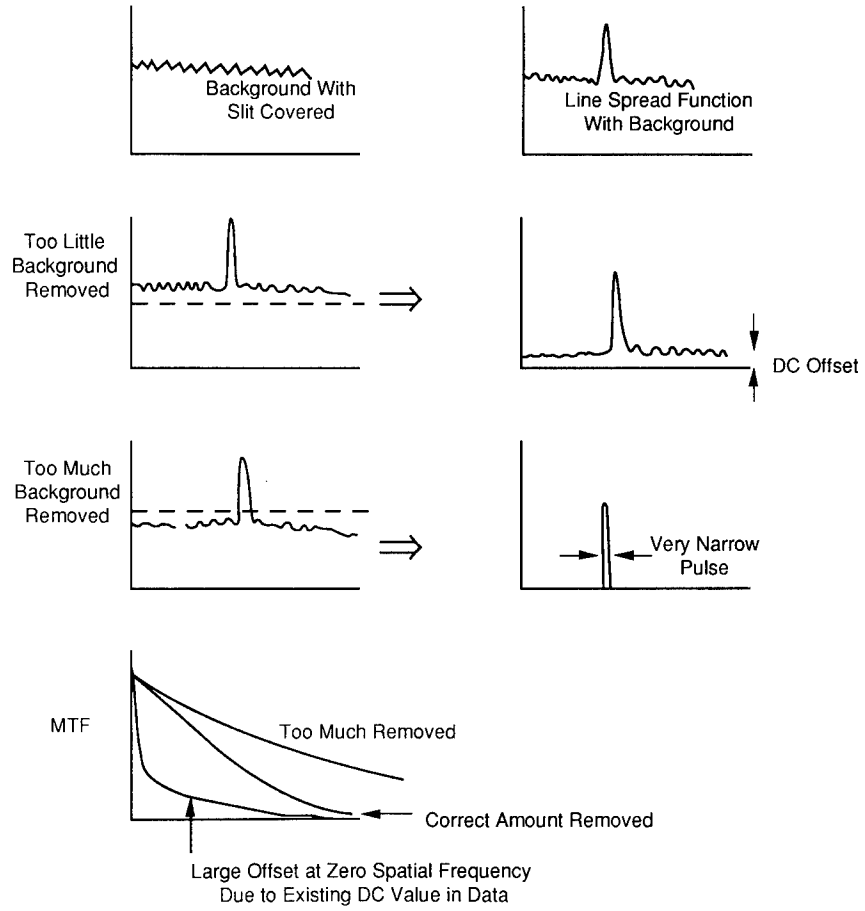


Fig. 4.29 Effects of the background removal on the MTF.

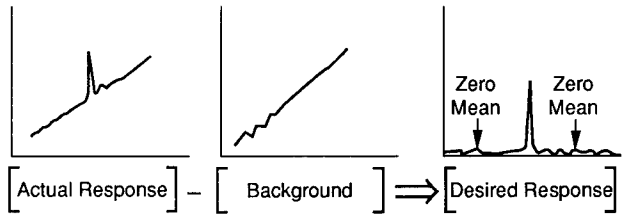


Fig. 4.30 Trend removal to obtain the LSF.

on the data (monotonically increasing or decreasing base line) (Fig. 4.30). Trends[15] can be removed by applying a first-order polynomial to the background data and subtracting this line from the LSF. The resultant trace should have zero mean on either side of the LSF response. Since the LSF is a high-frequency response, the data can be passed through a high-pass filter, such as the filter used to separate the high and low noise frequencies.

4.8.3 Jitter

With jitter (Fig. 4.31), the position of the LSF peak varies from frame to frame. The jitter may be a result of movement of the system relative to the target during the test (mechanical vibration) or inherent jitter in the system. If the jitter results from synchronization problems in the transient recorder, then the potential MTF degradation is a result of the measurement technique. Mechanical vibration can be minimized by placing the source, target, collimator, and infrared imaging system on a vibration isolation table.

If several frames are averaged together to improve the SNR, jitter in successive frames can cause broadening of the averaged line spread function and produce an MTF that is lower than theoretically possible. It may be desirable to leave the jitter in the measurements if it is a property of the system. The reason is that the observer sees the entire system response including any inherent jitter. The eye, and to some degree the display, averages system jitter. The removal of jitter yields the best possible system MTF but may not be representative of what the observer sees.

One way to align the traces is to integrate the individual LSF traces to obtain the edge spread function. The slope of the edge response is calculated via a least-squares approach [Eq. (4.12)]. The 50% point of the step response is the center of the step response and its location corresponds (assuming a symmetrical LSF) to the center of the LSF (Fig. 4.32). By matching the 50%

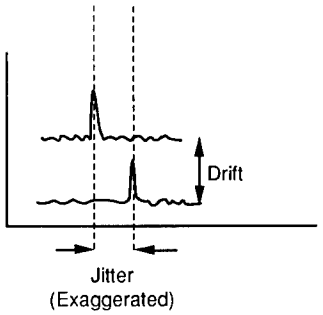


Fig. 4.31 LSF with jitter and drift.

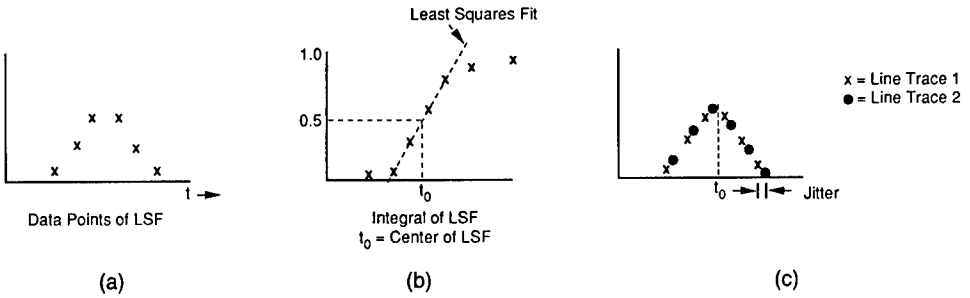


Fig. 4.32 Overlaying LSFs to remove jitter: (a) individual LSF, (b) determining center of LSF from the edge spread function, and (c) overlaid LSFs.

points and overlaying traces, the LSF is effectively filled with additional points. This increases resolution and increases the accuracy in obtaining the LSF. Care must be exercized in defining the location of the data points in time so that the Fourier transform is done properly. Because many Fourier transform algorithms require equispaced data points, it may be necessary to interpolate between the data points.

4.8.4 Noise

Noise introduces random errors and adds a positive bias to the MTF. The MTF in the presence of noise is always greater than that without noise. It has been reported that an SNR of 500 may be required for accurate MTF measurements.[28] To reduce noise, choose a high source intensity and, if possible, a low system gain. To ensure that the system is operating in a linear region, obtain the MTF for a variety of source intensities. Select the highest intensity for which the MTF is still well behaved. The noise bias can be removed by appropriate time averaging. Because of jitter and drift, the individual line spread functions vary in time and absolute voltage level and averaging must be done with care.

4.8.5 Fourier Transform Considerations

Because Fourier transform software is readily available, the actual computation is not discussed, but rather, we address the use of data manipulation techniques to achieve the desired results. The fast Fourier transform (FFT) does not denote a specific method of calculating the discrete Fourier transform but rather denotes a computationally efficient method of implementing a discrete transform. When operating on $2N$ data points, the Fourier transform provides real $\mathcal{R}$ and imaginary $\mathcal{I}$ parts of the complex transfer function, each containing N data points. The MTF for data point n is

$$\text{MTF}(n) = [\mathcal{R}^2(n) + \mathcal{I}^2(n)]^{1/2} \quad n = 1, 2, \ldots, N \ . \tag{4.25}$$

Each data point n is associated with a frequency given by

$$f_n = \frac{nf_s}{2N} \ , \tag{4.26}$$

where f_s is the sampling rate that digitized the LSF. For example, if the analog signal is digitized every 20 ns, then $f_s = 1/(20\text{ ns}) = 5$ MHz. The frequency resolution is $\Delta f = f_s/(2N)$. If the LSF and its background is digitized into 1024 samples, then $\Delta f = 4.88$ MHz. The maximum frequency (Nyquist criterion) that can be recovered is

$$f_{\max} = \frac{f_s}{2} \ . \tag{4.27}$$

Using a finite-length data record violates the Fourier transform requirement of continuity for all time. Discontinuities of partial cycles at the beginning or

end of the sample intervals cause errors. The errors (also called *sidelobes* or *leakage*) may mask small amplitude frequencies that are present. To minimize leakage, the data is often passed through a filter or window.[29] The most popular filter is the Hanning or raised cosine window. In general, the choice of windows requires knowledge of the data to be collected. The window also attenuates low-frequency components of the MTF. The MTF should be calculated with and without a window and the results compared. If the LSF is centered on the data set, the window is probably not necessary.

Noise introduces an error in the power estimation that is independent of the record length. Collecting more data points does not affect the magnitude of the error.[30] To reduce this error, multiple power spectra are computed and then are averaged at each frequency component to obtain a composite spectra:

$$\begin{aligned} P_1(f) &= \mathcal{R}_1^2(f) + \mathcal{I}_1^2(f) \ , \\ P_2(f) &= \mathcal{R}_2^2(f) + \mathcal{I}_2^2(f) \ , \\ &\vdots \qquad\quad \vdots \qquad\quad \vdots \\ P_k(f) &= \mathcal{R}_k^2(f) + \mathcal{I}_k^2(f) \ . \end{aligned} \tag{4.28}$$

The composite spectra is

$$\overline{P(f)} = \frac{\sum_{i=1}^{k} P_i(f)}{k} \ , \tag{4.29}$$

and the average MTF is the square root of the averaged power spectra:

$$\overline{\mathrm{MTF}(f)} = [\overline{P(f)}]^{1/2} \ . \tag{4.30}$$

4.8.6 Normalization

For passive linear systems, the MTF is normalized to unity at zero spatial frequency and it decreases as the frequency increases. However, image enhancement techniques and boost circuitry may increase the MTF at certain frequencies. With ac coupled systems, the dc (zero spatial frequency) component is suppressed and thus prevents normalization at zero. The problem can be avoided by normalizing at a spatial frequency one decade (10 times) above the ac coupling cut on frequency. Then the MTF curve appears to approach unity at zero spatial frequency and the increased MTF at the boost frequency is a reminder that the system contains peaking components. Note that the ac coupling cut-on frequency may be below the spatial resolution Δf of the Fourier transform and therefore may not be measured.

If the LSF is measured on the analog video (time domain), the FFT transforms the data into the electrical frequency domain (Hertz). The conversion of electrical frequency into spatial frequency requires knowledge of the field of view and the active time (TV) for one TV line:

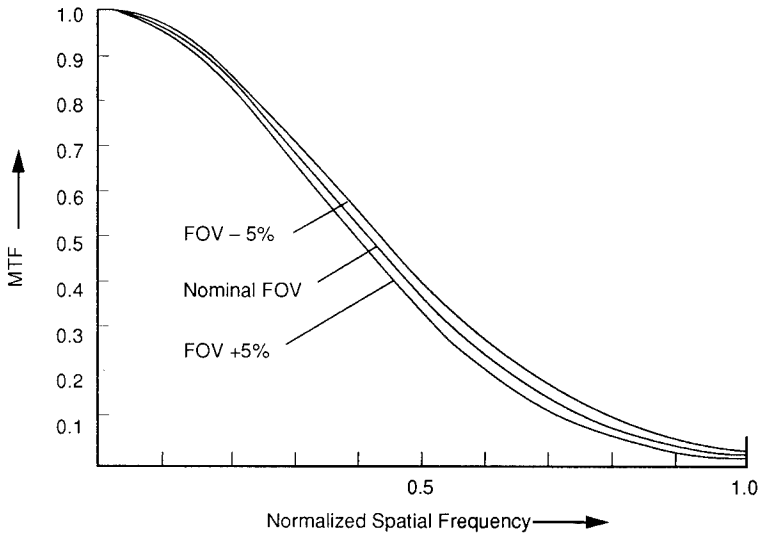


Fig. 4.33 Apparent MTF variations resulting from variations in the FOV. The variations are due to uncertainty in the scale factor.

$$f_x = \frac{\text{TV}}{\text{FOV}} f_{\text{Hz}} \ . \tag{4.31}$$

If the infrared imaging system operates in the digital domain and the clock rate is R_{clock} and the number of pixels is P per line, then

$$\text{TV} = \frac{P}{R_{\text{clock}}} \ . \tag{4.32}$$

The field of view should not be assumed but be measured. A 5% variation in FOV during manufacturing is not uncommon. This manifests itself approximately as a 5% variation in the scale factor. This, in turn, appears to shift the MTF curve such that there may be up to a 10% variation in the MTF values (Fig. 4.33).

The system MTF is the average MTF divided by the MTF of the test configuration. When using a slit, the MTF correction is $\sin(\pi\alpha f_x)/(\pi\alpha f_x)$, where α is the slit angular subtence measured in object space.

4.8.7 MTF Summary

MTF measurements, although straightforward, can be difficult to perform. Possible MTF variations are shown in Fig. 4.34 with possible causes given in Table 4.9.

Table 4.9 MTF Variations

MTF Variations	Possible Causes
Suppressed at zero spatial frequency [Fig. 4.34(a)]	ac coupling
Peak at low frequency [Fig. 4.34(b)]	Trends not removed
Peak in curve [Fig. 4.34(c)]	Boost circuitry, nonlinear image processing
Too low [Fig. 4.34(d)]	System out of focus System in nonlinear region Jitter broadened LSF (when averaging) Wrong frequency scale factor Insufficient samples on LSF
Too high [Fig. 4.34(e)]	Too much background removed Wrong frequency scale factor Noise bias Insufficient samples on LSF
Low-frequency components [Fig. 4.34(f)]	Insufficient background removed
Not reproducible	Variations in all of the above

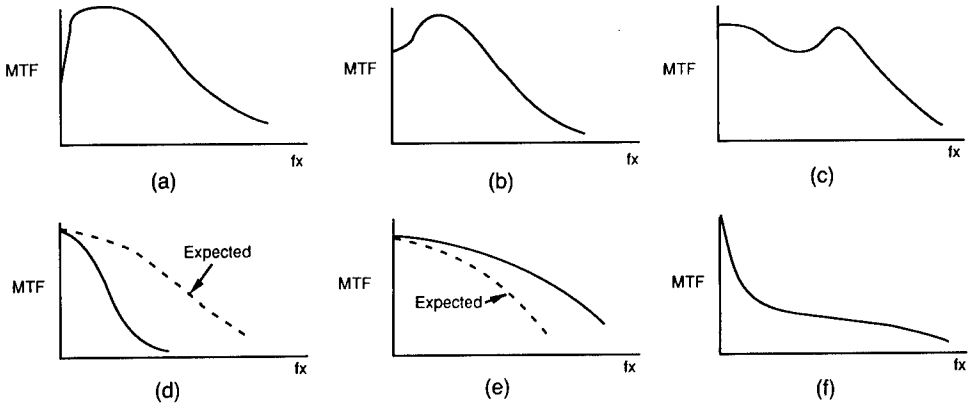


Fig. 4.34 Variations in the MTF (see Table 4.9 for explanation).

4.9 CONTRAST TRANSFER FUNCTION

With a large number of bar targets available and ease of fabrication, the CTF appears easy to measure and represents a convenient method to calculate the MTF. A square wave can be expressed as a series of sinusoids. The amplitude of the square wave at frequency f_x is as an infinite sum of sinusoid amplitudes:

$$\mathrm{CTF}(f_x) = \frac{4}{\pi}\left|\mathrm{MTF}(f_x) - \frac{\mathrm{MTF}(3f_x)}{3} + \frac{\mathrm{MTF}(5f_x)}{5} - \cdots\right|, \tag{4.33}$$

or

$$\mathrm{CTF}(f_x) = \frac{4}{\pi}\left|\sum_{k=1}^{\infty}(-1)^k \frac{\mathrm{MTF}[(2k+1)f_x]}{2k+1}\right|. \tag{4.34}$$

Conversely, the sine wave amplitude at frequency f can be expressed as an infinite sum of square-wave amplitudes[31,32]:

$$\mathrm{MTF}(f_x) = \frac{\pi}{4}\left|\mathrm{CTF}(f_x) + \frac{\mathrm{CTF}(3f_x)}{3} - \frac{\mathrm{CTF}(5f_x)}{5} + \frac{\mathrm{CTF}(7f_x)}{7} + \frac{\mathrm{CTF}(11f_x)}{11} + \text{irregular terms}\right| , \tag{4.35}$$

or

$$\mathrm{MTF}(f_x) = \frac{\pi}{4}\left|\sum_{k=0}^{\infty} B_k \frac{\mathrm{CTF}(kf_x)}{k}\right| , \tag{4.36}$$

where

$$B_k = (-1)^m(-1)^{(k-1)/2} \quad \text{for } r = m , \tag{4.37}$$

and $B_k = 0$ for $r < m$, $k = 1, 3, 5, \ldots$ (odd only); m is the total number of prime factors in k, and r is the number of different prime factors in k. The relationship between square-wave and sinusoidal amplitude responses is illustrated in Fig. 4.35.

A typical CTF test configuration is shown in Fig. 4.36. The CTF response is the difference between the maximum and minimum output value. Because of phasing effects,[33] the maximum and minimum values may not be adjacent to each other (Fig. 4.37). As a result, a target with many bars is required. If the data is captured by a digital recording device (frame grabber or transient recorder), its sampling rate should be high compared to the infrared imaging system sampling rate to avoid any signal degradation. The CTF must be normalized to unity at zero spatial frequency. The "zero" spatial frequency value can be determined by measuring the response to a very large target.

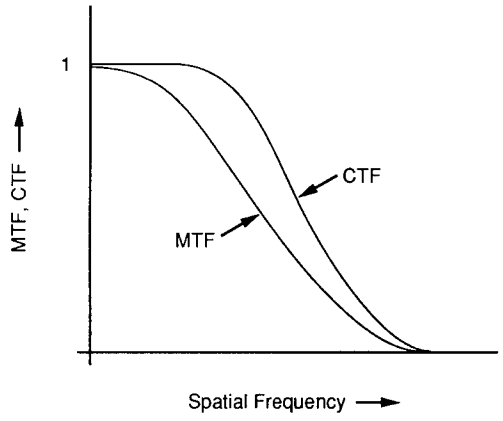


Fig. 4.35 Relationship between CTF and MTF.

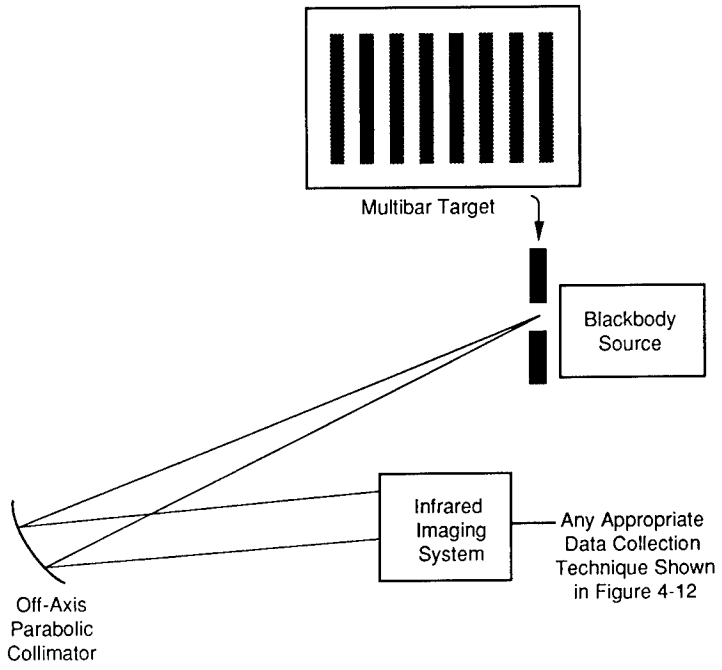


Fig. 4.36 CTF test configuration.

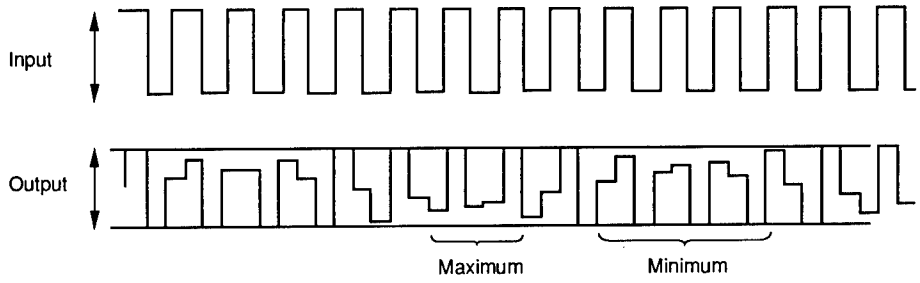


Fig. 4.37 Effects of phasing on output signal intensity.

To obtain the MTF, many square-wave responses must be obtained. The maximum number is limited by the system cutoff. For targets whose spatial frequency are above 1/3 system cutoff, the MTF is equal to $\pi/4$ times the measured CTF (i.e., only one target is necessary to provide the MTF). It is often difficult to determine exactly where cutoff is and, as a result, difficult to estimate where 1/3 cutoff is located. Unless there is evidence otherwise, the optical spatial frequency cutoff should be used as the system cutoff for purposes of calculating the MTF. For diffraction-limited optics, the optical spatial frequency cutoff is D_{sys}/λ, where λ is the system's spectral responsivity average wavelength.

The CTF data must be clearly labeled as the square-wave response, and, for completeness, the calculated MTF should also be shown on the same graph and appropriately labeled.

Example 3 MTF Calculations from Square-Wave Responses. An infrared imaging system has a cutoff at 11 c/mrad. The CTF is measured at 1, 3, 5, 7, and 9 c/mrad (Table 4.10). What is the corresponding MTF?

$$\mathrm{MTF}(1) = \frac{\pi}{4}\left[\mathrm{CTF}(1) + \frac{\mathrm{CTF}(3)}{3} - \frac{\mathrm{CTF}(5)}{5} + \frac{\mathrm{CTF}(7)}{7}\right]$$

$$\mathrm{MTF}(3) = \frac{\pi}{4}\left[\mathrm{CTF}(3) + \frac{\mathrm{CTF}(9)}{3}\right]$$

$$\mathrm{MTF}(5) = \frac{\pi}{4}\mathrm{CTF}(5)\ , \qquad \mathrm{MTF}(7) = \frac{\pi}{4}\mathrm{CTF}(7)\ , \qquad \mathrm{MTF}(9) = \frac{\pi}{4}\mathrm{CTF}(9)\ .$$

The relationship described by Eq. (4.35) was developed for an analog optical system. For sampled data systems, the mathematical relationship between MTF and CTF has not yet been developed. Therefore, Eq. (4.35) should be used cautiously for sampled data systems, if at all. CTF measurements are appropriate system performance validation. It is the conversion to the MTF that is challenged.

4.10 MINIMUM RESOLVABLE TEMPERATURE

MRT is a subjective measure of image quality. MRT is a measure of an observer's ability to resolve a four-bar pattern embedded in noise. It is a laboratory summary measure that combines visual sensitivity and resolution. The results depend on decisions made by an observer and the results vary with training, motivation, and visual capacity. Because of the large inter- and intraobserver variability, several observers are required.

Because each person is different, individual characteristics can affect the way people behave. Age, experience, personality, vision, mood, and intelligence are just a few. During the course of an evaluation, many changes take place that could influence the outcome of a test. Equipment may show signs of wear and may even become damaged (e.g., fingerprints on targets). People learn new methods, have new experiences, change moods, and become fatigued. Because so many changes can occur and because they, in general, cannot be controlled, the consistency from one session to another can be severely compromised. Variabilities as high as 50% are often cited in the measured MRT from laboratory to laboratory with 20% variability reported at any one laboratory.[34]

The eye's detection capability depends on the visual angle subtended by the image size and the distance from the monitor to the observer. Although the eye's contrast threshold is usually reported as J-shaped, the actual shape is very dependent on the noise power spectral density.[35–37] The observer's ability to see a specific spatial frequency target depends on the noise content in the neighborhood of that spatial frequency (Fig. 4.38). Low-spatial-frequency noise components interfere with detecting low-frequency targets, mid-spatial-frequency

Table 4.10 Measured CTF and Calculated MTF

Spatial Frequency	CTF	MTF(1)	MTF(3)	MTF(5)	MTF(7)	MTF(9)
1	0.97	CTF(1) = 0.970	—	—	—	—
3	0.83	CTF(3)/3 = 0.276	CTF(3) = 0.830	—	—	—
5	0.55	−CTF(5)/5 = −0.110	—	CTF(5) = 0.550	—	—
7	0.32	CTF(7)/7 = 0.046	—	—	CTF(7) = 0.320	—
9	0.10	—	CTF(9)/3 = 0.033	—	—	CTF(9) = 0.110
Sum		1.182	0.863	0.550	0.320	0.110
MTF = $\pi/4$ · sum		0.928	0.677	0.432	0.251	0.086

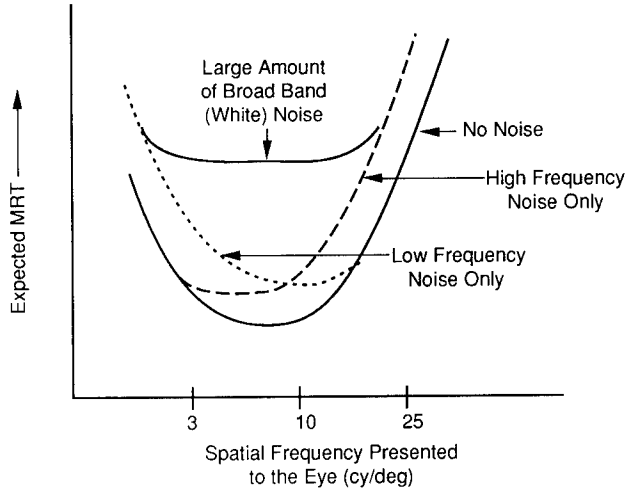


Fig. 4.38 Expected MRT as a function of noise.

noise increases the contrast threshold curve at mid-frequencies, and so on. When considerable white noise is present, the eye's contrast threshold curve becomes nearly flat from 0.5 to 20 cycles/deg.

In the laboratory, the distance to the monitor is not usually specified nor limited in any way. To maximize detection capability (stay on the minimum of the contrast sensitivity curve), an observer subconsciously moves closer to the monitor to perceive small targets (high spatial frequencies) and further away to see larger targets (low spatial frequencies). With head movement, the observer will report his lowest possible threshold for all spatial frequencies. On the other hand, if the infrared imaging system is designed for a specific application in which the observer sits a fixed distance away from the monitor, then it is appropriate to restrain head movement to those distances that are commensurate with the operational environment. Therefore, two measurement conditions are possible: (1) allowing the observer to move his head and (2) fixing the head in space. Head movement may yield different MRT values than if the head is fixed, depending on the amount of nonuniformity (low-frequency noise) present.

When uniform white noise is present, the MRT can approach zero as the distance to the monitor increases.[1] The phenomena by which the MRT can approach zero as the spatial frequency approaches zero is similar to the blending of ink dots in a half-tone image. With printing, there are no grays but rather ink dots of varying sizes that are below the eye's resolution. The eye blends the dots into a gray level. Similarly, if the noise had no significant low-frequency components, as the observer moved farther and farther away, the individual noise spikes would eventually be below the eye's resolution and the eye would blend the image into a uniform gray. Because the observer can optimize both the system and display, the MRT can approach zero. However, performance models[16,38,39] were developed for a noise-limited condition. Moving a substantial distance from the display violates the model assumptions. Although MRT may approach zero for extremely large viewing distances (in

excess of 15 ft), these responses may not be considered representative of actual system usage. The MRT low-frequency asymptote appears to depend on the amount of low-frequency noise present. There does not appear to be a universal scaling factor. However, many systems appear to have modest low-frequency noise such that the MRT asymptotes[2,3] somewhere between 0.3 and 0.7 times the NEDT.

Statistically speaking, at threshold, the target is detected 50% of the time by one observer or, equivalently, detected by 50% of the population. It is the variation of thresholds (frequency of seeing response) that accounts for the variation in detection values. The broader the frequency of seeing curve, the larger the variation in individual responses. Observers are considered "consistent" if their threshold value does not change on a periodic basis and if their frequency of seeing curve is "narrow." To appropriately average threshold responses, it is necessary to understand the underlying frequency of seeing distribution. Holst and Pickard[40] reported an experiment in which 76 observers detected standard four-bar targets of various spatial frequencies embedded in noise. It was determined that their 2700 detection responses could be mathematically described with a log-normal distribution.[a] With an underlying log-normal distribution, the average of N observations is the geometric average of the individual MRT_i values:

$$\mathrm{MRT}_{\mathrm{ave}} = \left(\prod_{i=1}^{N} \mathrm{MRT}_i \right)^{1/N} . \tag{4.38}$$

Infrared imaging systems are subject to sampling effects. The MRT does not have a unique value for each spatial frequency but rather has a range of values depending on the location of the targets with respect to the sampling lattice. As a result of phasing effects, it has become widely accepted to "peak up" the targets, that is, to adjust the targets to achieve the best visibility of the targets. It is important, however, that the observer count the number of bars to ensure that the required number are present. Phasing effects dramatically affect the appearance of the bars such that some bars appear with different widths or intensities. In the region of 0.6 to 0.8 times Nyquist frequency, no matter how the phase is adjusted, the bars do not appear equal in width or intensity (see Fig. 4.37). As a result the MRT is higher in this region[41] (Fig. 4.39). In apparent contradiction to sampling theory, it is possible to perceive four bar targets whose spatial frequencies are about 10% higher than the Nyquist spatial frequency. If an infinite number of bars were presented, then it would be obvious that the input spatial frequency was aliased down to a lower spatial frequency. But this is not obvious when only a few bars are

[a] A log-normal distribution appears as the usual Gaussian-shaped distribution when plotted on a logarithmic axis. Over a limited region, it can be approximated with a linear Guassian distribution with the same mean. The standard deviations are not related. The log-normal distribution detection threshold follows:

$$P(\mathrm{MRT}) = \frac{1}{\sqrt{2\pi}\,\log(\sigma)} \exp\left\{ -\frac{1}{2}\left[\frac{\log(\mathrm{MRT}) - \log(\overline{\mathrm{MRT}})}{\log(\sigma)} \right]^2 \right\} ,$$

where $\sigma = 1.58$.

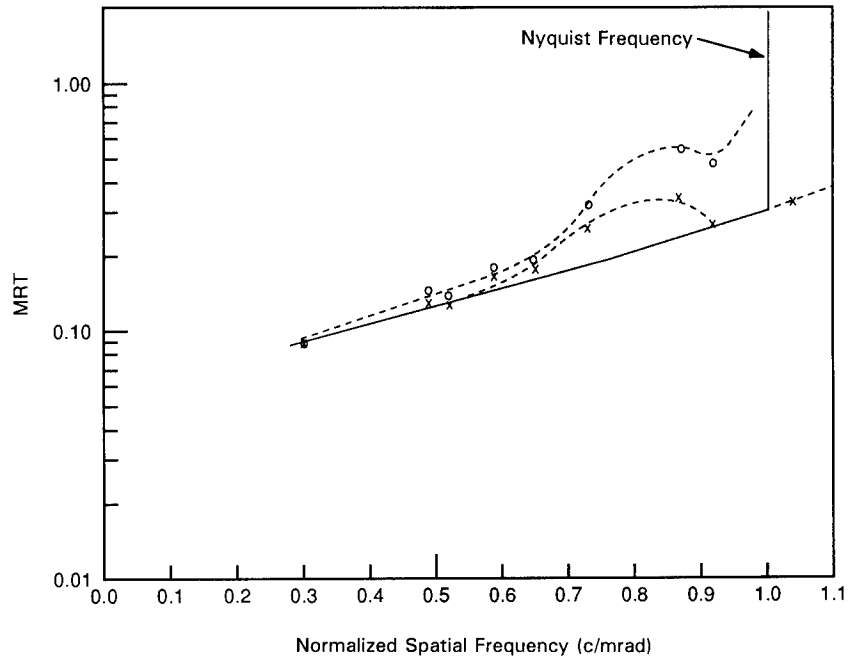


Fig. 4.39 MRT for a staring array. The anticipated response is shown with a solid line. The dashed lines are the experimentally obtained values for in-phase (crosses) and out-of-phase (circles) targets.[41]

presented. It is possible to select a phase such that a four-bar target above Nyquist can be reproduced with reasonable fidelity.

The test configuration is shown in Fig. 4.40. Each bar of the four-bar target has a 7:1 aspect ratio. The observer is allowed unlimited viewing time and can continually adjust the system gain and level to optimize the image for his detection criterion. Likewise, the observer can adjust the monitor contrast and brightness to maximize detection. Usually the observer adjusts the monitor to a relatively low brightness level and a relatively high contrast level to make the image rather noisy. The observer is permitted to readjust the system and the monitor at any time. The ambient lighting should approximately match the monitor luminance. The observer must be allowed sufficient time to dark adapt to the reduced ambient lighting before proceeding with the test. It is important that the observer not be influenced by the environment. This includes extraneous light sources, excessive noise, and other people in the room.

An offset between the actual temperature differential and the measured temperature differential may exist. This occurs when the SiTF does not pass through the origin. This offset is removed from the final MRT value by obtaining the MRT for both positive and negative contrast targets and averaging the results together.

The entire threshold detection versus spatial frequency curve should be measured. Because a discrete target set is used, the location at which the MRT asymptotes to infinity may not be measured. The curve may asymptote between

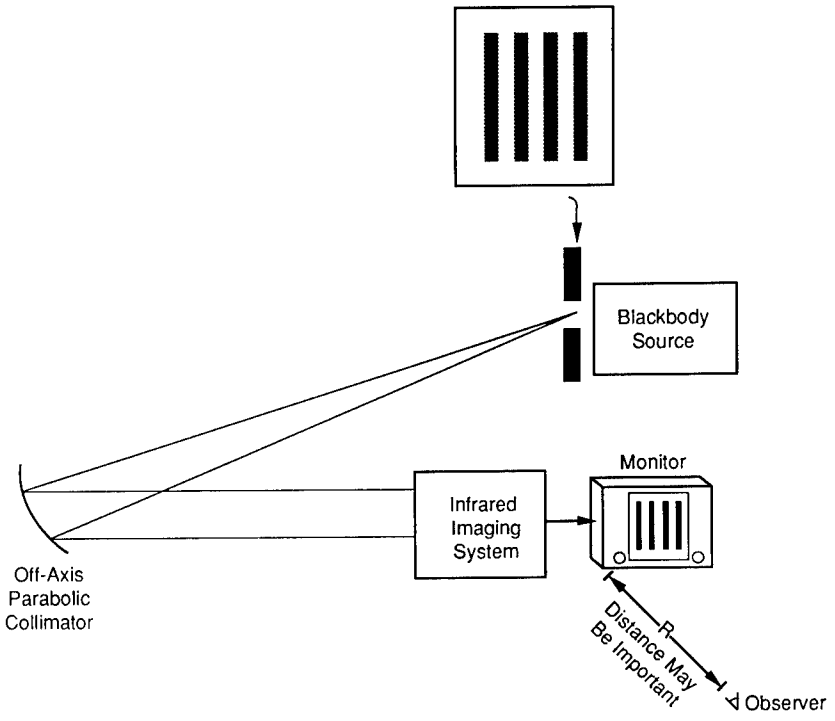


Fig. 4.40 MRT test configuration.

the last resolvable target and the next available target. It is therefore imperative that the first target that is not resolvable be recorded as CNR (cannot resolve). The MRT test procedure follows:

1. Position the target with the bars oriented vertically to obtain the horizontal MRT or horizontally to measure the vertical response.
2. Adjust target phase for maximum visibility.
3. Verify four bars are visible by counting them.
4. Establish a positive sub-MRT temperature differential and slowly increment the blackbody temperature differential.
5. Allow the observer to continually adjust system and monitor controls to optimize the image.
6. Record the temperature differential at which the observer can detect all four bars 50% of the time.
7. Establish a negative sub-MRT temperature differential and slowly decrement the blackbody temperature differential.
8. Allow the observer to continually adjust system and monitor controls to optimize the image.
9. Record the temperature differential at which the observer can detect all four bars 50% of the time.
10. Average the absolute values of positive and negative temperature differential recordings to obtain the MRT.

11. Repeat for other spatial frequencies.
12. If a target cannot be resolved, record CNR.
13. Multiply all observations by the average collimator transmittance and average atmospheric transmittance to determine the effective source intensity at the entrance aperture of the infrared imaging system.
14. Geometrically average the individual observer responses.

MRT testing takes a relatively long time (up to several hours) depending on the number of observers and the number of different spatial frequencies selected. Ambient temperature conditions can often change over this long period and therefore these measurements are particularly susceptible to ambient temperature changes (Fig. 4.9). Because of the potentially large frequency of the seeing curve, it is important to state the number of observers required in a specification (Table 4.11).

Table 4.11 Typical Specifications

The MRT shall not be greater than 0.5°C (average of three observations) at 5 cycles/mrad when the ambient temperature is at 20°C. If the measurement is performed at another background temperature, a correction factor must be applied (system specific).

References

1. J. M. Mooney, "Effect of spatial noise on the minimum resolvable temperature of a staring array," *Applied Optics* **30**(23), 3324–3332 (1991).
2. G. C. Holst, "Minimum resolvable temperature predictions, test methodology and data analysis," in *Infrared Technology XV,* I. Spiro, Ed., *Proceedings of the SPIE* **1157,** 208–218 (1989).
3. C. M. Webb and G. C. Holst, "Observer variables in minimum resolvable temperature difference," in *Infrared Imaging Systems: Design, Analysis, Modeling and Testing III,* G. Holst, Ed., *Proceedings of the SPIE* **1689,** 356–357 (1993).
4. G. R. McNeill and H. B. Macurda, "Infrared (IR) calibration measurement requirements: development from system requirements," in *Infrared Systems,* P. Sanman, Ed., *Proceedings of the SPIE* **256,** 84–90 (1980).
5. W. J. Dixon and F. J. Massey, *Introduction to Statistical Analysis,* McGraw-Hill, New York (1957).
6. J. S. Bendat and A. G. Piersol, *Random Data: Analysis and Measurement Procedures,* 2nd ed., John Wiley & Sons, New York (1986).
7. J. D. Vincent, *Fundamentals of Infrared Detector Operation and Testing,* Sec. 5.3, John Wiley & Sons, New York (1989).
8. American Society of Heating, Refrigeration and Airconditioning Engineers, "Applications of infrared sensing devices to the assessment of building heat loss characteristics," ANSI/ASHRAE Standard 101-1981, ASHRAE, Atlanta, GA (1983).
9. Y. M. Chang and R. A. Grot, "Performances measurements of infrared imaging systems used to assess thermal anomalies," in *Thermal Imaging,* I. Abel, Ed., *Proceedings of the SPIE* **636,** 17–30 (1986).
10. G. B. McIntosh and A. F. Filippone, "Minimum resolvable temperature difference (MRTD) testing: equipment specifications for building performance diagnostics," in *Thermosense IV,* R. Grot and J. Wood, Eds., *Proceedings of the SPIE* **313,** 102–111 (1981).
11. W. J. Smith, "Optical design," Sec. 8.10.2 in *The Infrared Handbook,* W. L. Wolfe and G. J. Zissis, editors, rev. ed., Environmental Research Institute of Michigan, Ann Arbor, MI (1985).
12. S. J. Briggs, "Photometric technique for deriving a 'best gamma' for displays," *Optical Engineering,* **20**(4), 651–657 (1981).

13. J. D'Agostino and C. Webb, "3-D analysis framework and measurement methodology for imaging system noise," in *Infrared Imaging Systems: Design, Analysis, Modeling and Testing II,* G. Holst, Ed., *Proceedings of the SPIE* **1488,** 110–121 (1991).

14. E. F. Cross and T. M. Reese, "Figures of merit to characterize integrating image sensors: a ten year update," in *Infrared Technology XIV,* I. Spiro, Ed., *Proceedings of the SPIE* **972,** 195–206 (1988).

15. J. S. Bendat and A. G. Piersol, *Random Data: Analysis and Measurement Procedures,* 2nd ed., pp. 362–365, John Wiley & Sons, New York (1986).

16. J. Ratches, W. R. Lawson, L. P. Obert, R. J. Bergemann, T. W. Cassidy, and J. M. Swenson, *Night Vision Laboratory Static Performance Model for Thermal Viewing Systems,* U.S. Army Electronics Command Report ECOM 7043, Ft. Monmouth, NJ (1975).

17. M. Schwartz, *Information Transmission, Modulation, and Noise,* 2nd ed., Chap. 6, McGraw-Hill, New York (1970).

18. D. A. Scribner, M. R. Kruer, J. C. Gridley, and K. Sarkady, "Physical limitations to nonuniformity correction in IR focal plane arrays," in *Focal Plane Arrays Technology and Applications,* J. Chatard, Ed., *Proceedings of the SPIE* **865,** 185–202 (1987).

19. W. J. Dixon and F. J. Massey, *Introduction to Statistical Analysis,* p. 55, McGraw-Hill, New York (1957).

20. J. S. Bendat and A. G. Piersol, *Random Data: Analysis and Measurement Procedures,* 2nd ed., Sec. 4.6, John Wiley & Sons, New York (1986).

21. W. Wittenstein, J. C. Fontanella, A. R. Newberry, and J. Baars, "The definition of the OTF and the measurement of aliasing for sampled imaging systems," *Optica Acta* **29**(1), 41–50 (1982).

22. S. E. Reichenbach, S. K. Park, and R. Narayanswamy, "Characterizing digital image acquisition devices," *Optical Engineering* **30**(2), 170–177 (1991).

23. E. M. Granger, "Image Quality analysis and testing for infrared systems," SPIE Tutorial Short Course Notes, T60, Orlando, FL (Mar. 1989).

24. J. C. Dainty and R. Shaw, *Image Science,* p. 204, Academic Press, New York (1974).

25. S. J. Pruchnic, G. P. Mayott and P. A. Bell, "Measurement of optical transfer function of discretely sampled thermal imaging systems," in *Infrared Imaging Systems: Design, Analysis, Modeling and Testing III,* G. Holst, Ed., *Proceedings of the SPIE* **1689,** 368–378 (1993).

26. T. L. Williams and N. T. Davidson, "Measurement of the MTF of IR staring imaging systems," in *Infrared Imaging Systems: Design, Analysis, Modeling and Testing III,* G. Holst, Ed., *Proceedings of the SPIE* **1689,** 53–63 (1993).

27. R. F. Rauchmiller and R. A. Schowengerdt, "Measurement of the Landsat thematic mapper modulation transfer function using an array of point sources," *Optical Engineering* **27**(4), 334–343 (1988).

28. H. P. Stahl, "Infrared MTF measurements of optical systems," *Lasers and Optronics* **10**(4), 71–72 (1991).

29. F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE* **66**(1), 51–83 (1978).

30. J. S. Bendat and A. G. Piersol, *Random Data: Analysis and Measurement Procedures,* 2nd ed., Sec. 8.5.4, John Wiley & Sons, New York (1986).

31. J. W. Coltman, "The specification of imaging properties by response to a sine wave input," *Journal of the Optical Society of America* **44**(6), 468–471 (1954).

32. I. Limansky, "A new resolution chart for imaging systems," *The Electronic Engineer* **27**(6), 50–55 (June 1968).

33. T. S. Lomheim, L. W. Schumann, R. M. Shima, J. S. Thompson, and W. F. Woodward, "Electro-optical hardware considerations in measuring the imaging capability of scanned time-delay-and-integrate charge-coupled imagers," *Optical Engineering* **29**(8), 911–927 (1990).

34. C. W. Hoover, Jr., and C. M. Webb, "What is an MRT? And how do I get one," in *Infrared Imaging Systems: Design, Analysis, Modeling and Testing II,* G. Holst, Ed., *Proceedings of the SPIE* **1488,** 280–288 (1991).

35. S. Daly, "Application of a noise adaptive contrast sensitivity function in image data compression," *Optical Engineering* **29**(8), 977–987 (1990).

36. H. Pollehn and H. Roehrig, "Effect of noise on the modulation transfer function of the visual channel," *Journal of the Optical Society of America* **60**(6), 842–848 (1970).

37. A. Van Meeteren and J. M. Valeton, "Effects of pictorial noise interfering with visual detection," *Journal of the Optical Society of America A* **5**(3), 438–444 (1988).

38. 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).

39. L. Scott and J. D'Agostino, "NVEOD FLIR92 thermal imaging systems performance model," in *Infrared Imaging Systems: Design, Analysis, Modeling and Testing III,* G. Holst, Ed., *Proceedings of the SPIE* **1689,** 194–203 (1993).

40. G. C. Holst and J. W. Pickard, "Analysis of observer minimum resolvable temperature responses," in *Imaging Infrared: Scene Simulation, Modeling, and Real Time Image Tracking,* A. J. Huber, M. J. Triplett, and J. R. Wolverton, Eds., *Proceedings of the SPIE* **1110,** 252–257 (1989).

41. G. C. Holst, "Effects of phasing on MRT target visibility," in *Infrared Imaging Systems: Design, Analysis, Modeling and Testing II,* G. Holst, Ed., *Proceedings of the SPIE* **1488,** 90–98 (1991).

CHAPTER 5

Tracking and Control Systems

Robert E. Nasburg
Hughes Aircraft Company
El Segundo, California

CONTENTS

5.1 INTRODUCTION

This chapter examines the problem of automatic video target tracking and describes design solutions based on modern estimation and control theory. As tracking system technology has evolved, it has progressed from designs constrained by hardware limitations[1] to solutions based on mathematical models of the tracking environment and sensor data. Using the tools of detection and estimation theory,[2–4] improved tracking systems can now be developed from such models. This model-based design methodology[5] is the foundation on which current and future generations of video tracking systems will be constructed and evaluated.

Automatic video tracking systems (AVTSs) are employed in a wide variety of missions and tracking environments. They are used in numerous applications, including fire control, guidance, navigation, passive range estimation, and automatic target discrimination among others. As a consequence, they tend to be highly specialized, resulting in many different hardware implementations. Rather than attempt to describe a multitude of hardware configurations, this chapter focuses on the basic signal processing functions that are common to most tracking systems.

Throughout the 1970s and early 1980s, a major challenge confronting automatic video tracker designers was the need to implement the signal processing within the limited capabilities of the available hardware. Tracking algorithms were usually implemented directly in dedicated hardware and were characterized by innovative hardware designs. By the mid-1980s, the advent of video analog-to-digital (A/D) converters, powerful programmable digital signal processors, and very dense application-specific integrated circuits alleviated many of these hardware constraints and shifted the design focus toward software solutions.[6] As a result, present-day tracking systems are distinguished from their predecessors by clever algorithms and software realizations. As a further consequence of this technology migration, modern detection and estimation theory now play a central role in the design of AVTSs.

The symbols and nomenclature used in this chapter are listed in Table 5.1. Units have not been listed because most of the symbols represent abstract mathematical quantities rather than physical variables.

5.1.1 Overview of an Automatic Video Tracking System

The purpose of an AVTS is to maintain a stable sensor-to-target line of sight (LOS) automatically, in the presence of both relative target motion and base motion disturbances to the sensor platform. The target is initially located (acquired) either by a human operator viewing a display or by some form of automatic target recognition system. After the target is acquired, the tracking subsystem locks onto it and thereafter maintains the LOS autonomously.

A typical AVTS consists of three major components (Fig. 5.1): (1) an imaging sensor, (2) a gimbal, and (3) a processor. A feedback control loop, called the *track loop,* continuously adjusts the gimbal to keep the target in the center of the sensor's field of view (FOV) or video gate. The processor closes this loop by computing the pointing error commands used to drive the gimbal. These components operate in the following sequence: (1) the processor locates the target signal in the video stream generated by the sensor, (2) the processor

Table 5.1 Symbols and Nomenclature

Symbols	Nomenclature
Az,El	Azimuth and elevation
$\mathbf{B}(\cdot)$	Background object plane
bias	Average error
$\mathbf{C}(\cdot)$	Two-dimensional sampling function
contrast	Difference between average target and background intensity
COV matrix	Covariance matrix
$\mathbf{d}_f,\mathbf{d}_g,\mathbf{d}_t$	Known or measurable functions
$\mathbf{e}$	Error vector
$E(\cdot)$	Expected value operator
e_{LOS}	LOS determination error
e_{p}	Pointing error
E_k	Innovation
e_{ss}	Steady-state error
$\mathbf{e}_t,\mathbf{e}_b,\mathbf{e}_d,\mathbf{e}_j$	Track, bias, drift, and jitter error vectors
E_t	Average per-pixel target image energy
e_{TL}	Target location estimation error
e_u	Error along u axis
FI matrix	Fisher information matrix
$\mathbf{f}_x,\mathbf{f}_t,\mathbf{f}_b$	State equations
$\mathbf{g}$	Measurement equation
g_u	Gain constant
G_1,G_2,G_3,G_4	Adaptation, LOS determination, loop compensation, and gimbal transfer functions
$\boldsymbol{Hist}_i$	Interior histogram
$\boldsymbol{Hist}_t,\boldsymbol{Hist}_b$	Target and background histograms
$\mathbf{HJ}_k$	Cost function Hessian matrix
HOT	Higher order terms
$I(i,j)$	Pixel intensity at (i,j)
IID	Independent identically distributed random variables
Int_t,Int_b	Target and background intensity
J	Cost function
J_k	Cost function after k frames
k	Frame number
$\mathbf{K}_k$	Two-dimensional correlation surface
$l(\cdot)$	Log-likelihood function
$\mathbf{n}_c$	Composite noise variable
$\mathbf{n}_k$	Sensor noise plus model error
N_G	Total number of pixels in gate
$\mathbf{N}_k$	Random pixel noise on frame k
N_t(level),N_b(level)	Number of target and background pixels with gray level
N_T,N_B	Number of true target and background pixels

Table 5.1 *(continued)*

Symbols	Nomenclature
Num_e	Number of exceedances
$\mathbf{P}(\cdot)$	Object plane image map
P_L	Probability of loss of lock
$\Pr(e_c\|B)$	Probability of misclassifying a background pixel
$\Pr(e_c\|T)$	Probability of misclassifying a target pixel
PSF$(\cdot)$	Point spread function
$\mathbf{Q}(\cdot)$	Target indicator function
$\mathbf{R}$	Rotation matrix
R_g	Gate
Z_k	Information set
$\mathbf{R}_k$	Target reference image
R_t, R_b	Target and background regions
s_1, s_2	Scaling parameters
$\mathbf{S}$	Scale matrix
$\mathbf{S}_k$	Sensor image
Std Dev	Standard deviation
SUM_u, SUM_v	Weighted index summation
(t_u, t_v)	Target location
(t_u^*, t_v^*)	Target location optimizing cost function
$(\hat{t}_u, \hat{t}_v)$	Target location estimate
$(\hat{t}_u^P, \hat{t}_v^P$	Predicted target location estimate
$\mathbf{T}$	Translation vector
$\mathbf{T}(\cdot)$	Target object plane map
T_f	Frame time
$\mathbf{T}_k$	True target image
$Total_t, Total_b$	Number of pixels classified as target and background
var	Error variance
VideoMap$(\cdot)$	Pixel intensity to binary map
$x_{\text{lgth}}, y_{\text{lgth}}$	Lengths
$x_{\text{size}}, y_{\text{size}}$	Target dimension along axes
x_{uT}	Target centroid, u axis location
α, β, γ	Parameters
Θ_c	Composite parameter vector
level	Pixel gray level
$\mathbf{x}$	State vector
$*$	Two-dimensional convolution operator
(i,j)	Pixel coordinates
θ	Angle of rotation
$\mathbf{w}_f, \mathbf{w}_g, \mathbf{w}_t$	State noises
Θ_t, Θ_b	Target and background parameters
r_{uT}^2, r_{uB}^2	Radius of gyration of target and background in u direction

(continued)

Table 5.1 *(continued)*

Symbols	Nomenclature
Θ_r, Θ_s	Rapidly and slowly varying parameters
$\partial J/\partial t_u$	Cost sensitivity along u axis
$\partial J/\partial s_1$	Cost sensitivity to change in u axis scale
Θ_s^p	Predicted slowly varying parameters
$\hat{\Theta}_c$	Estimated composite parameter vector
$\nabla \mathbf{J}_k$	Cost function gradient vector
σ^2	Variance
Δ	Determinant
w_l, w_f, w_i	Weighting functions
σ_x, σ_y	Standard deviations
Θ^*	Parameter vector optimizing cost function

estimates the target "state" (described later) and generates the gimbal commands based on state information, (3) the gimbal command is applied to the sensor's LOS, (4) the sensor generates a new video stream, and (5) the process repeats.

Although the three identified components are common to most AVTSs, each can assume any of several forms for a particular application, depending on the system requirements. For example, early analog centroid trackers employed analog comparators and digital counters to measure or estimate the location of the target's centroid in the video stream. In these systems, the target's LOS was based solely on its centroid, and the control law was implemented by means of a simple proportional control loop. In contrast, current tracking systems designed to track ground vehicles moving in heavily cluttered environments might use a multimode approach based on correlation and centroid algorithms. To cite another example, in order to accommodate the high dynamic rates typically encountered on airborne platforms, aircraft navigation

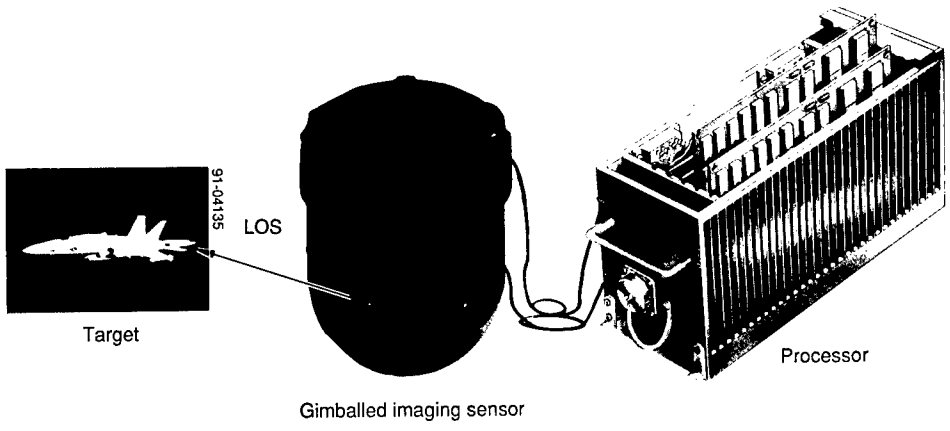


Fig. 5.1 Imaging sensor, gimbal, and processor.

information is used[7,8] as ancillary data in determining the LOS. This type of system requires multiple computers, a sophisticated video processor, and a high-speed interface to the aircraft's navigation system.

No "universal" tracking system design exists that can meet the performance requirements specified for all tracking environments of interest. Tracker designs tend to be environment sensitive. For example, a star tracker would be virtually useless for tracking a maneuvering, noncooperative target. Similarly, an up-looking tracker designed to track airborne or spaceborne targets against a sky or space background would be defeated by ground clutter if it were used in a down-looking surveillance system. It is therefore essential to incorporate into the tracker design an appropriate description of the expected target and background characteristics. Models are used to quantify the target and background characteristics. Sections 5.2 and 5.3 describe models used to characterize sensor data typical of various target and environmental parameters.

5.1.2 Track Loop Architecture

The feedback loop implementing the tracking system is inherently nonlinear and must operate with uncertain or noisy measurements. To simplify the design process for this type of feedback loop, the separation principle[9,10] from stochastic control theory is invoked. This principle defines a control architecture that separates state estimation, based on the sensor data, from control law generation (Fig. 5.2). The advantage of this separation is that it allows sophisticated and existing design procedures to be applied to both the estimation[11,12] and control law[9,13] aspects of the design. The control law can then be based on the principle of certainty equivalence,[4,10] which replaces random quantities by their expected or average values during the control law design. This simplification converts an extremely difficult control problem into a more manageable deterministic control law design problem. Although, strictly speaking, this is suboptimal, experience shows that it yields good tracking results.

To further reduce design complexity, state estimation is divided into two parts: (1) target location estimation and (2) LOS determination (Fig. 5.2). Target location estimation analyzes the sensor's imagery to determine the position

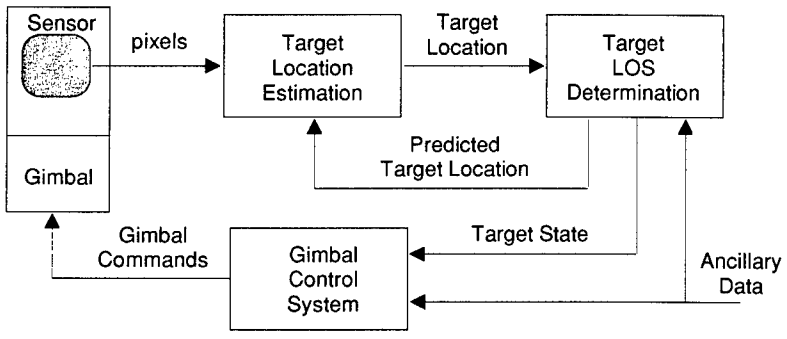


Fig. 5.2 Separation principle applied to track loop architecture.

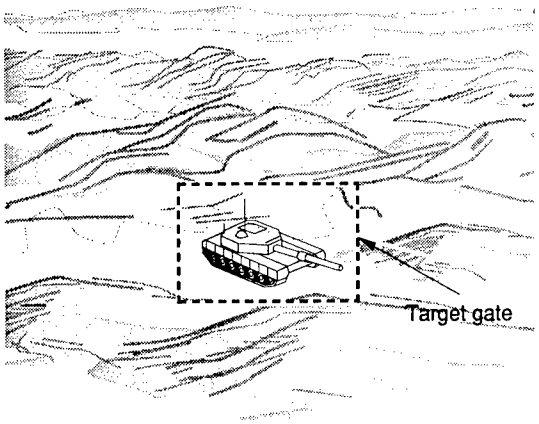


Fig. 5.3 Target gating suppresses clutter.

of the target in sensor coordinates. This is an image processing task that often requires extremely high processor throughput rates. The target location estimation function is discussed in Sec. 5.2.

The target LOS determination component[14] utilizes target location estimates and possibly inertial navigation system (INS) data to determine the target's state for use by the track loop controller (Sec. 5.4). Methods for implementing target LOS determination can vary over wide limits, depending on the operating requirements. For example, an earthbound sensor designed to track slowly moving vehicles might simply use the difference between the target's estimated location and the center of the sensor's FOV as the LOS error criterion for the gimbal controller. In contrast, a tracker that is required to follow maneuvering airborne targets from a high-performance aircraft platform might utilize the aircraft's INS to compensate for sensor platform motion, and a Kalman filter to make high-quality target location estimates. (Nonimaging sensor data, such as that from the INS, are referred to as *ancillary data.*)

The target location estimate is often computed on a small subimage called a *target gate* (or simply, *gate*). Figure 5.3 depicts a typical target gate surrounding a target. By processing only the portion of the scene lying inside the target gate, the effects of clutter and noise on the tracking process are largely suppressed, resulting in a more robust tracking system. Utilizing just the imagery in the gate for target location estimation also helps to limit the processing throughput requirement. As a rule, the gate (which is usually rectangular or square in shape) should have the smallest dimensions necessary to completely contain the target.[a] Analyses and techniques for target gating are discussed in Sec. 5.4.

[a]In sizing the gate, some allowance must be made for jitter and other angular disturbances. Under smooth tracking conditions, however, the necessary margin is normally a small fraction of the gate dimensions.

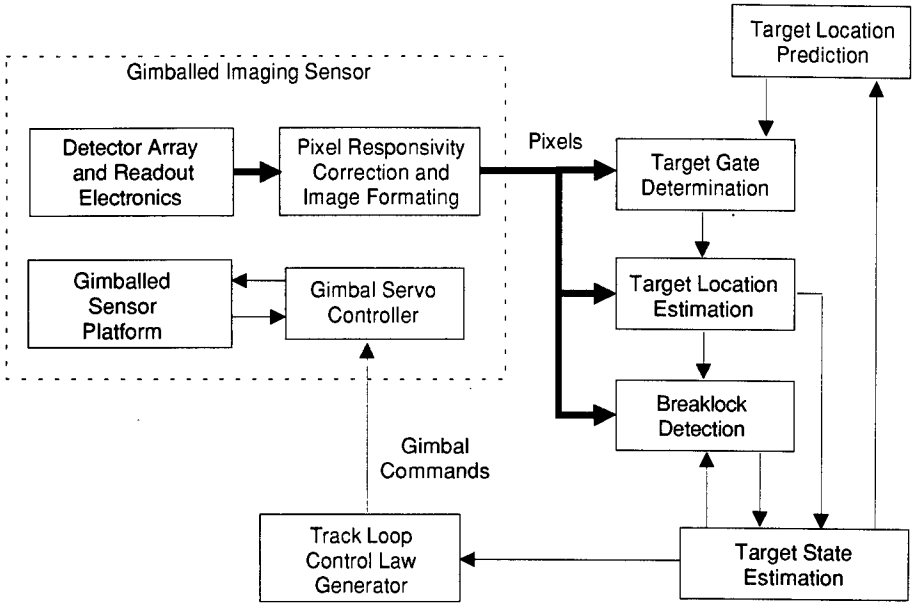


Fig. 5.4 General tracking system architecture.

The primary tracking system functions are shown in Fig. 5.2. Additional secondary functions are needed to support the primary functions. The two most important secondary functions are target gate determination (which supports target gating by measuring target size) and the breaklock detection function used by the LOS determination component. Figure 5.4 shows the general tracking system architecture, including both primary and secondary functions.

It is not uncommon to obtain a bad target location estimate from a failure of the target location estimator. This could occur, for example, if the target becomes obscured by ground clutter, leading to an unreliable location estimate. Obviously, the erroneous target location data generated under such conditions should not be used in determining the LOS. The breaklock detection function in Fig. 5.4 is used to sense when the target location estimate is unreliable. This ensures that erroneous or uncertain data will be ignored by the target LOS determination function, thereby improving track loop performance under stressful conditions.

When a breaklock condition occurs, the tracking system can continue to point the gimbal based on the predicted target state. In this "coast" mode, the track loop is open, i.e., the pointing axis continues to move in the direction and at the rate determined by the last target state estimate before breaklock. After the target has been reacquired and a reliable target location estimate has again been established, the breaklock function signals the LOS determination function to resume closed-loop operation.

Figure 5.4 illustrates a commonly used gimbal control technique that partitions the control function into LOS stabilization and track loop compensation. The LOS stabilization can be viewed as a high-speed inner loop working in

conjunction with a slower outer loop that incorporates a track loop compensator. The inner loop provides relatively high bandwidth closed-loop stabilization of the gimbal, based on gimbal position and possibly accelerometer inputs. Line-of-sight stabilization removes (or greatly attenuates) high-frequency sensor platform disturbances, and is often implemented as an analog signal processor. The track loop compensator implements the outer loop compensation using pointing information derived from sensor imagery. The compensator operates at the sensor frame rate, which is much slower than the sampling rate of the stabilization loop. Thus, track loop compensation is typically implemented in the digital processor,[15] along with LOS determination. Section 5.4 discusses both of these functions.

5.2 TARGET LOCATION ESTIMATION

Many different methods of target location estimation have been developed. The most common (and probably the best known) of these is the centroid tracker. Although this algorithm works well in some tracking environments, it performs poorly in others. The limitations of centroid tracking stem fundamentally from its inability to fully exploit the target signature information present in the image. To overcome this deficiency, maximum likelihood estimation theory can be invoked to identify better performing target location estimators. It will be shown that correlation and multimode trackers belong to the class of target location estimators. These trackers are often required to operate in highly dynamic clutter environments.

This section is concerned primarily with the mathematical foundations on which target location estimator analysis and design can be based. Many of the details needed to exploit this foundation in the design of practical tracking systems are beyond the scope of this chapter.

5.2.1 Centroid Trackers

5.2.1.1 Binary Centroid Tracker. Binary centroid trackers[16] employ the image processing architecture shown in Fig. 5.5.[b] This architecture does not require high-speed A/D converters or high-performance digital signal processors and can therefore be implemented with relatively simple hardware. For this reason, the binary centroid tracker is an attractive option for applications that require a low-cost tracker with only modest performance capabilities. Most early imaging trackers used this architecture because high-performance digital hardware was not available.

Consider the segmented image shown in Fig. 5.6, where each pixel has been assigned a value of 1 if it corresponds to the target, and a value of 0 otherwise. The binary centroid of a segmented image is analogous to the center of mass of a discrete distribution of unit masses. If the video map (Fig. 5.5) performs the segmentation by assigning a value of 1 to each target pixel and a value

[b]Binary centroid trackers are also called *silhouette* or *1-bit trackers*.

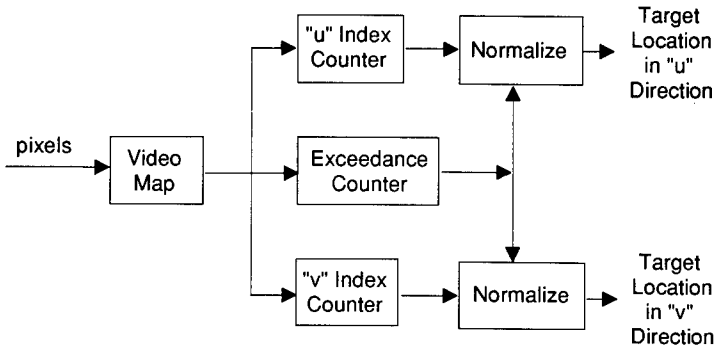


Fig. 5.5 Binary centroid tracking architecture.

of 0 to each background pixel, the target's centroid coordinates (location estimates) $(\hat{t}_u,\hat{t}_v)$ within the target gate R_g can be computed directly as follows:

$$\hat{t}_u = \frac{SUM_u}{Num_e} \ , \quad \hat{t}_v = \frac{SUM_v}{Num_e} \ , \tag{5.1}$$

where

$$SUM_u = \sum_{(i,j)\in R_g} u_i I(i,j)$$

$$SUM_v = \sum_{(i,j)\in R_g} v_j I(i,j)$$

$$I(i,j) = \begin{cases} 1 & \text{if the video map assigned the pixel at} \\ & (u_i,v_j) \text{ to the target} \\ 0 & \text{otherwise} \end{cases}$$

$$Num_e = \sum_{(i,j)\in R_g} I(i,j) = \text{total number of threshold exceedances (target pixels) in the gate}$$

R_g = region lying within the track gate.

For the example in Fig. 5.6, the total number of exceedances (Num_e) is 91. The weighted index sums SUM_u and SUM_v are 780 and 596, respectively. The binary centroid is formed by normalizing the weighted index sums by the total number of exceedances, yielding $\hat{t}_u = 780/91 = 8.57$ and $\hat{t}_v = 596/91 = 6.55$. The resulting binary centroid position is indicated in Fig. 5.6.

Because $I(.,.)$ is a binary variable (1 or 0), the summations SUM_u, SUM_v, and Num_e can be performed with binary adders. Two division operations then complete the calculation of $\hat{t}_u$ and $\hat{t}_v$.

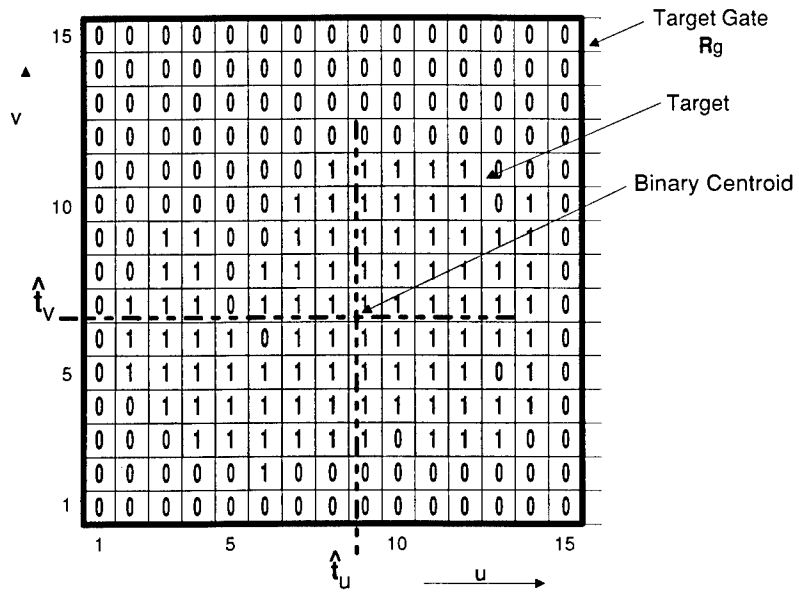


Fig. 5.6 Segmented target image.

To compute the target's centroid, the video map must segment the target from the background.[c] The performance of the binary centroid tracker depends critically on the ability of the video map to perform this task. A wide variety of video maps exist, ranging from the simplest, implemented by a single threshold, to a video map implemented as a lookup table. For the simple video map, the intensity is set equal to 1 if it exceeds the threshold (an exceedance) and to 0 otherwise. This implementation works well for targets that exhibit high contrast with respect to the background. On the other hand, a single threshold is inadequate for bimodal targets, which have areas of both positive and negative contrast. Such targets require a more complex video map.

The most general type of video map is implemented as a lookup table. A number of algorithmic approaches are available to specify the elements of the table adaptively. One representative technique based on intensity histogram data is outlined as follows[d]:

[c]The verb *to segment* means to extract the target from the background on a pixel-by-pixel basis. This function is performed by the processor (see Sec. 5.3.3).

[d]An intensity histogram displays the number of observed occurrences at a series of increasing intensity levels. Histograms are often used to estimate a probability density function.

$$\textbf{\textit{VideoMap}}(\text{level})_{k+1} = \begin{cases} 1 & \text{if } \textbf{\textit{Hist}}_t(\text{level})_{k+1} > \gamma \textbf{\textit{Hist}}_b(\text{level})_{k+1} \\ 0 & \text{otherwise} \end{cases}$$

$$\textbf{\textit{Hist}}_t(\text{level})_{k+1} = (1 - \alpha)\textbf{\textit{Hist}}_t(\text{level})_k + \alpha\left[\frac{N_t(\text{level})_k}{Total_t}\right] \tag{5.2}$$

$$\textbf{\textit{Hist}}_b(\text{level})_{k+1} = (1 - \beta)\textbf{\textit{Hist}}_b(\text{level})_k + \beta\left[\frac{N_b(\text{level})_k}{Total_b}\right],$$

where

$\textbf{\textit{VideoMap}}(\cdot)_k$	=	a function whose domain is the intensity gray levels and whose range is 0 or 1 (binary); it is applied to the image at frame number k
$\textbf{\textit{Hist}}_t(\text{level})_k$	=	normalized target histogram at frame number k; the histogram is normalized so that $\sum_{\text{levels}} \textbf{\textit{Hist}}_t(\text{level})_k = 1$
$\textbf{\textit{Hist}}_b(\text{level})_k$	=	normalized background intensity for the k'th frame
$N_t(\text{level})_k$	=	number of occurrences belonging to the target that have the specified gray level in the k'th frame
$N_b(\text{level})_k$	=	number of occurrences belonging to the background that have the specified gray level in the k'th frame
$Total_t$	=	total number of exceedances ($\textbf{\textit{VideoMap}} = 1$)
$Total_b$	=	total number of nonexceedances ($\textbf{\textit{VideoMap}} = 0$)
α and β	=	parameters that govern the adaptation rates of the target and background histograms, respectively
γ	=	another parameter; its value is usually determined experimentally to obtain the best performance characteristics
level	=	gray level or pixel intensity value.

In the video map algorithm of Eq. (5.2), the results of the target/background classification were used to develop the target and background histograms. But these histograms were then used to define the target/background classification, via the video map; hence, this algorithm is circular, i.e., the decision depends on itself. For example, if pixels were classified incorrectly because of noise,[e] the target and background histograms contain errors. As a result, the histograms are skewed to favor the classification errors, giving rise to further errors. In this way, errors continue to accumulate until the video map no longer yields reliable classifications, resulting in poor tracker performance. Clearly, the buildup of error in a circular algorithm must be tightly controlled if the algorithm [Eq. (5.2)] is to be useful in a practical tracking system. This is a major problem in algorithm design.

[e]Noise can give rise to two types of pixel classification errors: (1) *false alarms* (exceedance due to noise alone) and (2) *signal dropouts* (target signal driven below threshold by a negative noise spike).

The video map performs target/background classification by considering only one pixel at a time. Tracking can be improved by including information from surrounding pixels to reduce misclassifications, as is done by the target segmentation algorithms discussed in Sec. 5.3. A segmentation algorithm used in conjunction with the binary centroid algorithm can improve performance, but this expedient tends to nullify the principal advantage of the latter algorithm—its relatively simple implementation.

The tracking performance of the binary centroid algorithm can be quantified by its target location estimation error $\mathbf{e}(k)$, which is the difference between the estimated location of the target in the object plane and the target's true location. The estimation error is caused by sensor noise, and therefore is treated as a random variable. Ideally, a random variable is described by its probability density function, but because the true density function is seldom known, the location error is characterized by its mean value (or systematic bias) and its standard deviation.[17] If, as is often the case, the probability density function can reasonably be assumed to be Gaussian, then these parameters completely determine the properties of $\mathbf{e}(k)$.

In general, it is not possible to completely characterize the binary centroid error analytically because of the complex nonlinear function involving thresholds, products, and divisions. It is possible, however, to generate a first-order performance approximation, as in Eq. (5.3), by assuming that the target and background can be characterized by uniform intensity[f] values Int_t and Int_b, respectively ($Int_t \neq Int_b$), and that the pixel noise $\mathbf{N}(i,j)$ is spatially independent and identically distributed (IID).[g] Because the approximation can be very crude, this calculation should be carried out in conjunction with a detailed simulation analysis.

If $\mathbf{S}_k(i,j)$ denotes the pixel value at the coordinate position (i,j) on the k'th frame, then

$$\mathbf{S}_k(i,j) = \begin{cases} Int_t + \mathbf{N}_k(i,j) & \text{if } (i,j) \text{ belongs to the target} \\ Int_b + \mathbf{N}_k(i,j) & \text{otherwise} \end{cases} , \tag{5.3}$$

where

$\mathbf{S}_k(i,j)$ = pixel value (gray level) at location (i,j) for frame number k
Int_t = target intensity (gray level)
Int_b = background intensity (gray level)
$\mathbf{N}_k(i,j)$ = random pixel noise.

When $\mathbf{S}_k(i,j)$ is modeled in this way, it can be shown that the expected value (or bias) and the variance of the pointing error $\mathbf{e}(k)$ can be approximated as

$$\text{bias} = E\{\mathbf{e}_u(k)\} \approx -\frac{N_G}{N} x_{uT} \Pr(e_c|B) , \tag{5.4}$$

[f]In radiometric terminology, the concept of *radiance* (or *stearance*) is used to characterize the radiation signature of a resolved target, as well as the background. (In this context, the term *intensity* is reserved for unresolved targets.) When dealing with a spatially quantized image, however, it is customary to describe the target and background in terms of their equivalent pixel intensities, as is done throughout this chapter.

[g]This is more restrictive than spatially uncorrelated pixel noise.

$$\text{var}\{\mathbf{e}_u(k)\} \approx \frac{1}{N_T}\{r_{uT}^2 \Pr(e_c|T)[1 - \Pr(e_c|T)] + \frac{N_B}{N_T} r_{uB}^2 \Pr(e_c|B)[1 - \Pr(e_c|B)]\} , \tag{5.5}$$

where

$\mathbf{e}_u(k)$ = location error in the u direction based on the data in frame k; this error is defined to be the difference between the centroid algorithm's estimate of the target location and the target's true centroid location

x_{uT} = target's true centroid location

$\Pr(e_c|B)$ = probability that a background pixel is incorrectly classified as a target by the video map

$\Pr(e_c|T)$ = probability that a target pixel is incorrectly classified as background by the video map

N_G = total number of pixels in the gate

N_T = number of true target pixels in the gate

N_B = the number of true background pixels in the gate

r_{uT} = radius of gyration in the u direction for the target within the gate

$$r_{uT}^2 = \left(\frac{1}{N_T}\right) \sum_{\text{target}} (u - x_{uT})^2$$

r_{uB} = radius of gyration in the u direction for the background within the gate

$$r_{uB}^2 = \left(\frac{1}{N_B}\right) \sum_{\text{background}} (u - x_{uT})^2 .$$

From the foregoing definitions, it follows that N_T/N_G is the percentage of the target area within the gate and N_B/N_G is the percentage of background within the gate.

Equations (5.4) and (5.5) approximate the location error in the u direction. Similar expressions hold for the v direction.

Assuming that the sensor noise $\mathbf{N}_k(i,j)$ consists of a time sequence of white noise samples and that the video map does not change, $\{\mathbf{e}_u(k)\}$ is a white noise IID sequence. If video map adaptation is permitted, however, then the sequence $\{\mathbf{e}_u(k)\}$ is not IID as a result of correlation through the video map. However, the degree of correlation is often weak, so that $\{\mathbf{e}_u(k)\}$ can be approximated as a white noise sequence when analyzing track loop performance (Sec. 5.5).

In the foregoing analysis, the most restrictive assumptions are the uniform intensity of the target and the background. This can be generalized by regionalizing. Let R_g, R_t, and R_b represent the regions defined by the target gate, the target within the gate, and the background within the gate, respectively. Partition R_t and R_b into disjoint (nonoverlapping) subregions, i.e.,

$$R_t = R_{t_1} \cup R_{t_2} \ldots \cup R_{t_n} ,$$

$$R_b = R_{b_1} \cup R_{b_2} \ldots \cup R_{b_m} ,$$

where the pixel intensities associated with each subregion can be modeled as a constant intensity plus random IID noise. The hypothesis of Eq. (5.3) is satisfied in each subregion; hence, location estimation errors associated with each subregion are approximated by Eqs. (5.4) and (5.5). If it is further assumed that the random IID noise samples associated with each subregion are mutually independent, then the subregion error characterizations can be combined to determine the bias and variance of the target's location error:

$$\text{bias} = E\{\mathbf{e}_u(k)\} \approx \sum_{i=1}^{m} x_{uB_i} \frac{N_{B_i}}{N_T} \Pr(e_c|B_i) - \sum_{i=1}^{n} x_{uT_i} \frac{N_{T_i}}{N_T} \Pr(e_c|T_i) \ , \tag{5.6}$$

and

$$\text{var}\{\mathbf{e}_u(k)\} \approx \frac{1}{N_T} \sum_{i=1}^{n} \Pr(e_c|T_i)[1 - \Pr(e_c|T_i)] \frac{N_{T_i}}{N_T} (r_{uT_i}^2 + x_{uT_i}^2) + \frac{1}{N_T} \sum_{i=1}^{m} \Pr(e_c|B_i)[1 - \Pr(e_c|B_i)] \frac{N_{B_i}}{N_T} (r_{uB_i}^2 + X_{uB_i}^2) \ , \tag{5.7}$$

where the symbols have the same meanings as in Eqs. (5.4) and (5.5).

Although Eqs. (5.6) and (5.7) are an approximate description of the target location error, they are useful for studying the effect on centroid performance of varying target signatures and competing clutter. Here, the term *competing clutter* refers to areas of the background with intensities similar to that of the target, making it difficult for the video map to distinguish these areas from the true target. Equations (5.4) and (5.5) also provide insight into the effect of variations or uncertainty in a target's intensity map. For example, a military vehicle exhibits a different thermal signature when its engine compartment is hot than when it is cold.

5.2.1.2 Intensity Centroid. The architecture of the intensity centroid is similar to that for the binary centroid, but it does not employ the video map to classify target and background pixels.[18] Instead, it uses intensity information directly in the centroid calculation. The intensity centroid is defined by the architecture of Fig. 5.5, with the video map removed:

$$\hat{t}_u = \frac{SUM_u}{I_t} \ , \quad \hat{t}_v = \frac{SUM_v}{I_t} \ , \tag{5.8}$$

where

$(\hat{t}_u, \hat{t}_v)$ = intensity centroid estimate of target location

$$SUM_u = \sum_{(i,j) \in R_g} u_i I(i,j)$$

$$SUM_v = \sum_{(i,j) \in R_g} v_j I(i,j)$$

$I(i,j)$ = pixel intensity at location (u_i,v_j)

$$I_t = \sum_{(i,j)\in R_g} I(x_i,y_j) = \text{total intensity over the target gate.}$$

This algorithm can be implemented efficiently in digital signal processor architectures that have access to array index values within their pipeline.

One important application of the intensity centroid algorithm is high-accuracy point-target tracking. For this application, the sensor's optical point spread function (PSF) must span two or more pixels on the focal plane. If the PSF is approximately Gaussian in shape, the target's intensity distribution can be characterized by a two-dimensional Gaussian function. For this distribution, the target's binary and intensity centroids coincide. If the binary centroid were used, its video map would not realize any benefit from the high-intensity values in the region of the Gaussian peak because this algorithm assigns equal weight to all target pixels. The intensity centroid, on the other hand, makes better use of the target's signature by weighting these high-intensity pixels more than the lower intensity pixels. This results in better pointing accuracy for the intensity centroid.

The accuracy of the location estimate's intensity centroid is extremely sensitive to the size of the target gate. In the previous example, if the gate is large, then pixel values containing mostly noise are included in the centroid calculation, which degrades the accuracy. The binary centroid was found to be less sensitive to noise because the video map excluded most of these noisy pixels. This implies that the accuracy of the intensity centroid can be improved if a video map is used to eliminate pixel values not clearly associated with the target. The video map acts like a pixel-level gating function in the centroid calculation of Eq. (5.8). This type of intensity centroid is sometimes called a *thresholded intensity centroid.* It is computed from Eq. (5.8) using the following:

$$SUM_u = \sum_{(i,j)\in R_g} u_i[I(i,j)]\textbf{\textit{VideoMap}}[I(i,j)] \ ,$$

$$SUM_v = \sum_{(i,j)\in R_g} v_j[I(i,j)]\textbf{\textit{VideoMap}}[I(i,j)] \ ,$$

$$I_t = \sum_{(i,j)\in R_g} I(i,j)\textbf{\textit{VideoMap}}[(I(i,j)] \ .$$

Approximate analytic expressions for the location error of the intensity centroid and thresholded intensity centroid can be derived in a manner similar to that previously used for the binary centroid.[18] Taylor expansion and regionalization are effective analytic techniques for the intensity centroids.

To illustrate the relative target location estimation accuracy of these centroid algorithms, the accuracy, as measured by the standard deviation of the error value, is plotted as a function of the square root of the average per-pixel energy in the target image divided by the standard deviation of the sensor noise. This is one form of SNR for target imagery. The average per-pixel target image energy is given by

$$E_t = \sum_{(i,j)\in \text{target}} I^2(i,j)/(\text{pixels on target}) \ .$$

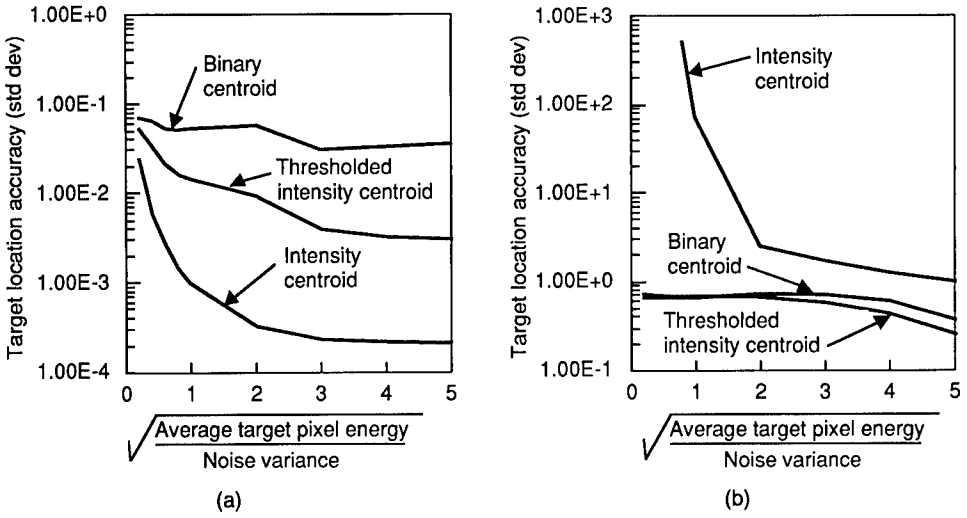


Fig. 5.7 Binary and intensity centroid algorithm performance: (a) Gaussian intensity distribution and (b) uniform intensity distribution.

For targets with uniform intensity distributions, this SNR definition becomes (target contrast)/(sensor noise standard deviation), which has commonly been used in specifying tracking systems.

Figure 5.7 illustrates the accuracy of the different centroid algorithms for target intensity with (a) a Gaussian distribution and (b) a uniform distribution. The Monte Carlo runs generating this figure included the effects of spatial quantization by the imaging system and blurring in the optics. The video map was implemented by a simple single threshold. No attempt was made to optimize the threshold setting. Furthermore, in this example, the target dimensions were a small fraction of the gate size, so the accuracy was influenced to a significant degree by the background. Gaussian random sensor noise was used in all Monte Carlo runs.

Comparing Figs. 5.7(a) and 5.7(b) illustrates that accuracy improves as SNR increases. However, the curves are strongly dependent on the target's intensity distribution. For the Gaussian case, Fig. 5.7(a), the binary centroid performed worse than the intensity centroids because it does not take advantage of the large intensity values near the peak of the intensity distribution. The thresholded intensity centroid splits the difference between the binary and intensity centroids because thresholding prevents it from utilizing target intensity values in the tails of the Gaussian intensity distribution.

For the uniform target intensity, the interrelationships among the centroid accuracy curves change completely. The binary and thresholded intensity centroids outperform the intensity threshold because they suppress background noise through thresholding. The thresholded centroid performs slightly better at high SNR because it can take advantage of the target's intensity signature modified by the optical blurring of the imaging system. The thresholded intensity centroid algorithm exploits the slight intensity slopes around the target boundary more effectively than the binary centroid.

Figure 5.7 also illustrates that the accuracy of centroid target location estimates depends on how effectively an estimation algorithm utilizes the target's intensity distribution (target's signature). Sections 5.2.2 and 5.2.3 address target location estimation algorithms that are explicitly optimized for a particular target signature. These algorithms are capable of higher accuracy than the centroid algorithms.

5.2.2 Maximum Likelihood Estimate of Target Location

Although the centroid algorithms perform quite well for simple targets in uncluttered backgrounds, a more powerful target location estimator is needed for more stressful environments. The maximum likelihood target location estimator is one such example.

The maximum likelihood principle (from mathematical statistics) provides a method of estimating target location from measured data.[3,5,10,12] The maximum likelihood estimate (MLE) of a target's parameters is obtained by maximizing the likelihood function, which is the conditional probability of the measured data for a given set of target parameter values.[19,20] The likelihood function is derived from the overlay model, which relates the target parameters to the measured data. Because the tracking algorithm is synthesized from the overlay model, maximum likelihood tracking belongs to the class of model-based algorithms.

Maximum likelihood tracking is optimal in the sense that it provides the most accurate unbiased estimates of the target's location. Even though the MLE algorithm frequently cannot be directly implemented in available hardware, it serves as a useful reference against which the performance of other tracking algorithms can be gauged. Thus, in a simulation study, any tracking algorithm whose performance approaches that of a maximum likelihood algorithm should be viewed as a strong candidate for implementation.

As a rule, accurate mathematical models, such as the overlay model, yield high-performance target tracking algorithms. But the price of this superior performance is often an algorithm that is complex and difficult to implement. To improve the implementability of the algorithm, the tracker design often entails simplifying the overlay model. The correlation tracker (Sec. 5.2.3) is an important example of this design philosophy; its implementation is less computationally intensive than an algorithm based rigorously on the overlay model.

Another way of improving the implementability of a tracking algorithm is to select an efficient mathematical optimization algorithm.[21] The optimization literature describes a variety of efficient algorithms applicable to real-time tracking.

5.2.2.1 Overlay Model. The overlay model [Eqs. (5.9) to (5.14)] is a composite of five interlinked submodels: (1) a submodel of the open-loop LOS dynamics, (2) a target parameter submodel, (3) a background parameter submodel, (4) an object plane submodel, and (5) a sensor submodel. Figure 5.8 shows the architecture of the overlay model and the linkages between the submodels.

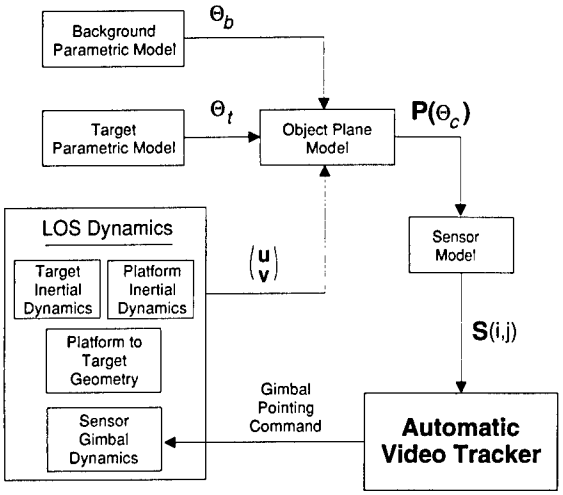


Fig. 5.8 Architecture of the overlay model.

Mathematically, the overlay model has the following form:

$$\mathbf{x}(k+1) = \mathbf{f}_x[\mathbf{x}(k),\mathbf{d}_f(k),k] + \mathbf{w}_f(k) \ , \tag{5.9}$$

$$[u(k),v(k)] = \mathbf{g}[\mathbf{x}(k),\mathbf{d}_g(k),k] + \mathbf{w}_g(k) \ , \tag{5.10}$$

$$\Theta_t(k) = \mathbf{f}_t[\Theta_t(k),\mathbf{d}_t(k),k] + \mathbf{w}_t(k) \ , \tag{5.11}$$

$$\Theta_b(k) = \mathbf{f}_b[\Theta_b(k),\mathbf{d}_b(k),k] + \mathbf{w}_b(k) \ , \tag{5.12}$$

$$\mathbf{P}(\Theta_c) = \mathbf{Q}[(u,v),\Theta_t]\mathbf{T}(\Theta_t) + \{1 - \mathbf{Q}[(u,v),\Theta_t]\}\mathbf{B}(\Theta_b) \ , \tag{5.13}$$

$$\mathbf{S}_k(i,j) = \mathbf{C}(u,v)\mathrm{PSF}^*\mathbf{P}(\Theta_c) + \mathbf{N}_k(i,j) \ , \tag{5.14}$$

where

$\mathbf{x}(k)$	= state variable characterizing the open-loop pointing system dynamics associated with motion of the target, sensor platform, and gimbal subsystems; this variable and the state equation modeling it describe the target LOS
$\mathbf{d}_t,\mathbf{d}_g,\mathbf{d}_b$	= known functions of time; these functions, which may be ancillary (nonimage) measurements, are inputs to the dynamic equations
$\mathbf{g}$	= transformation from state $\mathbf{x}$ to the LOS in the sensor's object plane coordinates
(u,v)	= coordinates of the sensor's object plane

$\mathbf{w}_f(k), \mathbf{w}_g(k), \mathbf{w}_t(k), \mathbf{w}_b(k)$	= the state error ("plant noise") associated with the dynamic equation for target LOS, the measurement error associated with the transformation to the sensor's object plane, the state error associated with the dynamic model for the target's parameters, and the state error associated with the dynamic model for the background parameters, respectively; it is assumed that these are mutually independent white noise sources
$\mathbf{T}_k(\Theta_t)$	= target's object plane intensity distribution parameterized by Θ_t
$\mathbf{B}_k(\Theta_b)$	= background object plane intensity distribution parameterized by Θ_b
$\mathbf{Q}[(u,v),\Theta_t]$	= target indicator function, which is unity for object plane locations corresponding to a target, and zero otherwise
$\mathbf{P}(\Theta_c)$	= sensor's object plane image formed by overlaying the target on the background
$*$	= two-dimensional convolution operator
$\mathrm{PSF}(u,v)$	= sensor's optical point spread function including the detector area and a high-frequency pointing jitter function
$\mathbf{C}(u,v)$	= two-dimensional comb or sampling function; this function converts the continuous representation of the sensor data to sampled pixel values
$\mathbf{N}_k(i,j)$	= sensor noise associated with the photon detection process and modeling errors; it is typically assumed to have Gaussian statistics and to be both spatially and temporally uncorrelated
$\mathbf{S}_k(i,j)$	= intensity value of the (i,j)'th pixel
$\Theta_t,\Theta_b,\Theta_c$	= target, background, and combined parameter vectors for the overlay model with $\Theta_c = \Theta_t \cup \Theta_b$ and target location $(t_u,t_v) \in \Theta_t$
$k(k = 1, 2, ..., n)$	= image frame number corresponding to the sampling times $t_1,t_2,...,t_n$.

The first dynamic submodel of the overlay model characterizes the open-loop LOS behavior of the pointing system. Equation (5.9) (the state equation) models the dynamics of the target, gimbal set, and sensor platform, whereas Eq. (5.10) transforms the state vector $\mathbf{x}$ into the sensor's object plane coordinates. In the language of control theory, this submodel is the *plant* around which the tracking system closes the loop.

The second submodel characterizes the target. The target's signature can be parameterized by factors such as size, shape, and contrast. These variables

depend on (1) range changes between the target and the sensor platform, (2) rotation of the sensor platform with respect to the target, (3) rotation of the target with respect to the background, (4) target signature changes resulting from changes in target orientation, and (5) heat loading of the target and background over time. A common feature of most of these changes is the fact that the signature variations occur slowly in comparison to the sensor's frame rate. The resulting temporal signature correlation (slow changes) can be modeled parametrically by difference equations [Eq.(5.11)]. For this reason, the overlay model employs difference equation submodels to characterize the target's signature variations. The same rationale is applied to the third submodel, which characterizes the background [Eq. (5.12)]. The target parameter and background parameter submodels are uncoupled because they arise from physically independent sources.

These dynamic submodels represent an important feature of the overlay model. Earlier tracking system models tended to be static, in the sense that temporal correlations between image frames were not modeled explicitly (cf., the correlation tracker model in Sec. 5.2.3). Explicitly modeling the frame-to-frame dynamics using the open-loop LOS dynamic submodel and the target and background parameter submodels can result in improved tracking system performance. This is particularly true for stressing scenarios that include maneuvering targets.

The sensor's object plane is characterized by the fourth submodel [Eq. (5.13)], which overlays the target onto the background. (This is the origin of the name *overlay model.*) Although its nonlinearities and complexities are analytically troublesome, the overlay does permit a very rich characterization of the target and the background. For example, the background scene may contain higher spatial frequencies than the target. By means of the overlay structure, this difference can be modeled and applied to the tracking analysis.

This submodel contains both target and background object plane representations: $\mathbf{T}(\Theta_t)$ maps the target parameters into the two-dimensional object plane and $\mathbf{B}(\Theta_b)$ performs a similar role for the background. The target indicator function $\mathbf{Q}[(u,v),\Theta_t]$ governs how the target and the background are combined into the scene viewed by the sensor; $\mathbf{Q}[(u,v),\Theta_t]$ can be partitioned into the product of two parts. One part depends only on Θ_t. It represents the target boundary and interior and is zero for all other values of (u,v). The other part of $\mathbf{Q}[(u,v),\Theta_t]$ represents those parts of the background that obscure the target. It is zero for these background sections. This indicator function then is capable of modeling both target movements on a fixed background and partial obscuration of the target by the background.

The fifth submodel [Eq. (5.14)] represents the sensor by its point spread function[22] (PSF). The pixel intensity map is modeled as the spatially sampled output of a linear filter (the PSF) acting on the object plane. Sensor noise and model uncertainty are modeled by additive pixel noise.

In the overlay model, the target is completely defined parametrically by its parameter vector Θ_t. Typical components of Θ_t might include the target position in Az (azimuth) and El (elevation) or (u,v) relative to the sensor's boresight axis; the average radiance contrast of the target with respect to the background; the target's angle of rotation θ about the LOS, relative to a convenient reference orientation; and the target dimensions. Figure 5.9 presents

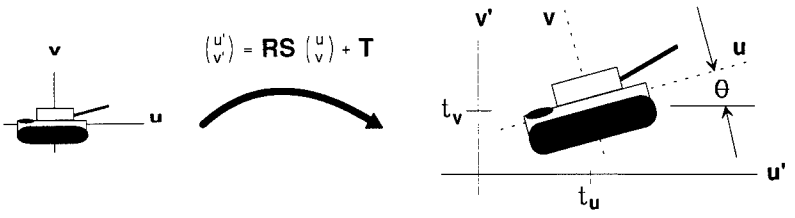


Fig. 5.9 *RST* transform used in a target parametric description.

an example of such a parametric target model. Here, an intensity template is specified and the target parameters modify the template to produce the target signature in the sensor's object plane. This approach permits the use of a simple parametric model to describe very complex objects. It also illustrates how a target's parametric description can be built up from templates. Templates can be very useful in target signature modeling because they allow signatures to be characterized more simply than by explicit mathematical descriptions.

The two-dimensional *RST* transform[23,24] used in Fig. 5.9 is given by Eq. (5.14a). This parametric form can be extended to three-dimensional transformations. In this model, the target's signature on the left side of Fig. 5.9 is modified by the transformation to produce the intensity map in the object plane. Note that in this process, the parameters of the *RST* transform are incorporated into the target's parameters.

$$I_0(u',v') = \text{contrast } I(u,v) \ ,$$

where

$$\begin{pmatrix} u' \\ v' \end{pmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} t_u \\ t_v \end{pmatrix}$$

θ = angular rotation (5.14a)
s_1, s_2 = scale changes in the u and v directions, respectively
t_u, t_v = translations in the u and v directions, respectively.

For LOS control, the translation components (t_u, t_v) are the output parameters, whereas the other symbols denote nuisance parameters, which are parameters other than the desired outputs; usually, they must be determined along with the translation components. For the preceding parameterization, the parameter vector $\Theta_t = (t_u, t_v, s_1, s_2, \theta, \text{contrast})$ would be estimated from the sensor's data to obtain the target location output parameters (t_u, t_v). In general, an estimator of Θ_c is developed and employed. The translation component of Θ_c is extracted for the target location estimate.

5.2.2.2 Maximum Likelihood Tracking. The maximum likelihood target location estimate is obtained by finding the value of the parameter vector Θ_c that yields the maximum value of the log-likelihood function $l(\Theta_c) = \ln[\Pr(Z_k|\Theta_c)]$, where $\Pr(\cdot|\cdot)$ is the conditional probability function and $Z_k = \{S_l(i,j)\}$;

$(i,j) \in R_g(l = 1, 2, ..., k)$ is the information set containing the pixel intensities on which the parameter estimate will be based.

The maximum likelihood estimate of the target's location for a broad class of target and background models has the following properties[3,5,12]:

1. The estimate is unbiased, i.e., on the average it reports the true target location.
2. The estimate has the smallest variance within the class of unbiased estimators, i.e., no estimator that, on the average, reports true target location has better accuracy.
3. The estimate can be represented approximately by Gaussian statistics. This property justifies the use of the standard deviation to characterize the target location accuracy.

A number of mathematical procedures exist for maximizing $l(\Theta_c)$. Tracker design involves selecting the best approach, given the computational constraints and the target dynamics. The difficulty that must be overcome in design stems from the complexity of the interrelations in Eqs. (5.8) through (5.14). One way of dealing with this problem is to express the target and background in terms of parameters Θ_r that change rapidly, such as the translation parameters (t_u,t_v), and other parameters Θ_s that change slowly, such as the target size. One procedure is to estimate Θ_r based on the previously determined value for Θ_s, thereby simplifying the optimization problem. Once Θ_r is determined, Θ_s is updated, treating Θ_r as a constant. This form of stepwise optimization reduces the computational cost of the MLE target location estimate.

The method can be formalized by defining the conditional log-likelihood functions $l(\Theta_r|\Theta_s) = \ln[\Pr(Z_n|\Theta_r,\Theta_s)]$ and $l(\Theta_s|\Theta_r) = \ln[\Pr(Z_n|\Theta_r,\Theta_s)]$. Because Θ_s changes slowly with respect to the sensor frame rate $\Theta_s(k) \cong \Theta_s(k - 1)$, it follows that $l(\Theta_r) \cong l[\Theta_r|\Theta_s(k - 1)]$. This implies that the optimization problem can be reduced to maximizing the simpler conditional log-likelihood function $l[\Theta_r|\Theta_s(k - 1)]$. The procedure is as follows:

Step 1. Initialize with $k = 1$ and $\hat{\Theta}_s(k - 1)$ set to *a priori* values chosen to be close to those of the actual parameters.
Step 2. Obtain image frame number k.
Step 3. Optimize $l[\Theta_r|\hat{\Theta}_s(k - 1)]$ to obtain $\hat{\Theta}_r(k)$.
Step 4. Optimize $l[\Theta_s|\hat{\Theta}_r(k)]$ to obtain the update $\hat{\Theta}_s(k)$.
Step 5. Set $k = k + 1$ and go to step 2.

To illustrate the method, consider a high-performance tracking system that uses a predictor/corrector algorithm derived from a gradient search procedure to optimize the conditional likelihood function. Predictor/corrector algorithms fit in well with the preceding procedure to yield implementable real-time tracking algorithms.[11,25]

For this example, assume Θ_b is the empty set and Θ_t is defined by Fig. 5.8 and Eq. (5.14a). Assume that (1) the distance (range) from the sensor to the target, (2) the contrast between the vehicle and the background, (3) the target rotation angle, and (4) the target dynamics change slowly compared to the sensor's sampling rate. The parameters $\Theta_s = (s_1,s_2,\text{contrast})$ will then change slowly with respect to the target location $\Theta_r = (t_u,t_v)$. Equation (5.15) gives the predictor/corrector equations for steps 3 and 4. For Eq. (5.15), assume that

the sensor noise $N_k(i,j)$ is temporally and spatially independent with a Gaussian probability density function. To limit computational requirements, only a single step in the gradient algorithm is performed for each new image frame. As long as the per-frame target movement with respect to the LOS is not excessive, little error will be introduced by this approximation.

$$\text{Step 3:}\quad \hat{t}_u(k+1) = \hat{t}_u^P(k+1) + \gamma_1 \left.\frac{\partial J}{\partial t_u}\right|_{[\hat{t}_u^P(k+1),\hat{t}_v^P(k+1)],\hat{\Theta}_s^P(k+1)}$$

$$\text{Step 4:}\quad \hat{S}_1(k+1) = \hat{S}_1^P(k+1) + \gamma_2 \left.\frac{\partial J}{\partial s_1}\right|_{[\hat{t}_u(k+1),\hat{t}_v(k+1)],\hat{\Theta}_s^P(k+1)}, \tag{5.15}$$

where

$\hat{t}_u$ = target location estimate after frame $k+1$

$[\hat{t}_u^P(k+1),\hat{t}_v^P(k+1)]$ = predicted target location based on frame (k); the target dynamics submodel [Eq. (5.11)] is used to generate both $[\hat{t}_u^P(k+1),\hat{t}_v^P(k+1)]$ and $\hat{\Theta}_s(k)$ from $[(\hat{t}_u(k),\hat{t}_v(k)]$ and $\hat{\Theta}_s(k)$. To save processor throughput requirements, $\Theta_s^P(k+1)$ is usually approximated with $\hat{\Theta}_s(k)$.

γ_1,γ_2 = constants usually chosen experimentally to give good tracking performance

J = part of the conditional log-likelihood function that depends on Θ_r:

$$J = \sum_{(i,j)\in R_g} \{\mathbf{S}(i,j) - \mathbf{C}(u,v)[\text{PSF} * \mathbf{P}(\theta_r,\theta_s)]\}^2$$

$\partial J/\partial t_u$ = rate of change of J with respect to t_u and is given by

$$-2 \sum_{(i,j)\in R_g} \left([E_{k+1}(i,j)] \times \left\{ \mathbf{C}(u,v) \left[\text{PSF} * \left(\frac{\partial \mathbf{P}(\theta_r,\theta_s)}{\partial t_u} \right) \right] \right\} \right)$$

$E_{k+1}(i,j)$ = innovation, or difference between the sampled imagery from the sensor and the predicted imagery based on $[\hat{t}_u^P(k+1),\hat{t}_v^P(k+1)]$ and $\hat{\Theta}_s^P(k+1)$ given by $\mathbf{S}_{k+1}(i,j) - \mathbf{C}(u,v)\text{PSF} * \mathbf{P}[(\hat{t}_u^P(k+1),\hat{t}_v^P(k+1)], \hat{\Theta}_s^P(k+1)]$

$[\partial \mathbf{P}(\theta_r,\theta_s)]/\partial t_u$ = rate of change of the target's object plane intensity with respect to t_u

$\partial J/\partial s_1$ = rate of change of J with respect to scale parameter s_1, given by

$$-2 \sum_{(i,j)\in R_g} \left\{ E_{k+1}(i,j)\mathbf{C}(u,v) \times \left[\mathrm{PSF}^* \left(\frac{\partial \mathbf{P}(\theta_r,\theta_s)}{\partial s_1} \right) \right] \right\}$$

$[\partial \mathbf{P}(\theta_r,\theta_s]/\partial s_1$ = rate of change of $\mathbf{P}(\theta_r,\theta_s)$ with respect to scale parameter s_1; can be determined from Eq. (5.14a).

The predictor/corrector equations for $\hat{t}_v(k+1)$, s_2, and contrast can be determined in a similar manner.

One reason that a single-step gradient algorithm such as Eq. (5.15) performs well is that, for a large class of practical problems, the log-likelihood surface is quadratic and fairly smooth near its maximum. Figure 5.10 shows contour curves for a typical log-likelihood surface about its maximum point. Gradient optimization generally works well on such surfaces. The location of the maximum can also be found by evaluating the log-likelihood function at a number of grid points such as those indicated by x's in Fig. 5.10, and then fitting a quadratic surface to these points. This procedure is an alternative to the predictor/corrector algorithm of Eq. (5.15). A combination of these two approaches is also possible. Surface fit is used to provide high-accuracy target location estimates, and target nuisance parameters are estimated using the predictor/corrector algorithm.

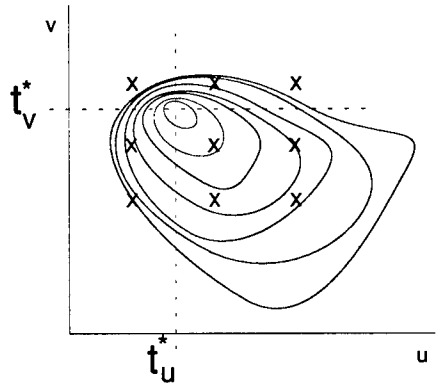


Fig. 5.10 Typical log-likelihood surface near the peak.

5.2.2.3 Maximum Likelihood Tracking Performance. The performance of maximum likelihood estimators is well documented in the literature, where the error covariance of these unbiased estimators is shown to be bounded from below by the Cramer-Rao (C-R) bound.[3,4,12] Experience shows that the C-R bound is fairly tight and therefore represents a good first approximation to the tracking accuracy. The C-R bound is given by Eq. (5.16) in terms of the Fisher information. Because Fisher information relates directly to tracking performance, it can be used to quantify tracking in terms of the overlay model parameters. The impact of different system and environmental parameters can therefore be evaluated in terms of this parameter.[26] Moreover, because Fisher information is monotonically related to tracking accuracy, it can be used to optimize the system in the same way that the SNR is used in communication system design. Fisher information serves in place of the SNR for imagery.

$$\mathbf{COV}(\Theta_c) = E\{[\hat{\Theta}_c - \Theta_c][\hat{\Theta}_c - \Theta_c]^T\} \geq [\mathbf{FI}]^{-1} , \tag{5.16}$$

where

$\hat{\Theta}_c$ = unbiased estimate of Θ_c
$\mathbf{COV}(\Theta_c)$ = covariance of Θ_c
$E\{\cdot\}$ = expectation operator
$[\ ]^T$ = transpose of matrix in the brackets
FI = Fisher information matrix about Θ_c, defined as follows:

$$\mathbf{FI} = E\left\{\left[\frac{dl(Z_n)}{d\Theta_c}\right]\left[\frac{dl(Z_n)}{d\Theta_c}\right]^T\right\} = -E\left[\frac{d^2 l(Z_n)}{d\theta_c^2}\right] .$$

The utility of Eq. (5.16) is illustrated by applying it to the assessment of target location accuracy, which is characterized by the standard deviation of t_u and t_v. If t_u is the i'th component of parameter vector Θ_c, the lower bound for the standard deviation of t_u [std dev (t_u)] is given by

$$\text{std dev}(t_u) \geq [(\mathbf{FI})^{-1}_{(i,i)}]^{1/2} , \tag{5.17}$$

where $\mathbf{A}_{(i,i)}$ refers to the (i,i) element of matrix **A**. A similar expression exists for the standard deviation of t_v.

For a few interesting tracking cases such as correlation trackers, it is possible to obtain an analytical expression for Fisher information directly from Eq. (5.16). For most tracking problems, however, numerical evaluations are necessary.

5.2.3 Correlation Trackers

Direct implementation of the maximum likelihood approach tends to be computationally intensive because of the number of nuisance parameters that must be determined. One way to reduce the computations required is to simplify the overlay model. This section describes the class of correlation trackers that can be considered as maximum likelihood trackers with simplified models. Correlation trackers tend to be less computationally stressful than the maximum likelihood location estimators obtained directly from the overlay model.

5.2.3.1 Correlation Tracker Formulations. If the following assumptions are made, the overlay model of Sec. 5.2.2 can be simplified:

1. Within the gate R_g, the number of target pixels is large compared to the number of background pixels. This allows the model to be simplified by eliminating nuisance parameters associated with the background.
2. Changes in target contrast, size, or orientation from frame to frame are insignificant. This reduces the target parameter set to the translation portion (t_u,t_v) of the RST transform.
3. Target spatial frequencies are large compared to the bandpass of the sensor's modulation transfer function (MTF). This allows the MTF of the overlay model to be approximated by the identity transfer function. This assumption can be weakened.

Based on assumptions 1 and 2, the overlay model reduces to the following:

$$\mathbf{S}_k(i,j) = \mathbf{C}(u,v)\mathrm{PSF}^*\mathbf{T}[\Theta_t(k)] + \mathbf{N}_k(i,j) \ , \tag{5.18}$$

where $\Theta_t(k) = (t_u,t_v)$ are the target parameters.

Assumption 3 allows Eq. (5.18) to be further simplified to Eq. (5.19), which is the correlation tracker model:

$$\mathbf{S}_k(i,j) = \mathbf{T}_k(i - t_u, j - t_v) + \mathbf{n}_k(i,j) \ . \tag{5.19}$$

The additive noise $\mathbf{n}_k(i,j)$, which is assumed to be IID Gaussian in Eq. (5.19), consists of two components: $\mathbf{N}_k(i,j)$ and model error. Sensor noise/pixel intensity uncertainty $\mathbf{N}_k(i,j)$ is as described for the overlay model (Sec. 5.2.2). The model error is introduced here to account for the uncertainty or error between the underlying structure generating the data and the simplified correlation tracker model of Eq. (5.19). The model error component accounts for errors arising from the three assumptions that were used to simplify the overlay model. It is difficult to determine the magnitude of the model error. One approach is to use a Monte Carlo simulation to assess the difference between the overlay model and the correlation tracker model.

For the correlation tracker model, the maximum likelihood estimator for target location (t_u^*,t_v^*) can be derived directly from Eq. (5.19). This estimate is given by the relationship in Eq. (5.20), in which $\mathbf{J}_k(t_u,t_v)$ is the least-squares cost function that quantifies the difference between the image $S_k(.,.)$ and the target reference $\mathbf{R}_k(.,.)$. Ideally $\mathbf{R}_k = \mathbf{T}_k$, so the target reference is the target's signature. The maximum likelihood estimate (t_u^*,t_v^*) then corresponds to the best least-squares fit of the reference to the sensor image.[22,27,28]

$$\begin{aligned} \mathbf{J}(t_u^*,t_v^*) &= \sum_{(i,j)\in R_g} [\mathbf{S}_k(i,j) - \mathbf{R}_k(i - t_u^*, j - t_v^*)]^2 \\ &\le \sum_{(i,j)\in R_g} [\mathbf{S}_k(i,j) - \mathbf{R}_k(i - t_u, j - t_v)]^2 = \mathbf{J}_k(t_u,t_v) \ . \end{aligned} \tag{5.20}$$

Other equivalent and useful formulations are possible. The sum-of-squares differences in Eq. (5.20) can be expanded as in Eq. (5.21). The first term to the right of the equals signs is the cross-correlation between the reference and the

image. The second and third terms on the right are the total energies in the reference and sensor images, respectively. If these terms do not depend on the target location, then maximizing the cross-correlation between the reference image and the current sensor image is equivalent to minimizing Eq. (5.20). Thus, the target location estimate (t_u^*, t_v^*) maximizes the cross-correlation between the reference and sensor images, accounting for the name *correlation tracker*.

$$\mathbf{J}_k(t_u,t_v) = -2 \sum_{(i,j)\in R_g} \mathbf{S}_k(i,j)\mathbf{R}_k(i - t_u, j - t_v) + \sum_{(i,j)\in R_g} [\mathbf{S}_k(i,j)]^2 + \sum_{(i,j)\in R_g} [\mathbf{R}_k(i,j)]^2 \ . \tag{5.21}$$

In Eq. (5.22), the cross-correlation is expressed as a discrete two-dimensional filtering of the image with the target reference. Hence, the correlation tracker can also be regarded as a matched filter in which the target location estimate corresponds to the peak of the filter's output, and the reference image corresponds to the match filter's impulse response.

$$\sum_{(i,j)\in R_g} \mathbf{S}_k(i,j)\mathbf{R}_k(i - t_u, j - t_v) = \mathbf{S}_k{*}\mathbf{R}_k = \mathbf{K}_k(t_u,t_v) \ , \tag{5.22}$$

where $\mathbf{S}_k{*}\mathbf{R}_k$ is the two-dimensional convolution of $\mathbf{S}_k(.,.)$ and $\mathbf{R}_k(.,.)$, and $\mathbf{K}_k(t_u,t_v)$ is the two-dimensional correlation surface of $\mathbf{S}_k(.,.)$ and $\mathbf{R}_k(.,.)$.

The quadratic loss function in Eq. (5.20) is a direct consequence of the Gaussian noise assumption. If the difference between the reference and the target image results from a non-Gaussian process such as gamma spikes (common in space-based sensors), the quadratic loss function may degrade performance. According to the Gaussian model, events falling into the tail of the distribution will occur so infrequently that a large weighting can safely be assigned to them. But if non-Gaussian processes are present (e.g., gamma events), they may occur with sufficient frequency to significantly degrade tracking accuracy[29] when coupled through large weighting factors. The use of more robust loss functions, such as the absolute value function |.| improves tracking accuracy in these environments.

In particular cases, there may be reasons to prefer one formulation over another. The least-squares formulation [Eq. (5.20)] is sensitive to the total energy in the image. Changes in energy (or contrast values) due to scintillation, fading, or gain variations associated with automatic gain control (AGC) can degrade track performance. The cross-correlation and the matched filter formulations are largely insensitive to image energy variations. On the other hand, if image energy is relatively constant, the least-squares formulation delivers improved tracking performance.

Selection of the correlation tracker formulation also depends on the computational environment.[30,31] The cross-correlation and the matched filter formulations both work well in a pipeline processor supporting a multiply-accumulate pipeline. The least-squares formulation requires a three-deep pipeline, which is usually fabricated in a subtract-multiply-accumulate configuration. If the number of pixels on target is large, the two-dimensional convolution [Eq. (5.22)] can be

implemented efficiently using a fast Fourier transform (FFT) processor. The matched filtering is then performed in the frequency domain.

The correlation algorithm for Eq. (5.20), (5.21) or (5.22) can easily be implemented with parallel processing. Parallel processes can be employed to evaluate discrete points on the correlation surface corresponding to a particular target location. If pointing accuracy requirements are not stressing, the point of maximum correlation can be used as the track point. If better pointing accuracy is required, the target location can be found by interpolating between points on the correlation surface.

One way to reduce computational requirements is to use an on-line version of the correlation algorithm.[32] Although the on-line correlation algorithm approximates true correlation, its tracking accuracy approaches that of Eq. (5.20). It also reduces processing latency because it does not require stored imagery.

To understand the on-line correlation algorithm, consider the tracking processes as a sequence of two modes: (1) an initial convergence mode and (2) a steady-state tracking mode. In the convergence mode, the sensor's initially poor pointing accuracy is continually improved until it reaches a steady state. In the steady-state tracking mode, corrections to the sensor's gimbal are usually small; i.e., the error between the predicted target location and the tracker algorithm's target location estimate is slight. By expanding the least-squares cost function $J_k(t_u,t_v)$ in a Taylor series about the predicted target location, the cost function can be expressed explicitly in terms of the true target location:

$$\mathbf{J}_k(t_u^*,t_v^*) = \mathbf{J}_k(t_u^p,t_v^p) + \nabla_{(t_u,t_v)}\mathbf{J}_k|_{(t_u^p,t_v^p)}[(t_u^*,t_v^*) - (t_u^p,t_v^p)] + \frac{1}{2}\mathbf{H}_{(t_u,t_v)}\mathbf{J}_k|_{(t_u^p,t_v^p)}[(t_u^*,t_v^*) - (t_u^p,t_v^p)]^2 + \text{HOT} \ , \tag{5.23}$$

where

(t_u^*,t_v^*)	= true target location
(t_u^p,t_v^p)	= predicted target location
$\nabla_{(t_u,t_v)}\mathbf{J}_k\|_{(t_u^p,t_v^p)}$	= gradient of $\mathbf{J}_k$ with respect to (t_u,t_v) at (t_u^p,t_v^p)
$\mathbf{H}_{(t_u,t_v)}\mathbf{J}_k\|_{(t_u^p,t_v^p)}$	= Hessian matrix of $\mathbf{J}_k$ with respect to (t_u,t_v) at (t_u^p,t_v^p)
HOT	= higher order terms.

As long as the prediction error $(t_u^*,t_v^*) - (t_u^p,t_v^p)$ is small (which is usually true in the steady-state tracking mode), the HOT can be ignored. If the Hessian matrix is approximated by a constant diagonal matrix, a simplified expression is obtained for the target location estimate:

$$t_u^* \cong t_u^p - g_u \left.\frac{d\mathbf{J}_k}{dt_u}\right|_{(t_u^p,t_v^p)} , \tag{5.24}$$

where g_u is a gain constant chosen for best closed-loop tracking performance and $(d\mathbf{J}_k/dt_u)|_{(t_u^p,t_v^p)}$ is the derivative of $\mathbf{J}_k$ with respect to t_u. This derivative can be approximated as

$$\sum_{(i,j)\in R_g} \mathbf{S}_k(i,j)\mathbf{R}_k(i - t_u^p + 1, j - t_v^p) - \sum_{(i,j)\in R_g} \mathbf{S}_k(i,j)\mathbf{R}_k(i - t_u^p, j - t_v^p) \ .$$

If the derivative approximation is used, the on-line correlation algorithm given by Eq. (5.24) reduces to the evaluation of two points on the correlation surface:

$$t_u^* = t_u^p + g_u \sum_{(i,j)\in R_g} \mathbf{S}_k(i,j)\mathbf{R}_k(i - t_u^p, j - t_v^p) - g_u \times \sum_{(i,j)\in R_g} \mathbf{S}_k(i,j)\mathbf{R}_k(i - t_u^p + 1, j - t_v^p) \ .$$

A similar expression holds for t_v^*.

The assumption on which this algorithm is based is the closeness of the predicted target location to (t_u^*, t_v^*). This condition is usually satisfied in the steady-state track mode; hence, the accuracy of the on-line formulation is expected to be close to that of the complete correlation algorithm. During the initial convergence, however, this condition can be violated, reducing the accuracy of the on-line correlation algorithm. Moreover, convergence is not guaranteed, so a detailed simulation study is needed to establish a zone of convergence.

5.2.3.2 Correlation-Tracking Performance. This section presents a performance analysis of the correlation-tracking algorithm. Because the correlation tracker was shown to be a particular case of maximum likelihood estimation, all analysis methodologies described in Sec. 5.2.2.2 are applicable to correlation tracking. Section 5.2.2.2 established that Fisher information could be used to evaluate the accuracy of any MLE tracker. For correlation tracking with IID Gaussian noise, the analytical expressions for the Fisher information and the error covariance are given by Eq. (5.25) and Eq. (5.26), respectively.

$$\mathbf{FI} = \frac{N_G}{\sigma^2} \begin{Bmatrix} \sum_{(i,j)\in R_g} \left[\frac{\partial T}{\partial u}(i,j)\right]^2 & \sum_{(i,j)\in R_g} \left[\frac{\partial T}{\partial u}(i,j)\frac{\partial T}{\partial v}(i,j)\right] \\ \sum_{(i,j)\in R_g} \left[\frac{\partial T}{\partial u}(i,j)\frac{\partial T}{\partial v}(i,j)\right] & \sum_{(i,j)\in R_g} \left[\frac{\partial T}{\partial u}(i,j)\right]^2 \end{Bmatrix} , \quad (5.25)$$

where

N_G = number of pixels in the gate R_g

σ^2 = variance of $n_k(i,j)$ in Eq. (5.25); the noise term is assumed to be temporally and spatially independent with a Gaussian probability density function

$\partial T/\partial u$, $\partial T/\partial v$ = rate of change of the target template $T(\cdot,\cdot)$ in the u and v directions, respectively.

The Cramer-Rao bound can be used to obtain an approximation for correlation tracking algorithm performance from Fisher information:

$$\mathbf{COV}(\hat{t}_u, \hat{t}_v) \geq \frac{\sigma^2}{N_G \Delta} \begin{Bmatrix} \sum_{(i,j)\in R_g} \left[\frac{\partial T}{\partial v}(i,j)\right]^2 & \sum_{(i,j)\in R_g} \left[\frac{\partial T}{\partial u}(i,j)\right]\left[\frac{\partial T}{\partial v}(i,j)\right] \\ \sum_{(i,j)\in R_g} \left[\frac{\partial T}{\partial u}(i,j)\right]\left[\frac{\partial T}{\partial v}(i,j)\right] & \sum_{(i,j)\in R_g} \left[\frac{\partial T}{\partial u}(i,j)\right]^2 \end{Bmatrix} , \quad (5.26)$$

where $\mathbf{COV}(\hat{t}_u,\hat{t}_v)$ is the covariance matrix of the correlation tracker's target location estimate $(\hat{t}_u,\hat{t}_v)$ and

$$\Delta = \sum_{(i,j)\in R_g}\left[\frac{\partial T}{\partial u}(i,j)\right]^2 \sum_{(i,j)\in R_g}\left[\frac{\partial T}{\partial v}(i,j)\right]^2 - \left\{\sum_{(i,j)\in R_g}\left[\frac{\partial T}{\partial u}(i,j)\right]\left[\frac{\partial T}{\partial v}(i,j)\right]\right\}^2 .$$

Equation (5.26) is adequate for first-order tracking system analysis, but it should be supplemented by a quality simulation study because the correlation tracker model [Eq. (5.18)] is often at variance with measured data.

5.2.4 Adaptation, Learning, and Drift

The tracking models described in Secs. 5.2.2 and 5.2.3 utilized parametric descriptions of the target and/or background that were assumed to be accurate within certain limits. Any uncertainty in these parametric models was included in the noise terms. But because tracking performance degrades as the magnitude of the noise term increases (see Secs. 5.2.2.3 and 5.2.3.2), such uncertainties are a source of tracking error. When the target's signature cannot be accurately modeled parametrically, the technique of target signature adaptation or reference adaptation, described in this subsection, can improve tracking performance.

Reference adaptation is usually implemented by estimating the target's predicted intensity distribution[33] using a temporal filter on previously registered target images. Either finite impulse response (FIR) filters [Eq. (5.27)] or infinite impulse response (IIR) filters [Eq. (5.28)] can be used. Weighting functions should sum to unity to ensure that the target and its reference have the same amplitude.

$$\hat{\mathbf{R}}_{k+1}(i,j) = \sum_{l=0}^{m} \mathbf{w}_l(l)\mathbf{S}_{k-l}(i - i_l, j - j_l) , \tag{5.27}$$

$$\hat{\mathbf{R}}_{k+1}(i,j) = \mathbf{w}_f(0)\mathbf{S}_k(i - i_k, j - j_k) + \sum_{l=1}^{m} \mathbf{w}_f(l)\hat{\mathbf{R}}_{k-i}(i,j) , \tag{5.28}$$

where

$\hat{\mathbf{R}}_k$ = estimated target's signature for frame number k
$\mathbf{w}_f(\cdot)$ = weighting function for the FIR filter
$\mathbf{w}_l(\cdot)$ = weighting function for the IIR filter
(i_l,j_l) = target location estimate for the l'th image.

In choosing the weighting function, both random sensor noise and the rate of change of the target signature must be taken into account. If the noise level is high and the signature changes slowly, then a weighting function with a long effective time constant is desirable. On the other hand, if little noise is present and the target signature changes rapidly, tracking performance will be improved by using a smaller effective time constant. The effect of noise on tracking performance can be seen by considering the instantaneous noise term $\mathbf{n}_c$ in the correlation tracker model of Eq. (5.18). This term can be regarded

as the sum of three components. The first component depends on the sensor's random noise $\mathbf{N}_s$, the second depends on random noise in the adaptive reference $\mathbf{N}_r$, and the third models target signature uncertainty in the reference $\mathbf{N}_u$. It is reasonable to assume that the three terms are mutually independent.

$$\mathbf{n}_c(i,j) = \mathbf{N}_s(i,j) + \mathbf{N}_r(i,j) + \mathbf{N}_u(i,j) \ . \tag{5.29}$$

A simple example provides some insight into the relationships involved. Consider a first-order FIR filter for the adaptive reference. Assume that the variability of the target signature results entirely from platform roll—a situation typical of some fire-and-forget missiles. For low roll rates, a smaller filter parameter is preferable because it reduces the effect of the second noise term of Eq. (5.29), and the low roll rate does not produce a dominant third term. If the roll rate is high, tracking performance is improved by setting the parameter to a larger value. An algorithm might be implemented to estimate roll rate, thereby allowing the filter parameter to adjust dynamically.

The foregoing analysis addressed tracker performance when the sequence of input images is perfectly registered (i.e., the target image is estimated perfectly). Short of performing complex image interpolations,[34] image alignment accuracy is limited by 1-pixel spatial quantization error. This implies that reference adaptation always entails registration errors. The impact of registration errors is aggravated by the circular nature of the correlation algorithm using reference adaptation. The circular sequence is as follows: (1) the adaptive reference is used to compute the target's location, (2) the computed target location is then utilized to generate the reference for the next frame, and (3) the cycle repeats. If an error (such as a registration error) occurs in the target location computation, it will propagate around the circular loop and accumulate. Consequently, errors (including registration errors) can build up, leading to loss of lock.

Tracker error growth resulting, for example, from misregistration in the adaptive reference algorithm is referred to as *tracker drift*. A human observer would see the tracker cross hairs drift slowly off the target until loss of lock occurred. Drift rate can be reduced by carefully selecting the filter parameters, but it cannot be totally eliminated. The drift error and random reference error are linked so that choosing a filter parameter value to reduce the drift rate leads to an increased random reference error. Unless other information is utilized to correct the alignment in the adaptation process, the registration errors continue to accumulate and degrade tracking performance. Section 5.2.5 describes a method for eliminating drift in the adaption process.

5.2.5 Multimode Tracking

Multimode tracking represents a synergistic approach that utilizes a variety of tracking techniques (centroid, correlation, etc.) to overcome the limitations inherent in any single-mode tracker. The multimode philosophy stems from the observation that certain trackers work better in some situations than others, and the best overall performance can be achieved by incorporating several algorithms into one tracker. Thus, if it were possible to determine in advance which tracking approach would work best for a particular situation,

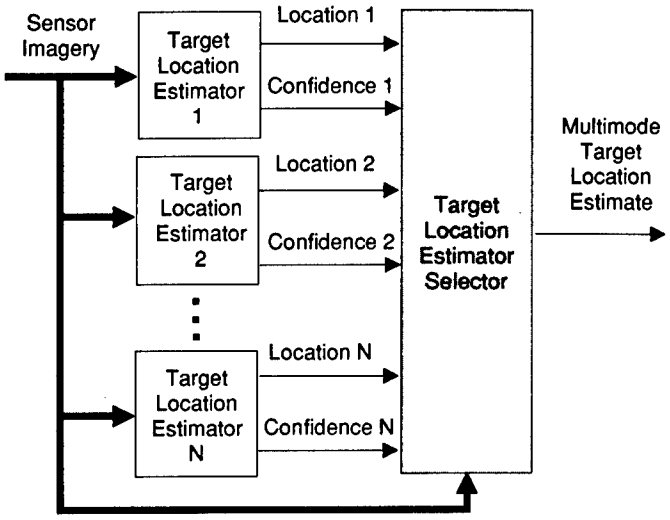


Fig. 5.11 Multimode tracking architecture using tracker selection.

optimum tracking performance could be realized by selecting it in that situation. Figure 5.11 shows a multimode tracker architecture based on this observation.

Multimode tracking can be illustrated in a typical scenario by considering a correlation tracker following a tank that is firing small, very bright flares. The correlation-tracking algorithm seeks to align the incoming target image with the reference image of the tank (without flares). While the target is firing its flares, the current imagery differs significantly from the model inherent in the correlation tracker. Hence, the correlation tracker may be unable to achieve its expected tracking performance. The binary centroid tracker, on the other hand, is relatively insensitive to intensity variations that occur within the target image (such as the "hot spot" caused by the flare), whereas the video gate suppresses effects due to changes that occur away from the target. Therefore, a binary centroid whose video map is set to eliminate the hot spot should provide satisfactory track performance, even while the flare is within the track gate. The multimode tracker would use the correlation algorithm until it detected the presence of the flare; then it would switch to the centroid algorithm until the flare left the vicinity of the target.

The design challenge for this type of multimode tracker is the selection algorithm. Designs for this algorithm are based on heuristic rather than analytical approaches. The heuristic method is dictated by the complexity of the target and background characteristics. The dynamic behavior of the target and/or background must be taken into account in the design of a multimode tracker. If there were no dynamic changes, a multimode solution would not be necessary. A model of the dynamic behavior is needed for simulation studies to validate heuristic algorithms.

If the correlation tracker model [Eq. (5.18)] could describe the sensor data perfectly, the tracker selection block should then select the correlation tracker.

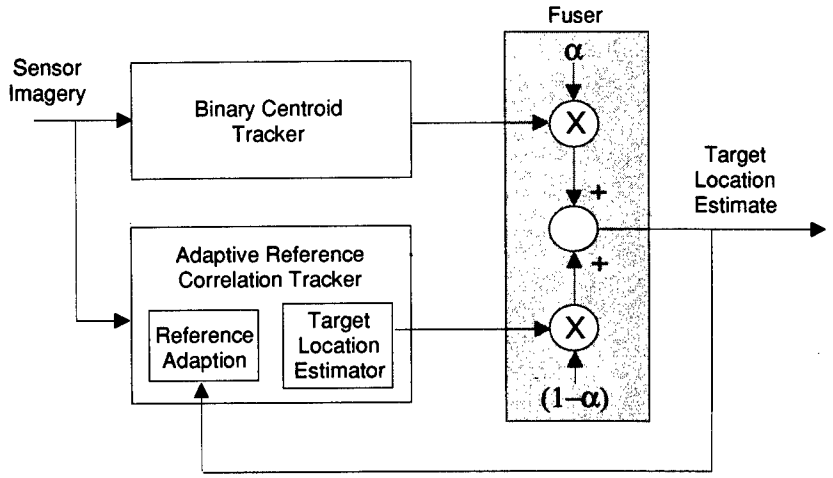


Fig. 5.12 Example of a multimode tracker using data fusion.

On the other hand, if significant errors exist between the sensor data and the correlation tracker model, track performance might be improved by choosing a different algorithm. Thus, the key to the tracker selection algorithm lies in examining and responding properly to differences between the model on which a tracking approach is based and the true sensor data. Complexities in the true sensor data make heuristic approaches attractive for the selection algorithm, but extensive simulation studies are needed to refine it.

An alternative approach to multimode tracking employs data fusion to combine the outputs of several different tracking algorithms.[25] For example, one version of a multimode tracker uses two tracking algorithms: (1) an adaptive reference correlation tracker and (2) a binary centroid-tracking algorithm. The adaptive reference correlation algorithm offers nearly optimal performance but suffers from the drift problem described in Sec. 5.2.4. The centroid algorithm, on the other hand, has no appreciable drift. If the fusion block (Fig. 5.12) combines the target location reports from both the centroid and correlation algorithms to define the target location estimate used to update the reference, errors do not accumulate in the reference. Thus, the multimode tracker does not drift. In this case, the centroid algorithm helps to overcome an inherent limitation of the adaptive reference correlation algorithm.

5.2.6 Comparison of Target Location Algorithms

A number of different target location estimation approaches have been presented. They cover a wide range of capabilities from simple hardware implementations to extremely accurate algorithms. It is difficult to make simple generalizations comparing target locations without becoming inaccurate and misleading. However, Table 5.2 does provide a high-level contrast of the different target location algorithms and can be used for first-pass algorithm selection.

Table 5.2 Comparison of Target Location Algorithms

Algorithm	Target	Background	Performance	Hardware Implementation
Binary centroid	Gray levels significantly different from background	Prefer uniform with small range of gray levels	Easily corrupted by structured clutter	Very simple
Intensity centroid	High intensities relative to background	Should be uniform with no significant intensities in gate	Very good for some target intensity distributions and bad for others	Simple
Thresholded intensity centroid	High intensities relative to background	Best if background intensities are different from target intensities	Can be very good; more robust than other centroid algorithms	Simple
Fixed reference correlation	Known signature	Best against uniform intensity distribution	Very high accuracy but poorly performing if target signature deviates from expected signature	May require moderate throughput
Adaptive reference correlation	Unknown signature	Best against uniform intensity	Very high accuracy; robust to changing target signature	May require large throughput
Multimode	Unknown signature	Can include unknown structured clutter	Very high accuracy; no drift; robust to structured clutter	May require large throughput
Maximum likelihood	Characterized parametrically	Characterized parametrically	Excellent	May require large throughput

Because, for a particular tracking application, it is often easy to modify any one of the target location algorithms to meet requirements, the trade-off should be expanded beyond what was selected on the first pass.

5.3 TARGET GATING AND BREAKLOCK DETECTION

Target gating and breaklock detection are two supporting functions that help to ensure good tracking system performance over a wide range of environments.

The target location estimation algorithms described in Sec. 5.2 presume that the target's intensity distribution is characterized in a way that differentiates it from the background. If this characterization deviates substantially from the true target signature, or if the target cannot be distinguished from the background (e.g., because of clutter), then significant error in the target location estimation can occur. The purpose of the breaklock function is to detect such conditions and prevent erroneous target location estimates from influencing the LOS control, thereby improving tracking performance. Section 5.3.4 describes approaches to breaklock detection.

Target gating has two purposes. First, it limits the amount of pixel data that must be processed. Because the target gate typically covers only a small fraction of the sensor's image, bus and processor loading is reduced. Second,

restricting the target location computation to the target gate reduces the likelihood that noise and competing clutter or other targetlike artifacts will corrupt the target location estimate.

The dimensions of the target gate are often adaptable to accommodate targets with changing sizes. Selection of a gate adaptation algorithm is a critical step in tracking system design because the target gate can significantly influence the performance of the tracking system. The relationship of tracking performance to target gating is the subject of Sec. 5.3.1. Sections 5.3.2 and 5.3.3 present an overview of some techniques and approaches that have been found useful for choosing a target gate algorithm.

5.3.1 Gate Construction and Track Performance

In the binary centroid tracker, the centroid is computed only over the target gate R_g. Hence, errors in target gate sizing can influence tracking performance. For example, if the gate is smaller than the target, the binary centroid will differ from the true area centroid because the target area outside the gate is excluded from the computation. The magnitude of this error is proportional to the degree of undersizing. Target location errors can also be caused by an oversized gate.

For the binary centroid tracker, the impact on target position accuracy can be estimated using the results of the regionalization analysis in Sec. 5.2.1.1. Figure 5.13 shows the predicted impact on binary centroid track accuracy in the presence of a uniform background derived using methods in Sec. 5.2.1.1. Note that the performance drops off gradually as the gate size increases, because of the increasing probability of noise exceedances as more nontarget pixels are enclosed by the gate. If the background is not uniform, an increase in gate size has a more pronounced impact on track performance. The sensitivity to gate size for specific backgrounds can also be assessed using the results of Sec. 5.2.1.1.

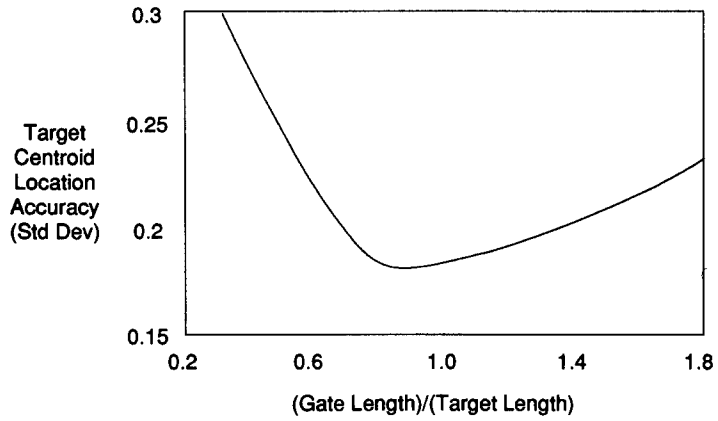


Fig. 5.13 Track accuracy as a function of gate size.

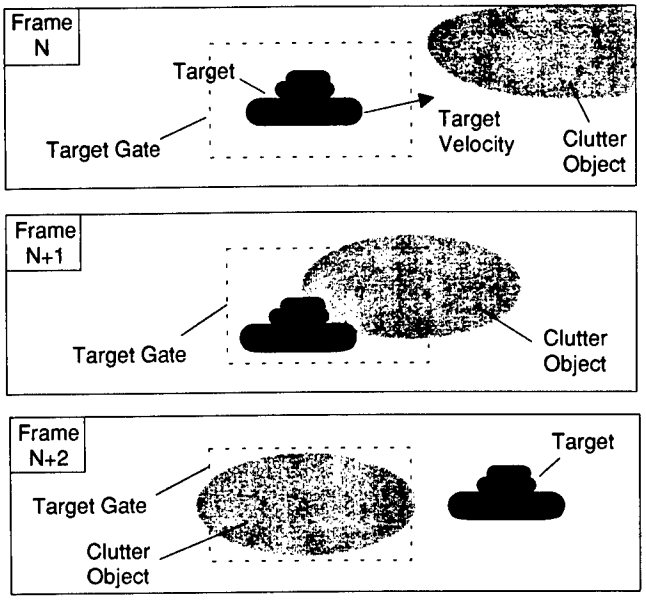


Fig. 5.14 Track sequence illustrating loss of lock caused by improperly sized gate.

The impact of gate size on the performance of the correlation algorithm is similar to its effect on the binary centroid algorithm. Tracking performance is optimized when the gate most closely matches the target, and falls off as the gate is either increased or decreased relative to the optimum size. When the gate is undersized, tracking jitter is increased because some target pixels are excluded from the target location computation. Tracking performance also degrades when the gate is oversized, because background pixels and additional noise then influence the target location estimate. As a rule, the gate should include the edges of the target, which usually contain a large intensity gradient. In Sec. 5.2.3.2, it was shown that tracker performance is directly related to intensity gradients in the image.

In the presence of a uniform background, tracking accuracy falls off gradually as the gate is increased beyond its optimum size. In a cluttered environment, however, tracking performance may degrade sharply if a portion of the clutter is allowed to enter the gate. This phenomenon is best illustrated by an example. Figure 5.14 shows a tracking sequence employing an oversized target gate moving past a large clutter object whose intensity is comparable to that of the target. When the target is close to the clutter object, the target location estimate is distorted by the presence of the clutter pixels included in the gate. Eventually, the target moves away from the clutter object, but in the meantime, the centroid algorithm has locked onto the clutter, causing loss of lock.

Methods of constructing the target gate can be divided into two classes, based on the target location estimation algorithm being supported. Class 1 includes target location algorithms that generate their results in a single pass

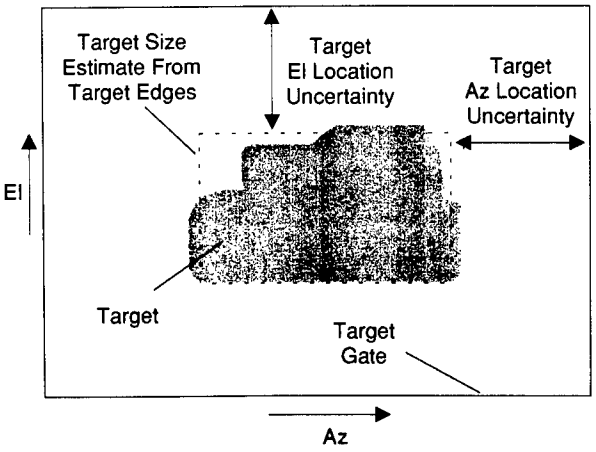


Fig. 5.15 Gate geometry for a single-pass target location estimation algorithm.

over the sensor's image data, such as the centroid algorithm. The gate for this class must be sized to ensure that it includes the target, with an allowance for the uncertainty in the target location estimate. Class 2 includes algorithms, such as the correlation algorithm, that search the image, point by point, for the target. For this class, the gate encompasses only the pixels needed to determine the likelihood that the target is at a particular location.

The manner in which the gate is used differs significantly between the two classes, and, as a result, the gate construction is also different. The difference centers on the way in which uncertainty about the predicted target location is utilized in the gate construction process. For the class 1 one-pass algorithms, the gate must be expanded to include the uncertainty in the target's predicted location. Figure 5.15 illustrates the gate geometry for this class. For class 2, the gate size is not influenced by uncertainty in the predicted location of the target. Instead, it is determined directly from the size of the target's image, whereas the predicted target location uncertainty is used to define the search grid, as shown in Fig. 5.16.

Gate construction can be described on a formal basis as follows. First, let $x_{\text{lgth}}(k)$ and $y_{\text{lgth}}(k)$ be the estimated target lengths in the two image plane directions, based on the k'th frame of data. Denote the predicted target location uncertainty by $\sigma_x(k+1)$ and $\sigma_y(k+1)$; these are the standard deviations of the prediction error at the time that the $(k+1)$'th image frame is received.

For class 1 (single-pass target location estimators), the gate R_g, defined as x_{size} by y_{size}, for use with frame $k+1$ is given by

$$x_{\text{size}} = \alpha\sigma_x(k+1) + \sum_{i=0}^{m} w_i x_{\text{lgth}}(k-i) \ ,$$

$$y_{\text{size}} = \alpha\sigma_y(k+1) + \sum_{i=0}^{m} w_i y_{\text{lgth}}(k-i) \ , \tag{5.30}$$

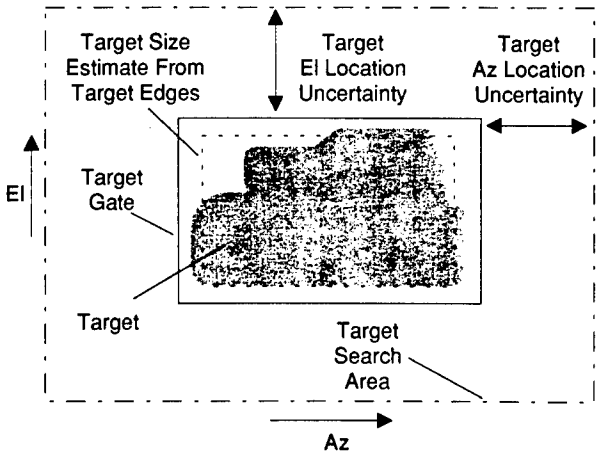


Fig. 5.16 Gate and search geometry for a searching target location estimation algorithm.

where the scale factor α is usually a value between 2 and 6 to ensure a high probability that the entire target will be included in the gate. The weighting function $\{w_i\}$, normalized so $\Sigma_{i=0}^{m} w_i = 1$, increases monotonically with decreasing i, so the most recent image frames are weighted more heavily than earlier frames. Performing a weighted average over the most recent $m + 1$ image frames serves to reduce the sensitivity of the target length computation to estimation errors. This reduction in sensitivity exploits the fact that target size tends to change slowly so that the size estimation error can be reduced by averaging. The integer m should be based on the rate of change of target size, e.g., if the target size changes rapidly, then a smaller value of m is preferable to a larger one.

Some earlier tracking system performance problems have been traced to poor gate sizing, particularly in a high noise and/or clutter environment. In these situations, inferior target gating combined with a poor SNR caused tracking breaklock, even though the target location estimation algorithm was still performing well. Algorithms such as that in Eq. (5.31) can improve tracking performance by reducing gate sensitivity to noise and local transient clutter.

For class 2 (searching target location algorithms), the gate can be constructed using an approach similar to Eq. (5.31):

$$\begin{aligned} x_{\text{size}} &= \sum_{i=1}^{m} w_i x_{\text{lgth}}(k - i) \ , \\ y_{\text{size}} &= \sum_{i=1}^{m} w_i y_{\text{lgth}}(k - i) \ , \end{aligned} \tag{5.31}$$

and the search size is given by

$$\begin{aligned} x_{\text{search}} &= \alpha\sigma_x(k + 1) \ , \\ y_{\text{search}} &= \alpha\sigma_y(k + 1) \ . \end{aligned} \tag{5.32}$$

The foregoing equations demonstrate that the gate sizing process must account for the uncertainty in the target location predicted for the next frame. High-performance on-line prediction methods, such as Kalman filtering, provide estimates of both the predicted value of the target location and its standard deviation. These can be used directly in the gate construction algorithms previously described. The relatively high computational cost of the full Kalman filter can be reduced by using an approximation for the predictor,[9] e.g., assuming that the target remains at the same pixel location estimated on the previous frame. This crude approximation for the target location prediction was employed in earlier tracking systems. Crude location predictions have larger standard deviations and yield larger gate sizes, which can increase the tracker's sensitivity to noise and clutter.

Although $\sigma_x(k+1)$ and $\sigma_y(k+1)$ are determined by the target location prediction algorithm, the target length estimates $x_{\text{lgth}}(k)$ and $y_{\text{lgth}}(k)$ are obtained through image analysis, which is discussed in Secs. 5.3.2 and 5.3.3.

5.3.2 Target Length Estimation Based on Histogram Analysis

If the edges of the target can be identified, then its length can be found from the edge locations. Because of the relative simplicity of the hardware implementation required, many early tracking systems employed the technique of histogram analysis to locate the target edges.

One example of histogram analysis for target edge determination is based on a geometry employing three gates. These gates contain (1) the target only (interior gate), (2) the local background only (background gate), and (3) both background and target (edge gate). Histogram data is collected over the background and interior gates. The resulting histograms are then used to classify pixels within the edge gate. If the interior-gate histogram value of an edge-gate pixel is larger than its background-histogram value, the pixel is classified as a target pixel, because its intensity more closely matches that of the target. Otherwise, the pixel is classified as a background pixel. The location of the target edge is then determined from the fraction of target pixels in the edge gate. Figure 5.17 illustrates this gate geometry and the estimated edge location based on this algorithm.

A formal description of this algorithm is

$$\text{offset} = \frac{\sum_{\text{edge gate}} \begin{cases} 1 & \text{if } \boldsymbol{Hist}_i[\mathbf{S}_k(i,j)] > \boldsymbol{Hist}_b[\mathbf{S}_k(i,j)] \\ 0 & \text{otherwise} \end{cases}}{\text{number of pixels in edge gate}}, \tag{5.33}$$

where $\boldsymbol{Hist}_i(\cdot)$ and $\boldsymbol{Hist}_b(\cdot)$ are the interior and background histograms, respectively.

Because this edge algorithm uses fewer target pixels than are used in the target location estimator, it is also noisier than the latter. Moreover, because poor target gating reduces the accuracy of the target location estimate, it follows that tracking performance may be limited by the accuracy of the target edge determination. Hence, in tracking system design, care must be exercised to minimize the noise associated with the edge algorithm.

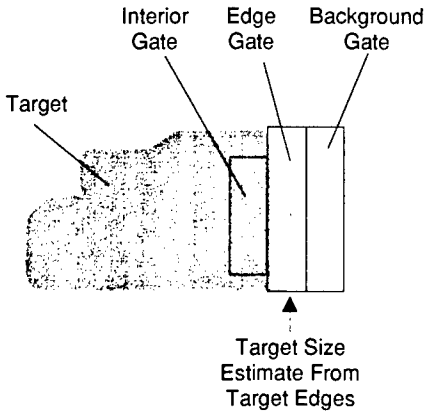


Fig. 5.17 Gate geometry for target edge location.

Noise effects can be minimized in several ways. First, the noise reduction can be achieved by increasing the size of the edge gate. But this improvement is offset by the fact that a larger edge gate may also reduce the accuracy of the estimate, because more background or clutter will be included in the computation. Another method of increasing noise immunity involves the interior and background histograms. Because typically, the background and target signatures do not change rapidly, the accuracy of the histograms can be improved by averaging over a number of previous sensor image frames.

The target gating approach previously described is predicated on the assumption that the target size and intensity distribution are relatively static. In certain severe tracking environments, however, this assumption is not valid. For example, a jinking aircraft produces an image whose size and shape repeatedly undergo sudden changes. To deal with such cases, the target gating approach can be supplemented with a target size model (part of the overlay model).[35] Another potential limitation of this approach concerns the shape of the target. The gating algorithm works well for blunt-edged targets and poorly for targets with long sloping edges.

If the edge algorithm cannot perform reliable target edge gating, then it may be necessary to employ sophisticated image processing algorithms from automatic target recognition to determine the target's boundary. These methods are described in Sec. 5.3.3.

5.3.3 Target Segmentation

Target gating is similar to the problem of target segmentation in automatic target recognition (ATR) systems. Target segmentation[36,37] classifies pixels into one of two categories: target or surrounding background. In this way, the target is "segmented" from the background. In ATR systems, this technique is used to obtain target shape information for discrimination purposes. In tracking systems, the target gate can be obtained directly from the segmented target.

A number of very sophisticated target segmentation algorithms have been developed in recent years. Two image processing techniques commonly used in target segmentation are edge following and region merging. If the target edges can be identified, then its perimeter can be determined by following the edges around the target. Region merging divides the subimage containing the target into regions (or blobs) based on intensity and/or texture. These regions are then merged together using a decision tree to obtain a set of regions that define the target. The literature contains many other segmentation algorithms.

In automatic target recognition systems, target segmentation must be carried out before recognition can be achieved. For tracking systems, because segmentation is used only to determine the target gate, it can be executed in a background mode. The target gate usually does not change rapidly, so segmentation results from previous frames can be used in the current frame's target location estimation. Furthermore, unless the target size is changing rapidly, the segmentation algorithm need not execute after every frame of sensor data. Because target segmentation tends to be computationally intensive, activating it in the background mode every few seconds unburdens the processor.

Another difference between the use of target segmentation for automatic target recognition and for tracking stems from the ability of a tracking system to enhance the image SNR. The SNR has a direct bearing on the segmenter's performance. Because the tracking system updates the target location estimate on each frame, the image SNR can be enhanced by averaging several properly registered (aligned) frames, using the target location estimate. For this reason, a segmentation routine that would prove unsatisfactory for automatic recognition of a faint target may perform well for tracking.

5.3.4 Breaklock Detection Algorithms

One way to gauge the performance of a tracking system is to assess the robustness of the track loop to changes in the target and/or background characteristics. Examples include signature changes that occur when (1) the target expels a flare as a countermeasure and (2) the target-to-background contrast decreases to a very low value. In both of these examples, a breakdown of the target location estimation is likely to occur. Such a breakdown can lead to track loop failure (breaklock) if it is not detected and compensated for. In at least some cases, a breaklock algorithm can be devised that will detect a breakdown condition just before the target location estimator becomes unreliable. By disconnecting the unreliable target location estimator, the robustness of the track loop is enhanced.

The design of a high-quality breaklock algorithm is complicated both by the random noise present in the sensor's output and by the wide variety of potential breaklock conditions. Because the conditions that the breaklock algorithm must protect against are often complex, heuristic approaches to breaklock algorithm design have usually been adopted. Sometimes this results in a crude algorithm that declares a breaklock condition prematurely, i.e., before the target location estimator actually becomes unreliable. A breaklock algorithm that behaves in this way limits the performance of the entire tracking loop. A breaklock algorithm may also prove unsatisfactory because it does not in-

dicate when the target location estimator has become unreliable. Careful analysis is required to balance the competing goals of reliable breaklock detection, robustness against breaklock, and algorithm complexity.

One heuristic approach involves a set of confidence parameters associated with the target location estimation process (Fig. 5.11). These confidence parameters are usually chosen for situations that have caused problems in previous tracking systems. For example, it is known that a high-clutter background can cause a centroid tracker to breaklock. To improve tracking performance under such conditions, a confidence parameter keyed to the standard deviation of the background might be employed. If the standard deviation was found to exceed a predetermined value, the confidence parameter would be lowered.

Another breaklock detection technique, which is based on classical detection theory, utilizes the model developed for the target location estimation algorithm. As described in Sec. 5.2.2, the estimate generated by the target location algorithm is based on the best fit of the model to the sensor's data. The degree of fit can be regarded as a measure of how well the target location estimation algorithm is performing,[28] and it can be used as a confidence parameter for breaklock detection. In the tank-expelling-a-flare example, the model (which does not include the flare) will be at variance with the image provided by the sensor. The existence of a large residual difference between the real image and the model then indicates a breaklock condition.

This approach can be expressed mathematically using the sensor's data model:

$$\mathbf{S} = \mathbf{M}(\Theta^*) + \mathbf{N} \ , \tag{5.34}$$

where

$\mathbf{S}$ = sensor data
$\mathbf{M}(\cdot)$ = mathematical model used for estimating the target's location
Θ^* = best-fit parameter vector containing target location
$\mathbf{N}$ = all the noise components contained in the system affecting the sensor data, including characteristics not represented by $\mathbf{M}(\cdot)$.

The breaklock test is formed by summing the square of the residuals over the target gate:

$$\sum_{(i,j)\in R_g} [\mathbf{S}(i,j) - \mathbf{M}(i,j)|_{\Theta^*}]^2 \quad \begin{cases} > Thr \Rightarrow \text{Breaklock} \\ \leq Thr \Rightarrow \text{No Breaklock} \end{cases} . \tag{5.35}$$

This test can often be performed with little extra computational loading because the summation can be performed as part of the target location estimation algorithm. Implicit in this breaklock algorithm is the assumption that the noise $\mathbf{N}$ is Gaussian. If this assumption is not appropriate, then a more general detection approach such as the generalized likelihood ratio test or the multiple model detection algorithm should be considered.

5.4 LOS DETERMINATION AND TRACK LOOP COMPENSATION

5.4.1 LOS Determination

In the tracking architecture of Sec. 5.1, the LOS determination component has two functions: (1) using the target location estimate and other measurements to determine the target LOS state vector and (2) predicting the next-frame target location for use in the target location estimator. As part of the target location prediction function, the LOS determination component might also provide the prediction accuracy for use in gating or search area definition (Sec. 5.3). LOS state vector determination and location prediction estimation are strongly related functions. If a Kalman filter is used for LOS determination, then both of these functions can be performed in the filter algorithm.[38]

The LOS state vector is commonly defined by the gimbal positions and velocities in azimuth and elevation. Other state definitions are sometimes used, but this definition is suitable for most tracking problems. The earliest tracking systems often did not estimate or use LOS velocities. In these systems, the LOS position determined by the target location estimator was forwarded directly to the loop compensator, as though the LOS velocity were zero. These early systems worked well for static pointing or for tracking slowly moving targets. To maintain a stable LOS against more rapidly moving targets and/or when tracking from a moving sensor platform, however, it is often necessary to include the LOS velocity in the state vector.[14]

Another input to the LOS determination component is the breaklock flag, which is set by the breaklock detector when the target location estimate is unreliable. When the breaklock flag is set, the LOS determination does not utilize the target location estimate to compute the LOS state vector. This opens the track loop. In this open-loop condition, the tracker is said to be operating in the *coast mode*. Coasting continues until the target is reacquired and the target location estimate again becomes reliable. Even then, however, target location prediction may not become sufficiently reliable for several frames, until the adaptation process in the target location estimation has settled down. As a consequence, if the target location estimation algorithm is adaptive, as discussed in Sec. 5.2, then a transition mode between the coasting and tracking modes may be advisable.

In the coast mode, the open-loop pointing commands are obtained exclusively from the LOS determination component. In effect, the LOS determination algorithm continually predicts the target's LOS, which is then transferred to the gimbal control to carry out the required pointing commands. In the absence of target location estimates, the error between the true target LOS and the output from the LOS determination component will continue to grow, as a result of both inaccuracies in the LOS determination gimbal pointing process and unpredicted target motion. Over time, this error grows so large that target reacquisition becomes impossible, when, for example, the target is no longer within the sensor's FOV. At this point, the tracking system should report "loss of track," which signifies that intervention by the operator or another system is required if tracking is to continue. An important performance measure for

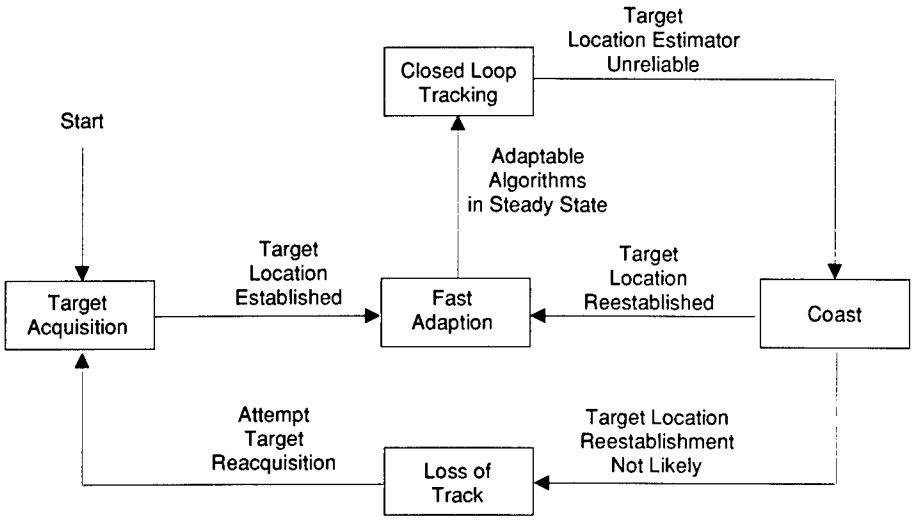


Fig. 5.18 Mode diagram of a typical tracking system.

a tracking system is the length of time it can be expected to coast before loss of track is declared. Figure 5.18 shows the mode diagram for a typical tracking system including the modes described here.

5.4.2 Track Loop Compensation

The track loop compensator uses the state vector developed in the LOS determination component to generate gimbal pointing commands. The compensator, which is a sampled data transfer function[13] with a sampling time equal to the sensor's frame time T_f, is designed to achieve the desired closed-loop error performance. In the general case, the entire LOS state vector is utilized in developing the gimbal pointing commands. In many cases of practical interest, however, little performance degradation results if the LOS vector is resolved into its orthogonal components, Az and El, and a separate loop compensator is constructed for each component. This greatly simplifies the design (Fig. 5.19).

The performance of the compensator (and hence, the tracker) is often limited both by the sensor's frame rate and by the amount of latency in the track loops (*loop latency* is also called *transport delay*). Increasing the sensor frame rate to improve compensator performance adds substantially to the computational requirements because the processing must then be completed in a shorter time. Similarly, reducing loop latency can also increase throughput requirements. In most tracking system implementations, loop latency tends to be the major influence on the closed-loop dynamics.

Loop latency arises from two major sources: the sensor and the processor. Figure 5.20 shows a loop latency time line typical of a staring sensor. In this figure, the frame time is the sum of the photon collection time of the sensor and the time required to transfer the image data to the processor. Loop latency is calculated from the midpoint of the photon collection period. Once the image data is transferred to the processor, the target location estimation, LOS de-

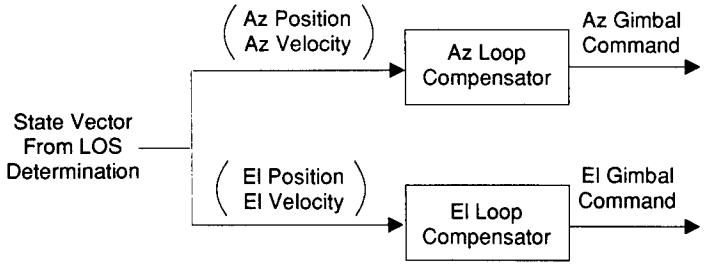


Fig. 5.19 Separable loop compensator architecture.

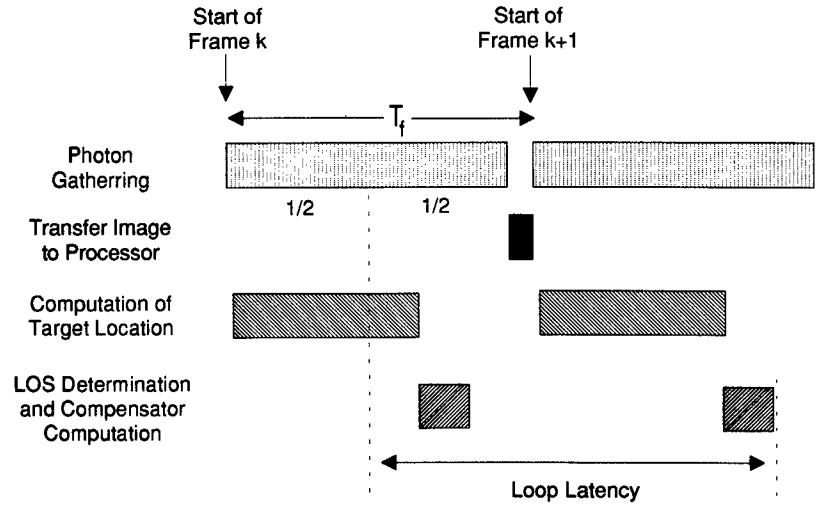


Fig. 5.20 Typical loop latency time line.

termination, and track loop compensation functions are all performed before the gimbal pointing command is initiated. The loop latency is therefore equal to ½ the photon collection time, image transfer time, and all of the processing time before the gimbal command. The processing latency part can be reduced by deferring any processing that is not directly required for generating the current frame's gimbal pointing command until after the gimbal command has been initiated. This includes tasks such as gate sizing and target location prediction. Increasing the frame rate will shorten the photon collection time and hence reduce loop latency.

Compensators are designed with regard to two key parameters: track loop bandwidth and steady-state tracking error. The track loop bandwidth governs how quickly the tracking loop responds to changes in target position. Compensator pole and zero locations and loop latency are adjusted during the design process to achieve the desired loop bandwidth. Steady-state tracking error is independent of loop latency but depends on open-loop gain. The larger the

open-loop gain, the smaller the tracking error. If zero steady-state tracking error is required, an open-loop integrator or type 1 feedback system should be considered. Another common design configuration comprises a type 2 loop with two open-loop integrations. A type 2 tracking loop tracks a target moving with constant LOS velocity, with no steady-state error. A trade-off exists between the loop type and the loop bandwidth.

5.5 TRACK LOOP PERFORMANCE PREDICTION

5.5.1 Performance Criteria for Tracking Systems

A tracking system usually constitutes only one component of a larger entity such as a weapon system. Hence, the criteria used to measure tracking performance must reflect the larger system. For example, in fire control applications, tracking accuracy is a critical measure of performance. There are two parameters that serve to quantify tracking performance in a large number of applications. These are tracking accuracy and loss-of-lock probability. Other performance criteria include the rate of convergence to steady-state tracking, robustness against countermeasures, and the ability to deal with multiple targets.

Tracking accuracy can be quantified by instantaneous track error $\mathbf{e}_t(k)$. The instantaneous track error can be represented by the algebraic sum of three error constituents:

$$\mathbf{e}_t(k) = \mathbf{e}_b(k) + \mathbf{e}_d(k) + \mathbf{e}_j(k) \ , \tag{5.36}$$

where

$\mathbf{e}_b(k)$ = systematic bias term
$\mathbf{e}_d(k)$ = drift error representing errors that grow with time
$\mathbf{e}_j(k)$ = random pointing (jitter) error resulting from noise; white noise is commonly assumed for analysis.

The bias term $\mathbf{e}_b(k)$ can arise from mechanical, servo, and/or gyro errors. Offsets in the target location estimation algorithm(s) can also lead to bias error. Drift error may result from adaptation in the target location algorithms and/or from gyro drift. The jitter error $\mathbf{e}_j(k)$ arises from noise in the sensor data, roundoff errors in the processor,[31] and sensor platform disturbances.

Loss of lock, quantified by the loss-of-lock probability P_L, can be caused by extreme target/platform motions and/or by imagery in which the target's location cannot be reliably estimated. Loss of lock caused by target motion usually results from a poorly performing LOS determination function and/or the gimbal control system. The specific cause can be found by the hybrid analysis method (Sec. 5.5.3), which employs characterizations of target location estimator performance. (These characterizations were described in Sec. 5.2.) The hybrid approach enables target motion parameters and servo parameters (e.g., acceleration and processor latency) to be related to P_L.

Analytically, it is difficult to determine the conditions for loss of lock resulting from an unreliable estimate of the target's location. The primary means of determining P_L for this case is simulation of the image processing algo-

rithm(s) using imagery containing the particular scene factor to be investigated. Scene factors that may result in loss of lock include:

- significant target signature changes, such as a gun blast
- target obscuration by ground clutter
- multiple targets in a crossing maneuver
- thermal merging or loss of contrast between parts of the target and the background
- dead or erratically behaving detectors or sensor readout electronics.

5.5.2 Analytical Pointing Accuracy Prediction

Performance predictions based on analytical models are often quite useful for design trade-off studies. Because many tracking algorithm combinations are possible, it is helpful to carry out trade-off studies at an early stage in the design process to narrow the design trade space. Analytical performance prediction is also useful after a design has been completed to aid in the evaluation of system test results and as a validation for simulation experiments.

Analytical prediction is difficult for any system as complex as a typical tracking system. Nevertheless, methods are available for obtaining answers to certain performance-related questions. Small-signal analysis is one such method. For this technique, it is assumed that all signal variations are small enough that nonlinear input-output processes can be modeled, to sufficient accuracy, by linearizing about the nominal operating point. As a consequence, the well-known procedures from linear systems theory can be applied.

Small-signal analysis is particularly effective for predicting the pointing accuracy of a tracking system. Pointing accuracy is normally specified for steady-state tracking conditions, i.e., after the tracking system has locked onto the target and initial transients have subsided. In the steady-state mode, LOS corrections are ordinarily small enough to justify the small-signal assumption.

Figure 5.21 is a block diagram of a linearized system model useful in pointing accuracy analysis. Note that the target location estimation component has

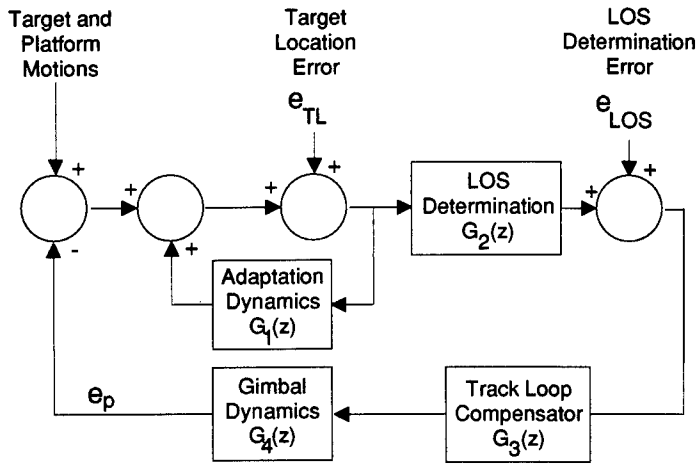


Fig. 5.21 Block diagram of the linearized small-signal model used for pointing prediction.

been represented by a dynamic linear component with additive noise. The dynamic model is used to account for adaptation associated with the target location estimation algorithm. Adaptation is modeled by the discrete time transfer function $G_1(z)$.

The LOS determination component is modeled by its linearized transfer function $G_2(z)$, and the track loop compensator and the gimbal dynamics are modeled by their discrete time or sampled data transfer functions $G_3(z)$ and $G_4(z)$, respectively. If the LOS determination algorithms contain geometric transformations, these are replaced by their derivatives evaluated at the nominal operating point. If the LOS determination component employs an extended Kalman filter, it is replaced by its linear steady-state form. The LOS error signal represents the effect of ancillary measurement errors on the LOS estimate. The LOS error signal can also contain random platform disturbances, which can be modeled by a random noise source and periodic signals.

The output signal y_p is the tracking system pointing error. Analytic methods from linear feedback analysis can be used to determine the relationship between e_p and the target location (TL) and LOS errors.[13]

$$e_p = \left(\frac{G_2G_3G_4}{1 - G_1 + G_2G_3G_4}\right)e_{\text{TL}} + \left[\frac{G_3G_4(1 - G_1)}{1 - G_1 + G_2G_3G_4}\right]e_{\text{LOS}} . \tag{5.37}$$

Tracking systems are sometimes characterized by their steady-state or asymptotic tracking error, which is typically calculated with the target location error and LOS error signals removed. The steady-state error is a measure of the tracking loop's ability to point the gimbals. For a stationary target, this steady state error is

$$e_{\text{ss}} = \lim_{n\to\infty} e(nT) = \frac{1}{1 + \lim_{z\to 1}(G_2G_3G_4)} , \tag{5.38}$$

where e_{ss} is the steady-state track error to a step change assuming that $\lim_{z\to 1} G_1 = 0$, which is true if the adaptation converges.

A special case of the steady-state tracking error is defined for a target whose angular velocity is constant with respect to the sensor platform. Assuming again that the adaptation converges, this error is given by

$$e_{\text{ss}} = \lim_{n\to\infty} e(nT) = \frac{T_f}{\lim_{z\to 1}(z - 1)[G_2G_3G_4/(1 - G_1)]} , \tag{5.39}$$

where e_{ss} is the steady-state track error to a target with constant velocity and T_f is the frame time (period) of the imaging sensor.

The small-signal assumption was invoked to permit use of the linear model on which the preceding results are based. If it is necessary to characterize the performance of the tracking system for large variations in target position, the described analytic approach usually breaks down. Performance predictions for large variations generally require a simulation study.

5.5.3 Tracking System Performance Prediction through Simulation

Because AVTSs are usually implemented as nonlinear feedback systems, it is difficult to predict tracking performance analytically, except in special cases where small-signal analysis is valid. In most cases of practical interest, performance predictions are obtained from Monte Carlo simulations that incorporate target and background data or models, the applicable tracking and LOS determination algorithms, and a nonlinear model for the gimbal control system and sensor platform. Platform disturbances can be modeled as colored noise with discrete harmonic components to represent periodic disturbances from sources such as a cryogenic pump. For the most part, Monte Carlo simulation experiments do not provide the insight needed to optimize the tracking system design, but they can be quite useful for evaluating its performance limits.

Simulation experiments are particularly effective in predicting the point at which tracking will break down. As long as the target's angular velocity remains constant or changes slowly, the tracking system will be able to follow it and maintain track. But if the target undergoes a sudden, large acceleration (e.g., a tactical aircraft switching on its afterburners), it may escape from the tracker's search field, resulting in loss of lock. It is often important to determine the point, falling between these limiting cases, at which the breaklock occurs, especially for systems tracking maneuvering targets.

Tracking systems that must follow maneuvering or accelerating targets typically exhibit large, nonlinear signals that cannot be ignored in the performance analysis. This is particularly true of the gimbal component, which often exhibits saturation effects.

A complete Monte Carlo simulation for a tracking system is computationally very intensive, and therefore costly. The primary reason for this stems from the need to simulate the nonlinear control components and the image processing components simultaneously. The extent to which simulation of a tracking system design can realistically be carried out is often sharply limited by the costs involved. Computational expense can be reduced somewhat by splitting the simulation into two parts comprising (1) image processing and (2) the closed-loop control system. Although closed-loop effects influence image processing performance, they can usually be modeled separately and accounted for without resorting to a complete closed-loop simulation. For example, a high angular rate between the target and the sensor LOS can smear the imagery produced by a staring FPA sensor, leading to diminished tracking performance. By applying the appropriate amount of smear to the input imagery, this closed-loop effect can be accounted for in the image processing part of the simulation.

The closed-loop control system simulation models the effect of the image processing component on the target location estimate reported to the LOS determination component, without including the processing algorithm itself. This model is coupled with an accurate simulation of the tracking loop's other components. Figure 5.22 shows the structure of this simulation. Here, the image processing part has been represented as the target LOS with respect to the sensor's boresight, with an additive random error. The analytical error characterization of Sec. 5.2 can be used for the error signal.

In Fig. 5.22, platform vibrations are modeled as colored noise with harmonics to represent periodic components, such as a cryopump. The target motion

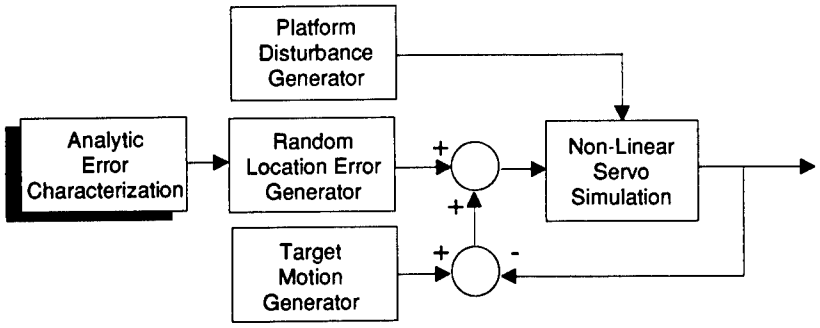


Fig. 5.22 Hybrid approach to tracking performance prediction.

generator models the target dynamics relative to the sensor platform. Target dynamics typically studied in this type of simulation include

- Step angular position change that models, for example, the shock caused by gunfire. This tests the system's transient response.
- Angular velocity with respect to the platform. This tests how well the track loop will follow nonaccelerating targets. The angular rate at which the track loop breaks down is an important performance specification for fire control systems.
- The Singer model for maneuvering targets.[14,39]

Acknowledgments

The author would like to thank both J. Garvey and M. Busch for the generous support they provided in generation of this chapter. Their contribution to readability and correctness is greatly appreciated.

References

1. D. W. Fisher, R. F. Leftwich, and H. W. Yates, "Survey of infrared trackers," *Applied Optics,* **5**(4), 507–515 (1966).
2. H. L. Van Trees, *Detection, Estimation, and Modulation Theory, Part I,* pp. 515–574, John Wiley & Sons, New York (1968).
3. M. D. Srinath and P. K. Rajasekaran, *An Introduction to Statistical Signal Processing with Applications,* pp. 156–201, 224–248, John Wiley & Sons, New York (1979).
4. P. S. Maybeck, *Stochastic Models, Estimation, and Control,* Vols. I, II, and III, Academic Press, New York (1979).
5. J. M. Mendel, *Maximum-Likelihood Deconvolution: A Journey into Model-Based Signal Processing,* Springer-Verlag, Berlin (1990).
6. R. E. Nasburg, S. M. Stillman, and N. M. Nguyen, "An integrated processor architecture for multisensor signal processing," *Proceedings of the SPIE* **1481,** 84–95 (1991).
7. J. M. Fitts, "Aided tracking as applied to high accuracy pointing systems," *IEEE Transactions on Aerospace and Electronic Systems* **AES-9**(3), 350–368 (1973).
8. E. Bakir, "Adaptive Kalman filter for maneuvering targets," *AIAA Journal of Guidance, Control, and Dynamics* **6**(6), 414–416 (1983).
9. H. Kwkernaak and R. Sivan, *Linear Optimal Control Systems,* Wiley-Interscience, New York (1972).
10. M. Aoki, *Optimization of Stochastic Systems,* pp. 20–70, 154–196, Academic Press, New York (1967).

11. C. K. Chui and G. Chen, *Kalman Filtering,* pp. 23–27, Springer-Verlag, Berlin (1987).
12. J. M. Mendel, *Lessons in Digital Estimation Theory,* pp. 81–99, 258–269, Prentice-Hall, Englewood Cliffs, NJ (1987).
13. S. S. Gupta and L. Hasdoff, *Fundamentals of Automatic Control,* pp. 217–221, John Wiley & Sons, New York (1970).
14. S. S. Blackman, *Multiple-Target Tracking with Radar Applications,* pp. 75–81, Artech House, Norwood, MA (1986).
15. J. S. Borland, L. J. Pinson, E. G. Peters, G. R. Kane, and W. W. Malcolm, "Design of a correlator for real-time video comparisons," *IEEE Transactions on Aerospace and Electronic Systems* **AES-15**(1), 11–20 (1979).
16. A. L. Gilbert, M. K. Giles, G. M. Flachs, R. B. Rodgers, and Y. Hsun U, "A real-time video tracking system," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAM-2**(1), 47–56 (1980).
17. D. R. Van Rheeden and R. A. Jones, "Noise effects on centroid tracker aim point estimation," *IEEE Transactions on Aerospace and Electronic Systems* **24**(2), 177–185 (1988).
18. Y. Bar-Shalom, H. M. Shertukde, and K. R. Pattipati, "Use of measurements from an imaging sensor for precision target tracking," *IEEE Transactions on Aerospace and Electronic Systems* **25**(6), 863–872 (1989).
19. R. Voles, "Image registration using the likelihood ratio," *Proceedings of the IEEE* **125**(7), 686–690 (1978).
20. W. Eversole and R. E. Nasburg, "Maximum likelihood estimation for image registration," *Proceedings of the SPIE* **432,** 190–194, San Diego (Sept. 1983).
21. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes,* pp. 274–334, Cambridge University Press, Cambridge (1986).
22. W. K. Pratt, *Digital Image Processing,* Wiley-Interscience, New York (1991).
23. H. Mostafavi and F. W. Smith, "Image correlation with geometric distortion, part I: acquisition performance," *IEEE Transactions on Aerospace and Electronic Systems* **AES-14**(3), 487–493 (1978).
24. H. Mostafavi and F. W. Smith,"Image correlation with geometric distortion, part II: effects on local accuracy," *IEEE Transactions on Aerospace and Electronic Systems* **AES-14**(3), 494–500 (1978).
25. D. M. Tobin and P. S. Maybeck, "Enhancement to a multiple model adaptive estimator/image-tracker, *IEEE Transactions on Aerospace and Electronic Systems* **24**(4), 417–426 (1988).
26. R. L. Frey, C. D. Nealy, L. M. Rubin, and R. M. Wilcox, "ATAC autocuer modeling analysis," Electro-Optical and Data Systems Division, Hughes Aircraft Company, Report No. FR-80-70-1325, El Segundo, CA (Jan. 1981).
27. W. K. Pratt, "Correlation techniques of image registration," *IEEE Transactions on Aerospace and Electronic Systems* **AES-10**(3), 353–358 (1974).
28. C. D. McGillem and M. Svedlow, "Image registration error variance as a measure of overlay quality," *IEEE Transactions on Geoscience Electronics* **GE-14**(1), 44–49 (1976).
29. D. F. Andrews, P. J. Bickel, F. R. Hampel, W. H. Rogers, and J. W. Tukey, *Robust Estimates of Location,* Princeton University Press, Princeton, NJ (1972).
30. D. I. Barnea and H. F. Silverman, "A class of algorithms for fast digital image registration," *IEEE Transactions on Computers* **C-21**(2), 179–186 (1972).
31. F. K. Bowers and R. J. Klingler, "Quantization noise of correlation spectrometers," *Astronomy and Astrophysics Supplement,* **15,** 373–380 (1974).
32. J. M. Fitts, "Correlation tracking via optimal weighting functions," Electro-Optical and Data Systems Division, Hughes Aircraft Company, Report No. P73-240, El Segundo, CA (Apr. 1973).
33. T. J. Willett and P. K. Raimondi, "Intelligent tracking techniques—a progress report," *Proceedings of the SPIE* **178,** 72–76 (1979).
34. D. P. Peterson and D. Middleton, "Sampling and reconstruction of wave-number-limited functions in N-dimensional Euclidean space," *Information and Control* **5,** 279–323 (Dec. 1962).
35. T. P. Wallace and O. R. Mitchell, "Analysis of three-dimensional movement using Fourier descripters," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAM-2**(6), 583–588 (1980).
36. W. A. Perkins, "Area segmenters of images using edge points," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAM-2**(1), 8–15 (1980).

37. R. M. Haralick and L. G. Shapiro, *Computer and Robot Vision,* pp. 13–91, 509–548, Addison-Wesley, Reading, MA (1992).

38. R. E. Nasburg, S. M. Stillman, and N. M. Nguyen, "A multiple Kalman filter approach for multiple sensor multiple target tracking," IRIS Specialty Group on Active Systems (Oct. 1990). (Unclassified paper available from authors).

39. R. A. Singer and K. W. Behnke, "Real-time tracking filter evaluation and selection for tactical applications," *IEEE Transactions on Aerospace and Electronic Systems* **AES-7**(1), 100–110 (1971).

CHAPTER 6

Signature Prediction and Modeling

John A. Conant
Malcolm A. LeCompte
Aerodyne Research, Inc.
Billerica, Massachusetts

CONTENTS

6.1 INTRODUCTION

Signature prediction is an important part of infrared systems analysis. Realistic assessments of IR system performance depend heavily on the use of signature models. These models should include and faithfully reproduce the effects of the most important radiative signature processes. Because IR systems typically utilize contrast between a target and its background, and because natural sources of target infrared radiation are frequently important, environmental as well as target radiative phenomena must be included in predictive analysis tools. The combined effects of environmental variability, target surface and thermal properties, and interactions between structural components or environmental features are sufficiently subtle to warrant the application of complex numerical models. The models themselves serve to enhance phenomenological understanding and provide a very cost effective means to examine a wide range of scenarios. However, this is effective only if the model has been validated by extensive comparison with measured data.

Although ultimately there are no substitutes for careful analysis, complex numerical analysis techniques alone may be unnecessarily time consuming, may fail to provide insight into the operation of fundamental processes, or may simply be unavailable to the analyst. In this event, simple analytic techniques can be applied that may provide a lower level of fidelity and accuracy, but may be adequate for a specific purpose.

Simple techniques may be appropriate for those occasions when it is useful to estimate the infrared intensities of objects (military targets, in particular) without resorting to complex and time-consuming computer modeling techniques. Estimation techniques should be derived from first principle approximations and should be judged by their ability to replicate results from data or from more complex analytic models. The analytic techniques presented here are all designed to be used given only the most cursory and fundamental information about the object whose signature is to be determined, such information as might be available in unclassified and generally available source publications.

The objective of this chapter is to provide a description of the most important phenomena affecting infrared radiation, especially from vehicles that are potential military targets. In addition, simplified equations are provided that allow infrared estimates to be made.

The complexities of infrared computations are such that many large computer models have been constructed for this purpose. We shall not discuss such models. Simplified calculations may serve several purposes, including

- to provide zeroth- or first-order infrared estimates
- to provide checks on the output of complex models
- to educate the analyst on the basic considerations and steps used by more complex models.

The last purpose should not be overlooked. The analyst will operate and utilize a complex model most accurately if he understands the phenomena and procedures on which it is based.

To achieve reasonably accurate estimates of object *signatures* (observable infrared radiation) it is most important to understand the complex phenomenology that determines the levels of radiation. The phenomenological pro-

cesses are particularly dependent on the wavelength of the radiation to be considered. The signature of an object is also dependent on the aspect presented to an observer.

The signature phenomenology of many objects changes markedly as the aspect viewed changes. For these reasons, it is not possible to specify a simple, single, rule of thumb that yields an accurate radiant intensity for all wavelengths or aspects. For example, tail views of aircraft can be dominated by either the plume or the engine hot parts, depending on whether the bandpass is restricted to spectral regions containing strong plume absorption bands. For views of the nose or beam, an aircraft's radiation may be governed by either the emissive or reflective characteristics of its coating, depending on the bandpass. In the mid-wave infrared (MWIR) band, typically from 3 μm (3300 cm^{-1}) to 5 μm (2000 cm^{-1}), exhaust plumes, solar radiation, and thermal emission all play significant roles in determining the total target signature. In the long-wave infrared (LWIR), usually 8 μm (1250 cm^{-1}) to 12.5 μm (800 cm^{-1}), body thermal emission and earthshine reflections usually play the dominant signature roles. Small changes in aspect significantly change a target's projected area as well as the dominant sources of radiation.

As a starting point, one should bear in mind that wavelengths shorter than 3.5 μm are typically dominated by hot sources of radiation—the sun, engine hot parts, and hot exhausts. Wavelengths longer than 5 μm are typically dominated by cooler surfaces, plus hot parts or exhausts. The troposphere is generally opaque in spectral regions around 2.7, 4.3, and 15 μm.

This chapter cannot cover all classes of military targets, and is limited to aircraft (fixed and rotary wing), ground vehicles, and ships. We refer to these as *objects* or *targets*. There is also no attempt to duplicate the detailed treatment of a state-of-the-art numerical infrared model. As stated above, the intent is to supplement such models with a more simple approach.

We can make an initial estimate of the infrared signature of a target of known projected area by assuming that it is a blackbody radiator at some equilibrium temperature, because all objects radiate in the infrared. In this case, the spectral blackbody radiance, given by the Planck function, is integrated over wavelength and multiplied by the target area projected onto a plane orthogonal to the observer's line of sight. This is illustrated in Example 1.

Example 1. An aircraft is shown in Fig. 6.1, superimposed on a geometric grid that can be used to estimate the target's projected area. The target is assumed to be a blackbody, radiating uniformly in all directions. For a typical tactical aircraft, the beam view projected area is about 30 m^2. If the target is at a uniform 300 K, the Planck blackbody radiance function, expressed in terms of wavelength,

$$B_\lambda(T) = \frac{2hc^2}{\lambda^5[\exp(hc/\lambda kT) - 1]} ,$$

yields a radiance of 0.003 W/m^2 sr cm^{-1} at 2220 cm^{-1} (4.5 μm) and 0.1 W/m^2 sr cm^{-1} at 1000 cm^{-1} (10 μm) wavelength. A typical mid-wavelength infrared sensor might have a bandpass from 4 to 5 μm. A long-wavelength sensor might utilize the spectral region from 8 to 12 μm. Integrating over each bandpass

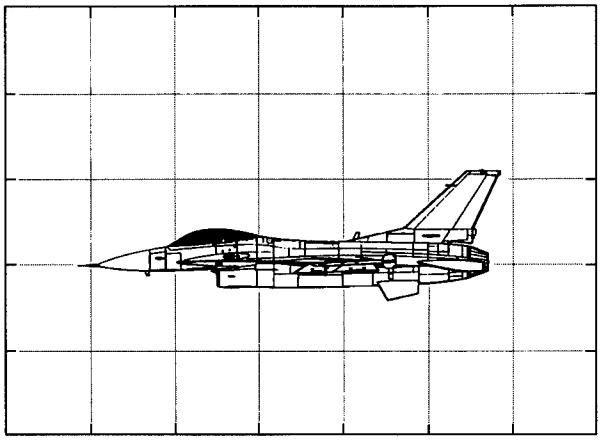


Fig. 6.1 Beam view of tactical fighter superimposed on grid to allow estimate of its area projected in the direction of an observer.

gives in-band radiances of about 1.5 and 42 W/m^2 sr, for the mid- and long-wave sensors. The approximate target source radiant intensity, from the beam perspective, is then about 45 and 1250 W/sr, respectively. These radiant intensity values might be representative of a target coated with a very low reflectance paint, seen at short range, and flying at moderate airspeed at low altitude or moving faster at higher altitude.□

This sample estimate was based on unrealistic assumptions about the optical properties and thermal conditions of the target surface. It also neglected the potentially significant target radiation resulting from exhausts, reflected earth, skyshine, and sunshine and the transmissivity of the atmosphere between the target and observer. It ignored aspect-dependent signature components, such as variation in target cross-sectional area, and the contribution resulting from engine exhaust ducts and plumes. It ignored the background radiation levels, which determine the contrast under which the target is seen. Each of these factors, in particular circumstances, can significantly impact the infrared signature of an object. In the sections that follow, we present techniques to estimate each of these effects.

6.2 OBSERVABLE INTENSITY RELATIONS

The end product of an infrared model, suitable for input to a sensor study, is usually an in-band irradiance as incident on the sensor foreoptics. For some studies, the *contrast* (target minus background) irradiance is useful. Because these irradiances contain a strong factor of R^{-2}, where R is a range to the object, the object's radiant intensity is often computed instead.

These radiometric quantities are several steps removed from the most basic model computation, which is of a spectral radiance from a single point on the object. The equations in this section illustrate the connection between them.

Once we have computed the spectral radiance $L_\lambda(\mathbf{x})$ of each visible point $\mathbf{x}$ on the object (the subject of the rest of this chapter), we can proceed to integrate

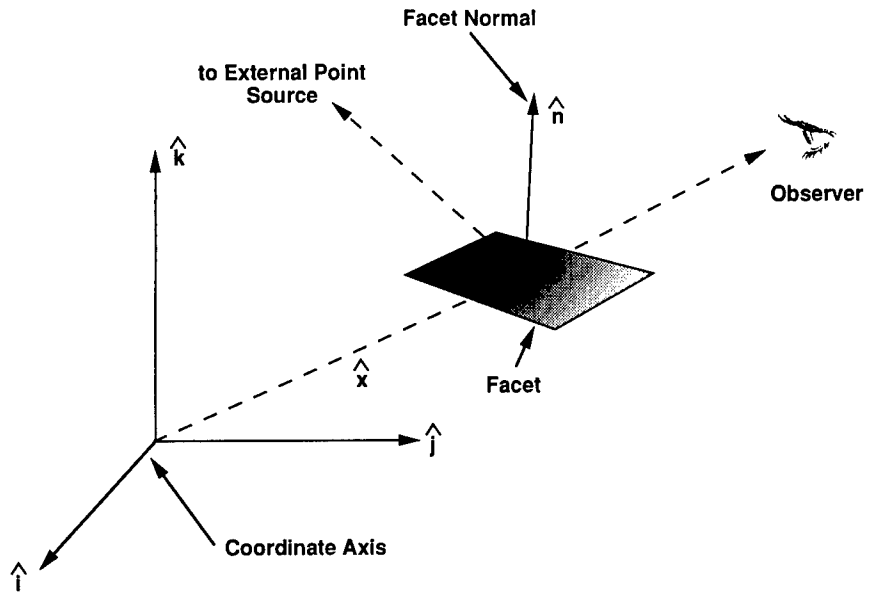


Fig. 6.2 Vector convention for object and observer description.

over the object's visible projected area (see Fig. 6.2). Each differential area dA is multiplied by a projection factor of $\hat{n}\cdot\hat{o}$, where $\hat{n}$ and $\hat{o}$ are unit vectors representing the local surface normal and the direction to the observer (sensor). In addition, the area should be multiplied by an obscuration factor of 0 or 1, depending on whether the point is, in fact, visible to the sensor. If the surface point is approximated with a locally flat facet, it is desirous to generalize the obscuration factor to a fraction visible, between 0 and 1. We denote a small, visible projected area as $dA(\mathbf{x})^*$ for the surface position $\mathbf{x}$.

Given the preceding, we can express the object's source spectral radiant intensity as

$$I_{\text{src}}(\lambda) = \int_{\text{object}} L(\mathbf{x},\lambda)\, dA(\mathbf{x})^* \ . \tag{6.1}$$

The term *source* (abbreviated as src) is used to denote the neglect of any transfer effects (absorption, scattering, emission) between the object and sensor. A second useful quantity is the object's total projected area:

$$A^*_{\text{object}} = \int_{\text{object}} dA(\mathbf{x})^* \ . \tag{6.2}$$

To convert from source to apparent quantities, we include foreground transmittance (τ_{fgd}) and radiance (L_{fgd}) by

$$I_{\text{app}}(\lambda) = I_{\text{src}}(\lambda)\tau_{\text{fgd}}(\lambda) + L_{\text{fgd}}(\lambda)A^*_{\text{object}} \ . \tag{6.3}$$

Note in Eq. (6.3) that we have assumed the foreground effects to be approximately constant across the object, as viewed at the sensor. A contrast radiant intensity, against a locally uniform background, is then computed as

$$I_{con}(\lambda) = I_{app}(\lambda) - L_{bkg}(\lambda)A^*_{object} \ , \tag{6.4}$$

where L_{bkg} is the background radiance, including atmospheric effects all the way up to the sensor. Finally, the preceding quantities can all be integrated spectrally, with an optional spectral response weighting function r_{sensor} for the sensor:

$$I_{con}(\Delta\lambda) = \int_{\Delta\lambda} I_{con}(\lambda) r_{sensor}(\lambda)\, d\lambda \ . \tag{6.5}$$

The radiance of any point can have contributions from emission, transmission, and reflection (scattering). For our present purposes we distinguish two primary types, namely, (1) opaque surfaces ("body") and (2) exhaust gases ("plume"). Opaque surfaces are considered to emit thermally, and to reflect their environment; any transmission is neglected. Plumes are considered to emit thermally, and to transmit light; scattering of light is neglected.

Body radiance is a sum of different components:

$$L_{tot} = L_{th} + L_{ss} + L_{es} + L_{sky} + L_{inter} + L_{totplume} \ , \tag{6.6}$$

where L_{tot} is a point's total spectral radiance [the same as $L(\mathbf{x},\lambda)$], L_{th} is the thermal emission, L_{ss} is the reflected sunshine, L_{es} is either reflected earthshine or cloudshine, L_{sky} is reflected skyshine or cloudshine, L_{inter} is the inter-reflections of the target, and $L_{totplume}$ is the total radiation, emitted and transmitted, emanating from the plume.

In the case of a ground vehicle or ship, the interactions term may include reflections of the object off of its environment as well as the reverse. For these cases, such reflections can be an important part of the object's total signature. A scene may also contain significant signature elements resulting from the recent passage of an object such as track marks or a ship's wake.

The spectral radiance from a plume has only two terms, an emission term and a transmittance term (neglecting the scattering term, which can be important, as for visible contrails):

$$L_{totplume} = L_{plumeemis} + \tau_{plume} L_{behind\ plume} \ , \tag{6.7}$$

where $L_{plumeemis}$ is the plume's own emission, τ_{plume} the plume's transmittance, and $L_{behind\ plume}$ the radiance of whatever is behind the plume from the observer's point of view (the target, background, etc.).

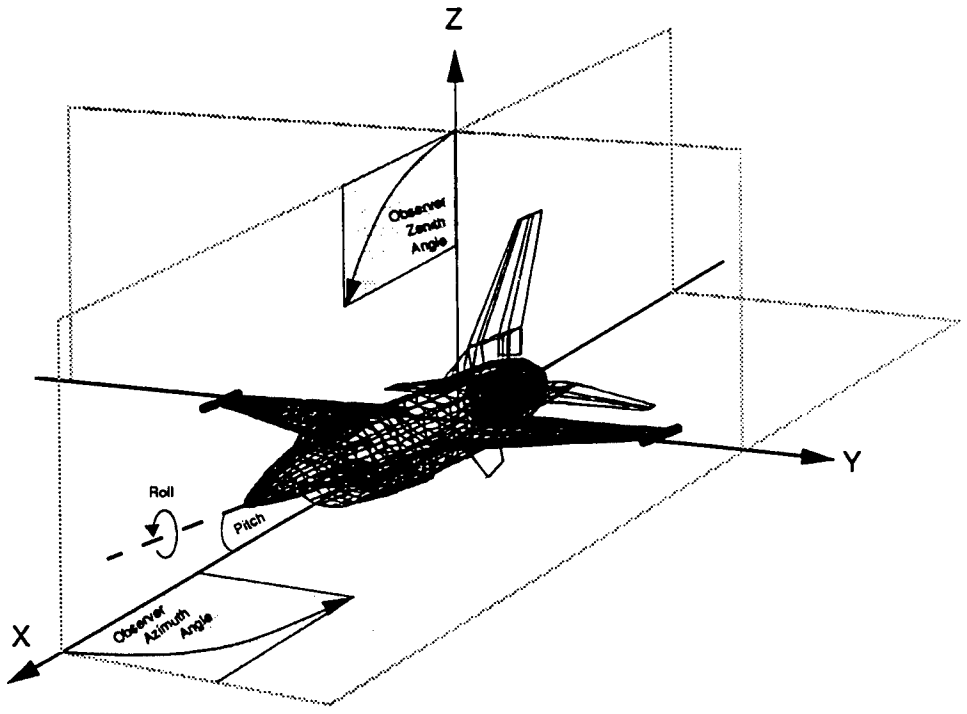
Each of these component radiation sources is discussed in turn. Moreover, techniques are introduced to estimate the contribution of thermal emission, environmental sources of reflected radiation, and plume radiation.

6.3 GEOMETRY DESCRIPTION

To perform infrared signature calculations, the relative positions and orientations of the target, the observer, and the environment must be known. The geometry of the target itself must also be known, both for the orientations of its various surfaces (relative to the emission sources and the observer), and for obscurations (hidden surfaces). These are discussed in the following.

6.3.1 Scene Coordinates

Assume a coordinate system called the basic coordinate system (BCS) centered at the target (see Fig. 6.3). It is a right-handed Cartesian system with the z axis pointed up along the local vertical. The x and y axes are in the local horizontal plane. Usually, the x axis points in the direction of vehicle heading (as if there is no pitch or angle of attack). The sun and observer are defined in spherical coordinates from the origin. Assume for simplicity, therefore, that the object is seen at long range and all surface-to-observer directionals are



Observer (reader) Zenith angle = 80°

Observer (reader) Azimuth angle = 25°

Fig. 6.3 Diagram of the basic coordinate system, showing the axes, angle of attack (AOA), and zenith (θ) and azimuth (ϕ) angles to either the sun or observer.

parallel. For a typical target, an observer range of 350 m is sufficient to keep the maximum angular error within 5 deg.

6.3.2 Target Coordinates

Another useful coordinate system is the target coordinate system (TCS). The TCS is a right-handed Cartesian system that is fixed to the target body, so that it rolls and/or pitches with the body. The recommended convention has the z axis "vertical" to the target, the x axis pointing forward, and the y axis pointing out the left, or port, side. These axes describe a right-handed system with azimuth angles measured positively counterclockwise from the nose; readers should not confuse this system with the common pilot's usage of azimuths measured clockwise, implying a left-handed system. Because the TCS is by definition fixed to the target, the target surface coordinates are always input in this system. An example of a faceted target description is also shown superimposed in the coordinate system in Fig. 6.3; this is known as a *wire-frame* drawing. Such geometric data bases are often generated from computer-aided design (CAD) systems or loft-line drawings, which may use a different coordinate system. One very common system has the x axis pointing out the rear of the vehicle, so that the x values increase from nose to tail. In this system, the x values denote target *stations,* the y values denote *butt lines* or *body lines,* and the z values denote *water lines,* in typical design parlance. If such a system is used, then observer and sun azimuth angles would have to be measured from the tail of the vehicle.

A target may be approximately described by a set of flat facets, which are often triangles or quadrilaterals. We use this technique in our examples as a computational aid when integrating across an object. Examples of simple models of a ship and a tank, whose polygonal surfaces can be constructed of triangular or quadrilateral facets, are shown in Figs. 6.4(a) and (b).

The visible area of each facet projected toward the observer can be obtained, as can the visible fraction of a facet's area. The first result is the term A^*, referred to in Sec. 6.2, and is needed to spatially integrate from radiances to radiant intensities. The sunlit area of each facet may be multiplied by the solar irradiance to obtain the amount of sunlight falling on that facet.

The first step in making an accurate estimate of the infrared signature of a target is to create a reasonably faithful rendering of the target shape. Ex-

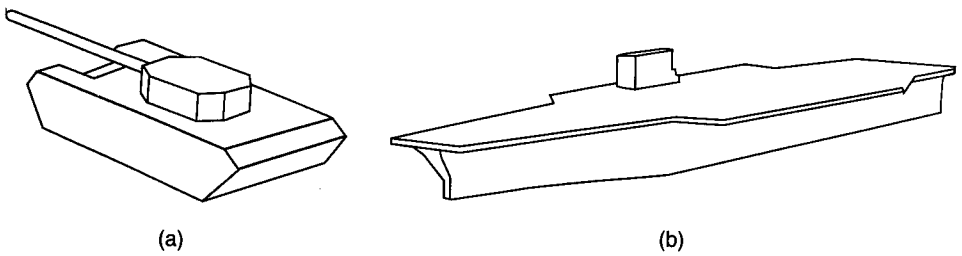


Fig. 6.4 Simple facetized rendering of (a) a tank and (b) aircraft carrier.

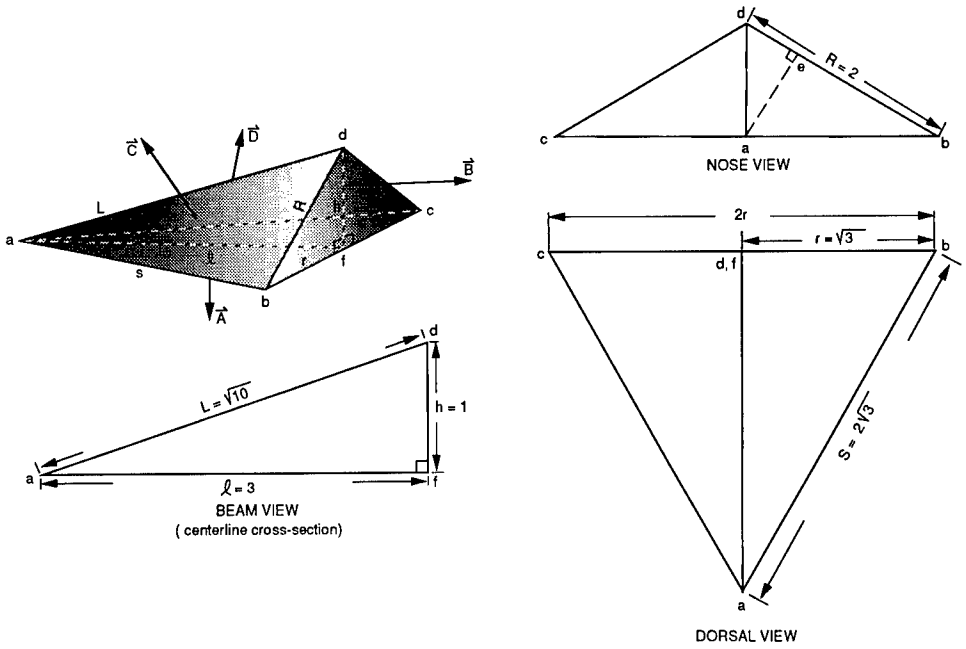


Fig. 6.5 Plan view and shaded 3-D drawing of sample tetrahedronally shaped object: flying wedge.

ample 2 provides the geometric basis for the target signature calculations in forthcoming sections.

Example 2. Figure 6.5 shows plan views of a simple geometric shape whose form is suggestive of a modern aircraft. The shape is that of a tetrahedron whose height is less than either length or width, and whose length is slightly greater than its "span." The dimensions and relative scale of the object are similar to the more complex tactical aircraft it is meant to represent. The value of this simple shape is that each of its surfaces can be treated separately and in detail to calculate its infrared signature. In this way, the techniques for calculating each signature component can be clearly illustrated. Although four triangular surfaces are far too few to adequately describe actual targets, many such shapes can be combined to create complex geometries. Also, the complexity of detail is easily incremented to incorporate progressively more complex signature phenomenology. We have christened this shape as the F-118 *flying wedge,* and remaining calculations are based on viewing its component surfaces. □

In the foregoing, a target-centered, spherical coordinate system was introduced, which specified the direction to observers and environmental sources of radiation like the sun, earth, moon, or sky. A rectilinear coordinate system is used to define vectors that characterize the orientation of each body surface. Azimuth and zenith of observer and radiation sources can be converted to

vectors in this coordinate system to accomplish the radiation calculations. The radiation apparent to an observer viewing any surface depends on the orientation of the surface normal with respect to the observer's line of sight.

The simple tetrahedron target has one face lying in the X-Y plane, with its normal in the negative z direction (downward). Another face lies in the Y-Z plane with its normal directed in the negative x direction (rearward). The figure is bilaterally symmetrical, with equal volume lying on either side of the X-Z plane. The tetrahedron's dimensions are determined by its height (z dimension), the length of the diagonal edge R in the Y-Z plane, and the tetrahedron's length along the x axis, which are in the ratio 1:2:3, respectively.

The projected area of a target surface to an observer is equal to the product of the surface area and the cosine of the angle formed by the surface normal and the vector to the observer. In the three-dimensional rectilinear coordinate system, this is equivalent to the product of the surface area with the dot product of the two vectors.

For the relative dimensions previously given, the area of each surface is determined by simple geometry, where the height of the object in dimensionless units is 1.

Ventral Surface Area *A*. The area of surface A is equal to the area of the triangle *abc:*

$$\text{area } \triangle abc = A_A = \frac{1}{2}(2r)l = rl \quad (= 3\sqrt{3}) \ .$$

Rear Surface Area *B*. The area of surface B is equal to the area of the triangle *bcd.*

$$\text{area } \triangle bcd = A_B = \frac{1}{2}(2r)h = rh \quad (= \sqrt{3}) \ .$$

Port and Starboard Dorsal Surface Areas *C* and *D*. The area of surfaces C (triangle *abd* with area A_C) and D (triangle *acd* with area A_D) can be found using the Pythagorean theorem and the shared altitude of the two right triangles, *abe* and *ade,* that are interior to the triangular surface *abd:*

$$A_C, A_D = \frac{1}{2}(Rt) \ ,$$

where t is the length of the shared altitude of right triangles *abe* and *ade*:

$$t = [(rh)^2 + (lh)^2 + (rl)^2]^{1/2}/(r^2 + h^2)^{1/2} \ ,$$

$$A_C, A_D = [(rh)^2 + (lh)^2 + (rl)^2]^{1/2}/2$$

$$\left(= \frac{\sqrt{3}}{2}\sqrt{13}\right) \ .$$

The normal unit vectors of each of the four surfaces are listed or derived below:

- Ventral surface normal $\hat{A}$: $\hat{A} = -\hat{k}$
- Rear surface normal $\hat{B}$: $\hat{B} = -\hat{i}$
- Port and starboard dorsal surface normals $\hat{C}$ and $\hat{D}$.

The unit normal vector for surface C ($\hat{C}$) is given by the cross product of vectors $\mathbf{q}$ (along edge S) and $\mathbf{r}$ (along edge L) divided by its absolute value (see, for example, Ref. 1 for an explanation of vector operations):

$$\hat{C} = (\mathbf{q} \times \mathbf{r})/|\mathbf{q} \times \mathbf{r}| ,$$

where

$$\mathbf{q} = -(l\hat{i} - r\hat{j}) \quad \text{and} \quad \mathbf{r} = -(l\hat{i} - h\hat{k}) ,$$

$$\hat{C} = (rh\hat{i} + lh\hat{j} + rl\hat{k})/[(lh)^2 + (rl)^2 + (rh)^2]^{1/2} ,$$

$$\hat{C} = (rh\hat{i} + lh\hat{j} + rl\hat{k})/2A_C ,$$

$$\hat{C} = (\hat{i} + \sqrt{3}\hat{j} + 3\hat{k})/\sqrt{13} .$$

Similarly, the unit vector normal to surface D ($\hat{D}$) is found to be

$$\hat{D} = (rh\hat{i} - lh\hat{j} + rl\hat{k})/[(lh)^2 + (rl)^2 + (rh)^2]^{1/2} ,$$

$$\hat{D} = (rh\hat{i} - lh\hat{j} + rl\hat{k})/2A_D ,$$

$$\hat{D} = (\hat{i} - \sqrt{3}\hat{j} + 3\hat{k})/\sqrt{13} .$$

Projected Area for Observer 1. The observer location is defined in the BCS, by specifying the zenith (θ) and azimuth (ϕ) angles and range from the object. Zenith and azimuth angles can be used to define a unit vector pointing to a point source or an observer, in a Cartesian system as shown:

$$\hat{O} = \cos\phi \sin\theta\hat{i} + \sin\phi \sin\theta\hat{j} + \cos\theta\hat{k} .$$

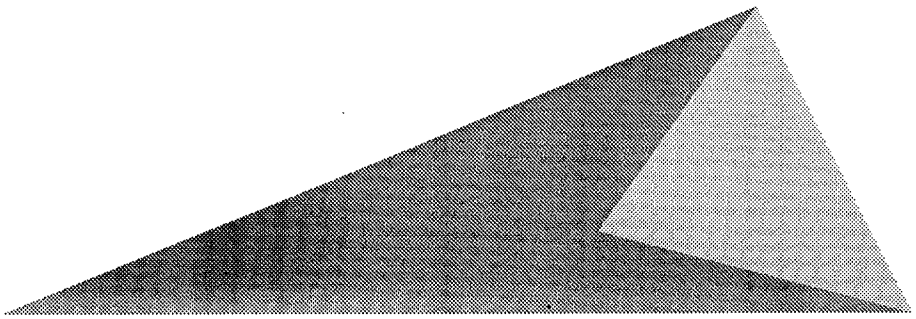
Consider an observer, abeam and slightly aft of the target, viewing it from just below, as in Fig. 6.6(a). The projected areas of surfaces A, B, and C are found by multiplying the dot product of the vector to the observer ($\hat{O}$) and the surface unit normals, $\hat{A}$, $\hat{B}$, and $\hat{C}$, with the three respective surface areas, as shown in the following. When the dot product goes negative the observer is *behind* the surface, and its projected area is taken as zero.

The projected area of the ventral surface A is given by

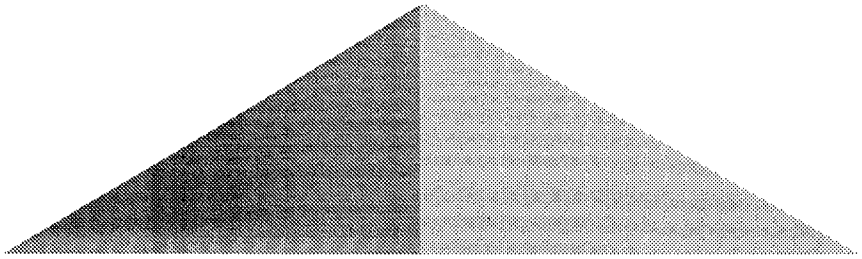
$$A_{PA} = A_A(\hat{O} \cdot \hat{A}) = rl[\hat{O} \cdot (-\hat{k})]$$

$$= -rl\cos\theta \quad (= -3\sqrt{3}\cos\theta) .$$

The projected area of the aft surface (B) is



(a)



(b)

Fig. 6.6 (a) Flying wedge viewed from slightly below and aft of a beam and (b) nose view of flying wedge.

$$A_{PB} = A_B(O \cdot B) = rh[O \cdot (-i)] = -rh \sin\theta \cos\phi$$
$$(= -\sqrt{3} \sin\theta \cos\phi) \ .$$

The projected area of the port dorsal surface (C) is obtained from

$$A_{PC} = A_C(\hat{O} \cdot \hat{C}) = A_C(rh \sin\theta \cos\phi + lh \sin\theta \sin\phi + rl \cos\theta)/(2A_C)$$
$$= \frac{\sqrt{3}}{2}(\sin\theta \cos\phi + \sqrt{3} \sin\theta \sin\phi + 3 \cos\theta) \ .$$

We now select the observer angles,

$$\theta = 96 \text{ deg} \quad \text{and} \quad \phi = 110 \text{ deg} \ .$$

The projected area of each face (in nondimensional units) is

ventral surface: $A_{PA} = 0.54$,

aft surface: $A_{PB} = 0.59$,

dorsal surface: $A_{PC} = 0.84$.

The object's total (dimensionless) area is about 2 square units. For a tetrahedron 5 m in height each square unit corresponds to 25 m^2 of area. The total projected area is therefore about 50 m^2. If the object is at a temperature of 300 K, then, assuming that it radiates as a blackbody, its radiant intensities at mid- and long-IR wavelengths (as in Example 1) are 75 and 2100 W/sr.

Projected Area for Observer 2. An observer with a direct view of the wedge's nose, as shown in Fig. 6.6(b), sees projected areas of surfaces C (A_{PC}) and D (A_{PD}). Expressions for the projected area in the observer's direction are found, as in the previous example, from the surface unit normals ($\hat{C}$, $\hat{D}$), areas (A_C, A_D), and the observer vector $\hat{O}$. In this case, $\theta = 90$ deg and $\phi = 0$ deg, so, $\hat{O} = \hat{\imath}$, and therefore;

$$A_{PC} = A_C(\hat{O} \cdot \hat{C}) = A_C(\hat{\imath} \cdot \hat{C}) = A_C\left(\frac{rh}{2A_C}\right) = \frac{rh}{2} .$$

We thus have

$A_{PC} = \sqrt{3}/2$,

$A_{PD} = \sqrt{3}/2$,

total area $= A_{PC} + A_{PD} = \sqrt{3}$.

For the wedge in the previous example the projected area is 43.5 m^2. The mid- and long-wavelength radiant intensities are 65 and 1830 W/sr.

The results for both observers are summarized in Tables 6.1 and 6.2, which list the projected area of each visible surface for each viewer and the surface blackbody radiant intensities at our exemplary MWIR and LWIR bandpasses.

The surface normal vectors, and the projected area values, derived for the nose and beam observers are used in radiation calculations in subsequent sections.

Table 6.1 Projected Area by Viewer in Nondimensional and Dimensionalized Units

	A_{PA}	A_{PB}	A_{PC}	A_{PD}	Total
Aft ventral beam view	0.54 / 13.5 m^2	0.59 / 14.8 m^2	0.84 / 21.0 m^2	— / —	1.97 / 49.3 m^2
Nose view	— / —	— / —	0.87 / 21.8 m^2	0.87 / 21.8 m^2	1.74 / 43.6 m^2

Table 6.2 Surface Blackbody Radiant Intensity (in watts per steradian) for MWIR and LWIR Bandpass

Radiant Intensity (W/sr)	L_A	L_B	L_C	L_D	Total
MWIR Bandpass					
Aft ventral beam view	20.3	22.2	31.5	—	74.0
Nose view	—	—	32.7	32.7	65.4
LWIR Bandpass					
Aft ventral beam view	567	622	882	—	2071
Nose view	—	—	916	916	1832

6.4 SIMPLIFIED RADIATION ENVIRONMENT MODEL

Because reflections of an object's environment can be important, it is necessary to have a model of that environment. As with the object itself, the environment requires a geometric model and a radiance model. The basic radiance formulas to use are the same as for the object surfaces, as described in Secs. 6.5 and 6.7. The geometric model must depend on the scenario.

The earth, sky, and clouds each provide both thermal emission and (in the daytime) solar reflections. Estimates of atmospheric emission and scattering are quite challenging. The analyst is urged to consult elsewhere in this handbook, and also to acquire the readily available computer model[2] MODTRAN.

Airborne objects can usually be considered to have uniform hemispheres or infinite planes above and below them, representing earth, sky, or clouds. Spatial variations along the earth, or across fields of broken clouds, are a second-order effect for reflection; they should be accounted for "on average," when choosing values of temperature and reflectance/emittance for the hemispheres. The sun appears as a disk subtending about one-half of a degree of angle. It is nearly always sufficient to represent it as a point source, although the disk model is just about as easy to use.

Objects that are at, or near, ground level often have a full 3-D environment, with elements such as walls, trees, or hills. The analyst must decide how much detail to include, and what geometric description to use. Both the element's radiance and its angle, subtended from the object, should be considered. Note that angle is more important than size—a mountain in the far distance is a small irradiance source compared with a close-range boulder. Also, if the radiance of an environmental element is close to that of surrounding elements, they should likely be combined.

There are many instances when the object under study reflects an appreciable amount off of its environment; this usually occurs when they are physically close. Examples of this effect are ship reflections in the water, or car reflections off of a road. In such cases, the object and its environment cannot be clearly separated in the calculation. A good approach to this effect is to include a local portion of the environment as part of the object calculations.

A number of environmental parameters must be collected before its radiances can be computed, for subsequent reflection off of the object. These parameters include temperatures, emittances, diffuse and specular reflectances,

the solar irradiance, and the atmospheric transmittance/radiance between environment and object. Material on most of these may be found elsewhere in this handbook.

Temperatures, of course, are dynamic, and can be hard to estimate. Healthy foliage tends to stay within a few degrees (up to 10) of air temperature because of evaporation.[3] Branches and soil can easily heat up 10°C or more above the air in sunny conditions. Bodies of water are thermally complex, but large ones exhibit surface temperatures that are nearly independent of the diurnal cycle because of their size and consequent thermal inertia. The date, locale, and recent weather history of the water can often provide an analyst with clues for a common-sense estimate.

Clouds are also complex. An estimate of their temperature is the ambient air temperature at the cloud altitude (taken from data, or from a standard model atmosphere of the sort mentioned elsewhere in this manual). This is more accurate for stratiform clouds than for cumuliform ones, which exhibit large amounts of vertical convection.

Given the radiance of an environmental element reaching the object, we need to know that element's total irradiance on the object's surfaces. For a point source (the sun) one needs merely to project the source's intrinsic irradiance (normal to its direction) onto the object surface, with its unit normal vector $\hat{N}$:

$$E_{\text{surface}} = E_{\text{sol}}(\hat{N} \cdot \hat{S}) \ ,$$

where $\hat{S}$ is a unit vector toward the sun. For a Lambertian infinite-plane source (earth, sky) with unit vector normal to the plane $\hat{N}_{\text{src}}$ ($+\hat{k}$ for the earth, $-\hat{k}$ for the sky) we find that

$$E_{\text{surface}} = \frac{\pi}{2} L_{\text{source}}[1 - (\hat{N}_{\text{src}} \cdot \hat{N})] \ .$$

These values would be reduced by transmission losses resulting from atmospheric scattering and absorption occurring between the source and surface. Clearly, the shorter the column length or the lower the density of intervening atmosphere, the smaller the transmission losses. Atmospheric emission tends to offset the losses. Calculations using the unattenuated irradiance usually yield upper limits to surface illumination levels.

Example 3. Solar Irradiance on Object Surfaces. The magnitude of the solar irradiance on the surface of the flying wedge can be obtained for a specific time of day and target orientation. It is given by

$$E_{\text{surface}} = E_{\text{sol}}(\hat{N} \cdot \hat{S}) \ ,$$

where $\hat{N}$ is the unit surface normal vector and $\hat{S}$ is the unit vector toward the sun. As before, negative dot products indicate the sun is behind the surface. The position of the sun selected for this sample calculation anticipates reflected radiation calculations to be performed in succeeding sections and $\hat{S}$ is taken to be

$$S_x = [(rh)^2 - 2A_C^2]/2A_C^2 \qquad (= -11/13) ,$$

$$S_y = lrh^2/2A_C^2 \qquad (= 2\sqrt{3}/13) ,$$

$$S_z = lhr^2/2A_C^2 \qquad (= 6/13) ,$$

$$\hat{S} = -(11\hat{i} - 2\sqrt{3}\hat{j} - 6\hat{k})/13 .$$

This choice, corresponding to zenith and elevation angles of 62.5 and 162.5 deg, respectively, anticipates an example in Sec. 6.7 that demonstrates the relative importance of solar illumination to an object's IR signature, while maintaining calculational simplicity. The sun is positioned to provide a sun glint toward an observer viewing the object nose. Because the solar irradiance varies with the orientation of the illuminated surface, the magnitude and importance of solar radiation reflected from a surface will also vary.

The target's coordinate system is taken to be superimposed on the inertial frame, its nose at the origin, and no pitch, roll, or yaw. In this case, only the rear and dorsal surfaces are illuminated by the sun. The unit normal vectors for these surfaces are

$$\hat{B} = -\hat{i} ,$$

$$\hat{C} = (rh\hat{i} + lh\hat{j} + rl\hat{k})/2A_C$$

$$= (\hat{i} + \sqrt{3}\hat{j} + 3\hat{k})/\sqrt{13} ,$$

$$\hat{D} = (rh\hat{i} - lh\hat{j} + rl\hat{k})/2A_C$$

$$= (\hat{i} - \sqrt{3}\hat{j} + 3\hat{k})/\sqrt{13} .$$

Taking the dot product of source direction and surface normal:

$$\hat{S} \cdot \hat{B} = -[(rh)^2 - 2A_C^2]/2A_C^2 \quad (= 11/13) ,$$

$$\hat{S} \cdot \hat{C} = rh/2A_C \qquad (= 1/\sqrt{13}) ,$$

$$\hat{S} \cdot \hat{D} = [A_C^2 - (lh)^2]rh/2A_C^3 \quad (= 1/13\sqrt{13}) .$$

For the flying wedge, the solar irradiance on each surface is then

$$E_B = E_{\text{sol}}(11/13) ,$$

$$E_C = E_{\text{sol}}(1/\sqrt{13}) ,$$

$$E_D = E_{\text{sol}}(1/13\sqrt{13}) .$$

We can approximate the radiance of the sun by taking it to be roughly equivalent to a blackbody at 5800 K. In this case, integrating over the mid-wavelength infrared band of 4 to 5 μm, E_{sol} is about 6 W/m^2. Integrating from 8 to 12 μm, E_{sol} is estimated to be about 1 W/m^2. □

Example 4. Earth and Sky Irradiance on Object Surfaces. The earth and sky (and over- or underlying cloud layers) are extended sources of radiation that may either form a background or be reflected off of object surfaces. If these extended sources are simulated by infinite planes at constant temperatures over or under the object, the source directions are always oriented in either the positive or negative z direction. The expression for irradiance of an object surface resulting from an infinite plane source is

$$E_{\text{surf}} = \pi L_{\text{source}}[1 - (\hat{N}_{\text{src}} \cdot \hat{N}_{\text{surf}})]/2 \ .$$

This expression indicates that sky radiation illuminates the dorsal and rear surfaces of our sample object. Earthshine provides illumination to its ventral and rear surfaces.

For the flying wedge, dorsal surface irradiance resulting from skyshine and earthshine is found by knowledge of the source directions,

$$\hat{N}_{\text{src}} = -\hat{k} \quad \text{for skyshine and} \quad \hat{N}_{\text{src}} = \hat{k} \quad \text{for earthshine} \ ,$$

and surface normals,

$$\hat{N}_{\text{surf}} = \hat{C} \quad \text{or} \quad \hat{D} \ .$$

The irradiance on surfaces C and D resulting from sky and earthshine is then

$$E_{C,D} = \pi L_{\text{sky}}[1 - (-\hat{k} \cdot \hat{C}, \hat{D})]/2$$

$$+ \pi L_{\text{earth}}[1 - (\hat{k} \cdot \hat{C}, \hat{D})]/2 \ ,$$

$$E_{C,D} = \pi L_{\text{sky}}(1 + rl/2A_C)/2$$

$$+ \pi L_{\text{earth}}(1 - rl/2A_C)/2 \ ,$$

$$E_{C,D} = \pi L_{\text{sky}}(1 + 3/\sqrt{13})/2$$

$$+ \pi L_{\text{earth}}(1 - 3/\sqrt{13})/2 \ .$$

Similarly, ventral surface $[\hat{N}_{\text{surf}}(= \hat{A} = -\hat{k})]$ irradiance resulting from earthshine is given by

$$E_A = \pi L_{\text{earth}}[1 - (\hat{k} \cdot -\hat{k})]/2 = \pi L_{\text{earth}} \ .$$

Irradiance on the aft surface of the flying wedge results from contributions from both earth and skyshine and is given by

$$\hat{N}_{\text{surf}} = \hat{B} = -\hat{i} \ ,$$

$$E_B = \pi\{L_{\text{earth}}[1 - (\hat{k} \cdot -\hat{i})] + L_{\text{sky}}[1 - (-\hat{k} \cdot -\hat{i})]\}/2 \ ,$$

$$E_B = \pi(L_{\text{earth}} + L_{\text{sky}})/2 \ .$$

Earthshine can be first approximated by assuming that the earth radiates as a blackbody with an emissivity appropriate to the type of terrain below the object. Assuming a temperature of 305 K and an emissivity of 0.9, earthshine radiance is about 2 W/sr m^2 at MWIR wavelengths and about 40 W/sr m^2 at LWIR wavelengths.

The radiance from a clear sky is more difficult to estimate because one must solve an equation of radiative transfer through a nonhomogeneous scattering medium. From the earth's surface, the clear sky appears cold when the optical path is near the local zenith, and warmer when viewing the horizon. If an object is located beneath an optically thick cloud deck, however, illumination of the upper surfaces can be treated in a manner similar to the earthshine. For example, a low hanging cloud at 285 K would produce a sky radiance of about 1 and 30 W/sr m^2 in the MWIR and LWIR, respectively. Clear sky radiances would be smaller. Table 6.3 shows the unattenuated irradiance on each surface of the flying wedge from solar radiation and by earth and skyshine.

These values are reduced significantly by atmospheric attenuation. With the sun at the 62.5-deg zenith angle, and an object at 1-km altitude, the solar irradiance is diminished by about 75% in the MWIR and by about 50% at LWIR wavelengths. Even at low altitudes, atmospheric attenuation of earthshine can be significant. For an object at 1 km, the average attenuation, estimated by considering a pathlength of 45-deg nadir is about one-half at MWIR wavelengths and about one-third in the LWIR. If the cloud deck lies 1 km above the object, atmospheric attenuation reduces its irradiance by about half at MWIR wavelengths and by about one-fifth in the LWIR.

Table 6.4 shows the values of the surface irradiance for the flying wedge accounting for atmospheric attenuation of the incident environmental radiation.

Table 6.3 Estimates for the Unattenuated MWIR and LWIR Irradiance of Each Surface of the Flying Wedge Resulting from Illumination by Solar Radiation and Earth- and Skyshine

MWIR Irradiance			
Surface Irradiance (W/m^2)	Solar E_{sol} = 6 W/m^2	Earthshine L_{earth} = 2 W/m^2 sr	Skyshine L_{sky} = 1 W/m^2 sr
E_A	0	6	0
E_B	5	3	2
E_C	2	0.5	3
E_D	0.1	0.5	3
LWIR Irradiance			
Surface Irradiance (W/m^2)	Solar E_{sol} = 1 W/m^2	Earthshine L_{earth} = 40 W/m^2 sr	Skyshine L_{sky} = 30 W/m^2 sr
E_A	0	125	0
E_B	0.8	65	45
E_C	0.3	10	85
E_D	0.02	10	85

Table 6.4 Estimates for the Attenuated MWIR and LWIR Irradiance of Each Surface of the Flying Wedge Resulting from Illumination by Solar Radiation and Earth- and Skyshine, with Atmospheric Attenuation

MWIR Irradiance			
Surface Irradiance (W/m^2)	Solar* E_{sol} = 1.5 W/m^2	Earthshine L_{earth} = 1.0 W/m^2 sr	Skyshine L_{sky} = 0.5 W/m^2 sr
E_A	0	3.00	0
E_B	1.30	1.50	1.00
E_C	0.50	0.25	1.50
E_D	0.03	0.25	1.50
LWIR Irradiance			
Surface Irradiance (W/m^2)	Solar* E_{sol} = 0.5 W/m^2	Earthshine L_{earth} = 27.0 W/m^2 sr	Skyshine L_{sky} = 24.0 W/m^2 sr
E_A	0	84	0
E_B	0.4	44	36
E_C	0.15	7	68
E_D	0.01	7	68

*Sun at zenith angle of 62.5 deg and azimuth angle at 162.5 deg. □

The attenuated values are used in calculations performed in subsequent sections. With the radiation environment characterized, the next step is to present analytic tools to define the interaction of an object surface with its environment.

6.5 SIMPLIFIED BRDF/EMISSIVITY MODEL

The bidirectional reflectance distribution function (BRDF) is an excellent tool for reflectance model development, both because of its generality and because of its definition as the ratio of two convenient radiometric quantities. The BRDF is defined as

$$f(\theta_i,\phi_i;\theta_r,\phi_r;\lambda) = \frac{dL_r(\theta_r,\phi_r;\lambda)}{dE_i(\theta_i,\phi_i;\lambda)} \quad . \tag{6.8}$$

Thus, the BRDF is the ratio of reflected radiance (dL_r) to the incident irradiance (dE_i). The primary purpose of this section is to provide an analytic BRDF model that is simple enough for our purposes and that contains the most important reflection phenomena. It is important to note the need for a concurrent emissivity model that conserves energy with the BRDF and that obeys Kirchhoff's law. We neglect surface transmittance in the current model.

Spectral reflectance variations are of major practical importance, but usually occur over relatively broad wavelengths compared with other infrared phenomena, especially atmospheric transmission. Because the modeling approach used in this section can easily be made spectral, we consider the reflectance to be constant over local spectral regions and drop all spectral subscripts for convenience.

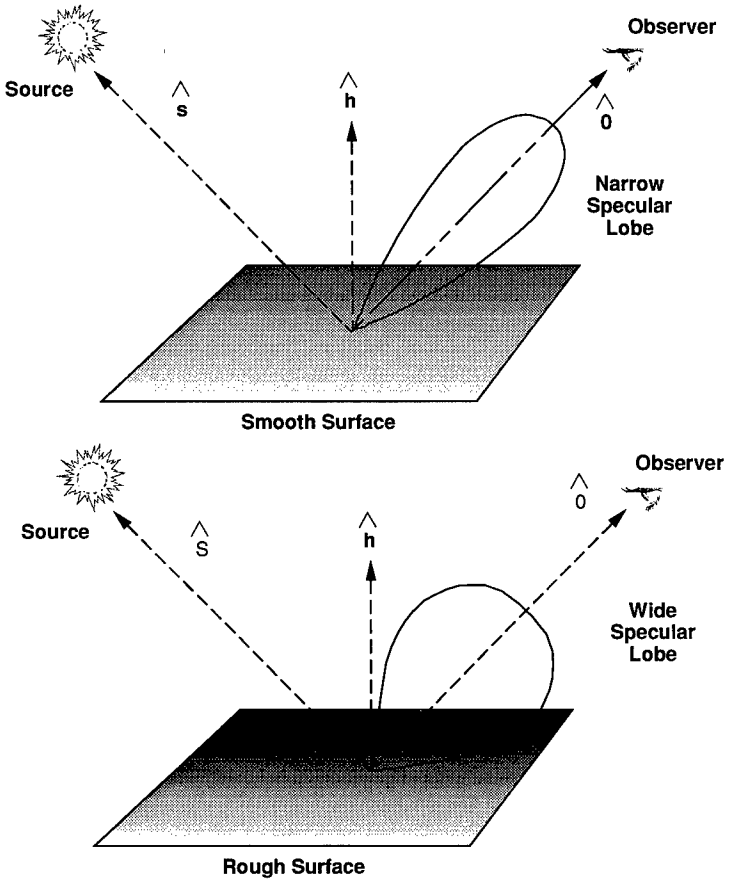


Fig. 6.7 Effect of surface roughness on specular lobe widths.

Many authors have pointed out the phenomenon of *quasispecular* reflection, in which light incident from a single direction is reflected over a spread of directions near the classical Snell's law—incidence angle equals reflected angle—direction.[4] These specular reflections typically form a *lobe* of finite width, with peak reflectance in the Snell's law direction. Real surface materials exhibit lobes of greatly different widths, depending on surface roughness (see Fig. 6.7). Lobe widths may range from a fraction of a degree to tens of degrees. For the present model, we idealize to lobes that have constant BRDF within an angular region and are 0 elsewhere, to give a simple representation that captures the essence of specularity. We denote the lobe's full-width angle as e.

The quasidiffuse component of reflection is sometimes used to represent surface reflections from microfacets whose normals are far from the bulk surface normal,[5] and sometimes to represent volume reflections of light that multiply scatters inside the surface.[6] Whatever the physical model, the analytical model is certainly a useful one. Models that explore diffuse phenomena provide useful angular reflectance functions, but the variations with angle are usually mild. We thus take a common course herein, and approximate the diffuse

component with a constant Lambertian BRDF. Other reflectance characteristics, which we do not consider, include off-specular peaks, polarization, and grazing-angle shadowing.

The ideal limit of a pure Lambertian BRDF was described by Nicodemus et al.[7] We merely combine this limit with an idealized specular lobe to allow a variable-specular surface. The extension to spectral variations is trivial, if required. Our BRDF model is, thus,

$$f(\theta_i,\phi_i;\theta_r,\phi_r;\lambda) = \frac{1}{\pi}\rho_d(\lambda) + \rho_s(\lambda)g(\hat{N}_{\text{src}},\hat{N}_{\text{surf}},\hat{O}) \ , \tag{6.9}$$

where the function g has a value of $1/[\pi \sin^2(e/2)]$ whenever $\hat{O}$ is within an angle of $e/2$ of the ideal glint direction and g is 0 elsewhere. Note the approximation that the glint lobe is always entirely above the surface plane.

The analysis is not complete until the emittance has also been modeled. If the material's transmittance is taken as 0, then conservation of energy requires that all incident light be either reflected or absorbed. The total reflected light is computed by the integral of the BRDF over all reflection solid angles. The light absorbed from an incident direction, from Kirchhoff's law, must be equal to the emittance in that direction. We can thus write

$$\int\int f(\theta_i,\phi_i;\theta_r,\phi_r;\lambda)\cos(\theta_r)\sin(\theta_r)\,d\theta_r\,d\phi_r + \varepsilon(\theta_i,\phi_i;\lambda) = 1 \ . \tag{6.10}$$

Substituting our simplified BRDF model, Eq. (6.9), we easily obtain

$$\rho_d(\lambda) + \rho_s(\lambda) + \varepsilon(\theta_i,\phi_i;\lambda) = 1 \ . \tag{6.11}$$

The reflectance model of Eq. (6.9) thus requires that the emittance be constant (Lambertian) with angle, or

$$\varepsilon(\lambda) = 1 - \rho_d(\lambda) - \rho_s(\lambda) \ .$$

This simple expression significantly mitigates the complexity in estimating radiation signatures. It retains diffuse, specular, and emittance terms; however, the magnitude of any two can be varied and energy will be conserved. For the example calculations we have used, $\varepsilon = 0.85$, $\rho_d = 0.05$, and $\rho_s = 0.10$.

6.6 BODY TEMPERATURES

In spectral regions beyond 4 μm having fairly high transmittance, the emission from the object surface is usually the largest contributor to the target source signature. These emissions are typically nongraybody radiation emitted from a non-Lambertian surface at a temperature T.

Thermal calculations for an object should account for heat transfer via conduction, convection, and radiation as well as evaporation and condensation. This topic deserves a volume to itself; for present purposes we discuss approximations for different classes of objects.

Fixed-wing aircraft can be the simplest of vehicles to model thermally to a first approximation. They are of necessity built with a thin skin, which equi-

librates quickly. The plane's motion causes convection to dominate other heat transfer mechanisms, and most of the outer airframe is close to the effective temperature of the air's boundary layer, the recovery temperature T_r:

$$T_r = T_{\text{amb}}\left(1 + \frac{\gamma - 1}{2} r M^2\right) , \tag{6.12}$$

where T_{amb} is the ambient air temperature (on an absolute scale, such as kelvin), γ is the air specific heat ratio (close to 1.4 unless greatly heated), r is a recovery factor, and M is the vehicle Mach number. The recovery factor accounts for the fact that the effective boundary layer temperature is less than the air *total temperature* (also called *stagnation temperature*) T_t. Values of T_t are achieved only when all of the kinetic energy of the air's speed is converted into thermal energy within the boundary layer. When some of that energy dissipates to the airstream, the boundary layer achieves only the cooler value of T_r. Experience shows that, at moderate subsonic Mach numbers, a recovery factor of 0.85 is a good estimate; at slower speeds, the recovery factor should approach unity. Places where the air impinges perpendicular to the airframe (*stagnation points*) will have a boundary layer at T_t. For an aircraft in subsonic cruise, T_r is often about 20°C above T_{amb}. Stagnation points are generally small, and are only slightly hotter, and so can be neglected to a first approximation.

Radiative cooling is partially offset by radiative absorption of earth and sky emissions, until it becomes large when the aircraft travels supersonically. When a turbine engine afterburner is on, however, the exhaust emissions typically dwarf those of the airframe, except for near-nose viewing. Radiative terms are discussed later in the section.

Turbine engine nozzles are, of course, important emitters. Although nozzles are thermally complex, a reasonable upper bound temperature (which is often close to reality) can be gained from the recovery temperature of the exhaust gases in contact with it. For turbofan engines, one must consider whether it is the hot core flow (from the turbine exhaust) or the warm fan flow that contacts the nozzle walls. If the turbine exit is directly visible, the exhaust total temperature is a good estimate for it.

It is extremely difficult to estimate exhaust gas temperatures with any accuracy without a complex model of the engine thermal cycle. Such models are routinely constructed by engine manufacturers, and are termed *cycle decks*. The exhaust temperatures of any given engine vary strongly with many factors, including ambient air conditions; flight speed; aircraft pitch, roll, and trim; airframe configuration and weight; and (of course) the pilot's power lever setting. For systems analysis purposes, one should consider turbine exhaust gas temperatures of 450 to 600 K for cruise flight, 600 to 800 K for maximum unaugmented thrust, and 1000 to 1500 K for augmented (afterburning) thrust. Fan exhaust temperatures may range from about 350 to 450 K.

Rotary-wing aircraft have the added complexities of rotor downwash and the ability to hover, complicating convection estimates. A reasonable first estimate is to compute a recovery temperature based on a vector addition of forward and downwash speeds.

Missiles can pose a difficult thermal problem because of their high accelerations and Mach numbers. In most cases, however, their enormous exhaust

plume emission dominates the body signature (even in spectral bands such as 8 to 12 μm). However, the analyst must sometimes consider scenarios with a missile postburn phase, where the plume has stopped, or with trajectories where low air pressure has cooled the plume sufficiently to stop any combustion in the external plume, giving it a weak signature. The latter phenomenon depends on many parameters and is very hard to predict.

When missile body emissions are important, they are transient in nature and quite dependent on missile materials and trajectory. During the boost phase, the missile accelerates continuously, leading to continuous increase in the recovery temperature. The skin temperature T will lag behind T_r because of finite thermal response time and radiative heat loss. An upper bound temperature is certainly given by T_r. For another level of accuracy, one must assume the skin to have infinitely small response time, and to balance the convective and radiative terms:

$$h(T_r - T) - \varepsilon\sigma T^4 + \varepsilon_{\text{env}}\sigma T_{\text{env}}^4 = 0 \ , \tag{6.13}$$

where T is the desired skin temperature, T_r the (instantaneous) recovery temperature, h the convective heat transfer coefficient, ε the body thermal emissivity, and σ is the Stefan-Boltzmann constant. The factor T_{env} represents an average environmental temperature (earth, sky, clouds); its value is not critical. Similarly, ε_{env} is the average environmental emissivity. For simplicity we take a simple estimate of h for a flat plate in a turbulent flow:

$$h = C_p \rho U S \ , \tag{6.14}$$

where C_p is the ambient air heat capacity at constant pressure, ρ is the ambient air density, U the flight speed, and S the Stanton number.[8] A value for S of 1×10^{-3} is sufficient for present purposes. The Stanton number for a flat plate in a turbulent flow depends on density, velocity, and distance from a leading edge, but only to the one-fifth power. A change in flow parameters by a factor of 32 is thus needed to halve or double S.

As an example, consider the typical values of $h = 100$ W/m^2 K^{-1}, $T_{\text{env}} = 280$ K, and $\varepsilon = 0.9$. With $\sigma = 5.75 \times 10^{-8}$ W/m^2 K^{-4} and $T_{\text{amb}} = 280$ K, we can estimate T versus the flight Mach number, using Eqs. (6.12) and (6.13). Table 6.5 lists T and T_r for these conditions and three different Mach numbers. The example shows how both the recovery temperature T_r and the radiative cooling (the difference between T and T_r) grow sharply with Mach.

Solar heating depends largely on the amount of atmosphere the sunlight passes through (a function of altitude and solar elevation angle), surface ori-

Table 6.5 Estimates of Skin Temperature T and Recovery Temperature T_r Versus Mach Number M for a Set of Example Conditions

M	0.8	1.5	2.5
T_r (degrees kelvin)	310.5	387.1	577.5
T (degrees kelvin)	309.3	379.9	537.8

entation to the sun, and the convective rate. When convection dominates conductive and radiative losses, the temperature rise resulting from the sun can be approximated by

$$\Delta T_{\text{sun}} = \frac{\alpha_{\text{sun}} E_{\text{sol}} (\hat{N}_{\text{sun}} \cdot \hat{N}_{\text{surf}})}{h} .$$

In this equation, α_{sun} represents the surface's solar absorptivity (between 0 and 1); E_{sol} is the solar irradiance at the object, normal to the sun's direction (integrated over all wavelengths); and $\hat{N}_{\text{sun}}$ and $\hat{N}_{\text{surf}}$ are unit vectors toward the sun and normal to the surface. The factor h is the convective heat transfer coefficient, as previously described. Typical values of these factors are $\alpha_{\text{sun}} = 0.7, E_{\text{sol}} = 1000\ \text{W/m}^2$, and $h = 100\ \text{W/m}^2\,\text{K}^{-1}$. The corresponding temperature increase resulting from the sun would then be 7 K for a surface directly facing the sun.

Temperatures of ground vehicles and ships are usually difficult to estimate for a number of reasons. First, the absence of high-speed motion reduces the dominance of convection on the temperature; the convective term may be quite large or quite small. Second, thick armor on such vehicles has a large heat capacity, leading to nonequilibrium conditions. Reduced convection leads to increased solar heating, which is strongly directional; the object will be nonuniformly heated. The combination of low convection and thick materials leads to significant three-dimensional heat transfer effects. Internal engine heating, for example, can spread through a vehicle over long time periods. Many other effects can also be important, such as condensation, evaporation, rain, snow, mud, sea spray, and shadowing by local objects.

These considerations warn us against attempting an accurate thermal estimate without a sophisticated model. To achieve even a first-order estimate, the analyst must consider the

- materials, thicknesses, and orientations of the surfaces in view
- prevailing weather and its recent history
- object's location and motion.

The most simple estimate would be of steady-state conditions, made by a thermal balance of convective, conductive, and radiative terms. However, unless one surface is in strong thermal contact with another one, or with a heat source, a first-order balance can usually neglect conduction. Typically, however, one must consider radiative absorption from the sun, sky, and the local terrain.

For a systems estimate, one can expect a sunlit surface to be heated by 5 to 50°C above the ambient air temperature, in the absence of rain, spray, etc., on a day with at least partial sun. During a clear night, most surfaces can radiate heat more efficiently than the sky does, giving a net cooling effect. Temperatures below that of the air are common. Cloud cover lessens both the amount of solar heating and of nighttime cooling.

6.7 OPAQUE SURFACE REFLECTIONS AND EMISSIONS

The general equation for reflection of light from an incident direction (θ_i, ϕ_i) into direction (θ_0, ϕ_0) is

$$dL_{\text{ref}}(\theta_0,\phi_0) = f(\theta_i,\phi_i;\theta_0,\phi_0)dE_{\text{inc}}(\theta_i,\phi_i) \ , \tag{6.15}$$

where all angles are measured relative to the reflecting surface (i.e., the zenith angles θ are measured from the surface's normal vector). In this equation, dL_{ref} is the differential radiance reflected from the incident direction, dE_{inc} is the differential incident irradiance, and f_r is the surface bidirectional reflectance distribution function (see Sec. 6.5). All three of these terms are spectrally dependent. To get the total radiance reflected into the observer direction one must integrate over all incident directions:

$$L_{\text{ref}}(\theta_0,\phi_0) = \int f(\theta_i,\phi_i;\theta_0,\phi_0)\ dE_{\text{inc}}(\theta_i,\phi_i) \ . \tag{6.16}$$

The irradiance from any source can be expressed in terms of that source's radiance as

$$dE_{\text{inc}} = L_{\text{inc}}\ d\omega\ \cos(\theta_i) \ , \tag{6.17}$$

where L_{inc} is the source radiance toward the reflector and $d\omega$ is the differential solid angle given by

$$d\omega = \sin(\theta)\ d\theta\ d\phi \ . \tag{6.18}$$

We can then write

$$L_{\text{ref}}(\theta_0,\phi_0) = \int f(\theta_i,\phi_i;\theta_0,\phi_0)L_{\text{inc}}(\theta_i,\phi_i)\ \cos(\theta_i)\ \sin(\theta_i)\ d\theta_i\ d\phi_i \ . \tag{6.19}$$

The final form of this equation depends on the geometry of the source; a point source emits from a single direction, a spherical source emits from all directions. The resulting equations are given in the following for the various sources in the environment. The result also depends on the form of the BRDF function. In Sec. 6.7.4, we provide both ideal specular and diffuse terms.

6.7.1 Solar Reflections

The sun is a uniform circular source that subtends an angle of about 0.5 deg from the earth. In most cases, it can accurately be considered as a point source. The extraterrestrial solar irradiance (measured perpendicular to the rays) is denoted by $E_{\text{sol}}(\lambda)$. To get the radiance of the sun we must formally multiply E_{sol} by a Dirac delta function to express that E_{sol} comes from a single direction in space. Finally, we must include a factor of the atmospheric transmittance from the sun to the target, giving the expression

$$dL_{\text{inc}}(\theta_i,\phi_i) = E_{\text{sol}}\tau\delta(\theta_i - \theta_s)\delta(\phi_i - \phi_s)/\sin(\theta_s) \ . \tag{6.20}$$

The factor of $1/\sin(\theta_s)$ goes with the delta function to give the correct integral over $d\omega$. Inserting Eq. (6.20) into Eq. (6.19) gives

$$L_{\text{ref}}(\theta_0,\phi_0) = E_{\text{sol}}\tau\cos(\theta_s)f(\theta_s,\phi_s;\theta_0,\phi_0) \ . \tag{6.21}$$

6.7.2 Earthshine Reflections

The earth is an extended source that can be modeled as an infinite flat plane for endoatmospheric targets. The radiance from the earth consists mainly of thermal emission and solar reflection, plus atmospheric transmittance and radiance. Although it is obviously nonuniform, a good earthshine approximation is to assume a uniform temperature and reflectance appropriate to the region under the target. We further approximate the reflectance and emittance as being Lambertian, with spectrally varying values. One can then write

$$L_{\text{earth}} = [L^o(T_e)\varepsilon_e + E_{\text{sol}}\tau_s \cos(\theta_s)\rho_e/\pi]\tau + L_{\text{atm}} , \tag{6.22}$$

where T_e, ε_e, and ρ_e are the earth's temperature, emissivity, and reflectivity, respectively. In this case, τ_s is the transmittance from the sun to the earth. The atmospheric transmittance and radiance terms τ and L_{atm} can be simplified by computations that use a single representative path between the earth and the target.

6.7.3 Skyshine Reflections

The sky above a target can, to a first approximation, be considered an infinite uniform source, as was the earth. This approximation will hold for either clear sky or a solid cloud deck above the aircraft. Broken clouds would require a more sophisticated approach. For clear sky, the only source is atmospheric radiance:

$$L_{\text{sky}} = L_{\text{atm}} . \tag{6.23}$$

A single representative zenith angle can be used for the atmospheric path. For a solid cloud deck, the source function is

$$L_{\text{cloud above}} = L^o(T_c)\varepsilon_c\tau + L_{\text{atm}} . \tag{6.24}$$

We again assume a Lambertian emitter for the cloud. For simplicity the cloud deck is assumed to be thick to eliminate solar radiation transmitted through the cloud.

6.7.4 Simplified Equations

The BRDF model of Sec. 6.5 can now be combined with the general reflectance/emittance equations and the simplified environmental radiance model of Sec. 6.4 to provide the equations we use in our model. The integral in Eq. (6.19) can be distributed over the two terms of the BRDF model, each of which can be integrated easily in closed form. The resulting equation is

$$\begin{aligned} L_{\text{surf}}(\theta_r,\phi_r;\lambda) = {} & \varepsilon(\lambda)L^o(T_{\text{surf}};\lambda) + \rho_d(\lambda)\left[L_{\text{earth}}(\lambda)\sin^2\left(\frac{\theta_n}{2}\right)\right. \\ & \left. + L_{\text{sky}}(\lambda)\cos^2\left(\frac{\theta_n}{2}\right)\right] + \frac{\rho_d(\lambda)}{\pi}(\hat{N}\cdot\hat{S})E_{\text{sol}}(\lambda) \\ & + \rho_s(\lambda)L_{\text{glint}}(\lambda) , \end{aligned} \tag{6.25}$$

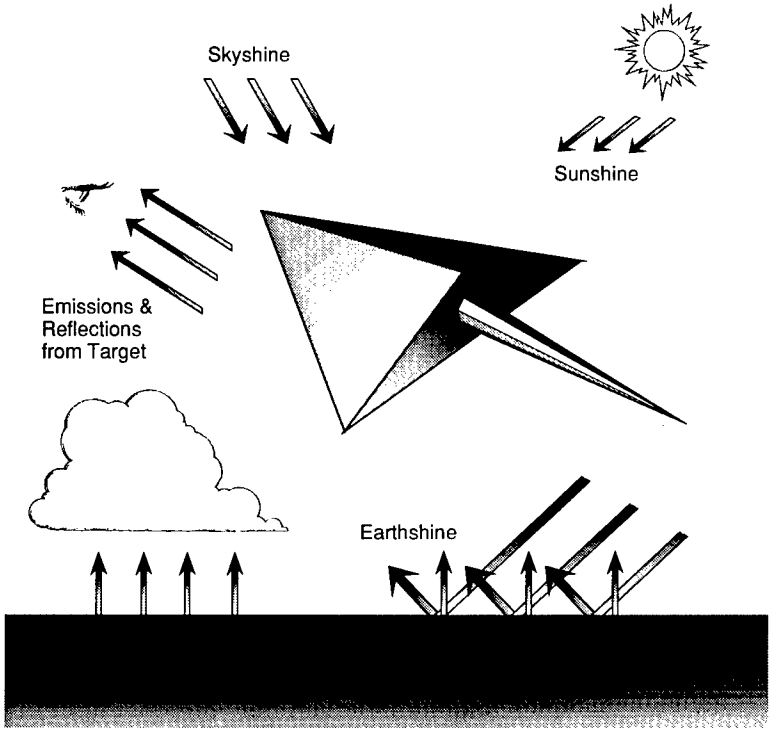


Fig. 6.8 Environmental sources of radiation.

where L_{glint} represents the radiance of the environment source in the ideal glint direction for the surface, either earth or sky (see Fig. 6.8). If the glint is from the sun, Eq. (6.25) holds if an effective solar radiance is used:

$$L_{\text{sun}}(\lambda) = E_{\text{sol}}(\lambda)(\hat{N} \cdot \hat{S})/\Omega_{\text{eff}} \ . \tag{6.26}$$

The factor Ω_{eff} is the effective solid angle for the glint lobe, using the specular BRDF of Eq. (6.9):

$$\Omega_{\text{eff}} = \pi \sin^2(e/2) \ . \tag{6.27}$$

A glint will come from the sun whenever the glint direction is within $\pm e/2$ of the sun's direction.

Knowledge of environmental radiation sources, surface temperatures, and surface optical properties enables us to calculate the radiance from any object surface.

In Sec. 6.4, the irradiance on each of the surfaces of the flying wedge was calculated for both a medium- and long-wavelength bandpass. Equation (6.25), the expression for surface radiance, can be written in terms of the irradiance values calculated in Sec. 6.4. If we assume spectrally averaged values over our MWIR and LWIR bandpasses,

$$L_{\text{surf}} = \varepsilon L^{o}(T) + \rho_d(E_{\text{earth}} + E_{\text{sky}} + E_{\text{sun}})/\pi + \rho_s E_{\text{sun}}/\Omega_{\text{eff}} + \rho_s E_{\text{extended}}/\pi \ , \tag{6.28}$$

where Ω_{eff} is the solid angle from Eq. (6.27) and E_{extended} is the surface irradiance resulting from either earthshine or skyshine in the specular direction,

$$E_{\text{extended}} = \pi L_{\text{extended}} \ . \tag{6.29}$$

For our example calculations, e has been taken to be 10 deg. Thus, surface radiance, resulting from thermal emission, and reflections of earth, sky, and solar radiation can be estimated for each surface and for the total object.

The results of calculating the radiance of each surface of the flying wedge, using this expression and previously derived values for attenuated surface irradiance (whose values are listed in Table 6.4), are displayed in Tables 6.6 and 6.7, which show the radiance contributed by each source for each of the object's surfaces. The product of the total radiance from all sources for each surface with its area yields the radiant intensity for that surface. Summing the radiant intensity of each surface provides the object's total radiant intensity. In Table 6.8, radiant intensity values for the entire flying wedge, as well as each surface, derived from the radiance values in Tables 6.6 and 6.7, are compared with the corresponding blackbody (BB) values from Sec. 6.3 for two views.

Table 6.6 Radiance by Source and Surface for Aft-Ventral Beam View (Attenuated)

Surface Radiance (W/m^2 sr)	Thermal Emission	Earth		Sky		Solar		Total
		ρ_d	ρ_s	ρ_d	ρ_s	ρ_d	ρ_s	
MWIR								
L_A	1.28	0.05	0.10	—	—	—	—	1.43
L_B	1.28	0.02	—	0.02	0.05	0.02	—	1.39
L_C	1.28	0.00	—	0.02	0.05	0.008	—	1.36
L_D	—	surface not observed				—	—	—
LWIR								
L_A	35.7	1.3	2.7	—	—	—	—	39.7
L_B	35.7	0.7	—	0.6	2.4	0.0	—	39.4
L_C	35.7	0.1	—	1.1	2.4	0.0	—	39.3
L_D	—	surface not observed				—	—	—

Note: T = 300 K, $\varepsilon \sim 0.85$, $\rho_d \sim 0.05$, and $\rho_s \sim 0.10$.

Table 6.7 Radiance by Source and Surface for Nose View (Attenuated)

Surface Radiance (W/m^2 sr)	Thermal Emission	Earth		Sky		Solar		Total
		ρ_d	ρ_s	ρ_d	ρ_s	ρ_d	ρ_s	
MWIR Sources								
L_A	—	—	surface not observed				—	—
L_B	—	—	surface not observed				—	—
L_C	1.28	0.00	—	0.02	0.05	0.008	20.7	22.1
L_D	1.28	0.00	—	0.02	0.05	0.00	—	1.35
LWIR Sources								
L_A	—	—	surface not observed				—	—
L_B	—	—	surface not observed				—	—
L_C	35.7	0.1	—	1.1	2.4	0.0	6.2	45.5
L_D	35.7	0.1	—	1.1	2.4	0.0	—	39.3

Numerical 0.0 indicates too small to include; — indicates component does not exist for this view.

Table 6.8 Total and Surface Radiant Intensities Computed for a Blackbody (BB) and by Use of Eq. (6.28) (*Em*-Ref)

Radiant Intensity (W/sr)	I_A	I_B	I_C	I_D	Total
MWIR					
Beam view: BB	20.3	22.2	31.5	—	74
Em-Ref	19.3	20.6	28.6	—	68
Nose view: BB	—	—	32.7	32.7	65
Em-Ref	—	—	481.8	29.4	511
LWIR					
Beam view: BB	567.0	621.6	882.0	—	2071
Em-Ref	536.0	583.1	825.1	—	1944
Nose view: BB	—	—	915.6	915.6	1831
Em-Ref	—	—	991.9	856.4	1848

6.8 ENGINE EXHAUST PLUMES

Turbine engine and rocket motor exhaust gases, often called a *plume,* constitute an important source of infrared radiation. To determine the plume radiation accurately the flow both inside and outside the nozzle must be modeled. Exhaust plume transmission is also important, because the airframe, nozzle, and background can be partially viewed through the plume, which partially absorbs their radiance. Complex numerical techniques, utilizing computational fluid dynamics, are required to make radiometrically accurate plume predictions. Consequently, the approximation techniques presented here are useful only for introducing plume radiation methodologies, and for making rough radiometric estimates. The estimation techniques shown here are most readily applied to aircraft (turbine engine) plumes. In contrast, when applied to missile plumes, estimates become very coarse due to the existence of shock structure and nonequilibrium chemistry.

Turbine engine operating conditions are generally modeled with a *cycle deck,* which is tuned for each specific type of engine. Cycle decks provide estimates of the exhaust flow parameters (i.e., temperature, pressure, velocity, composition) at the nozzle exit plane where an external exhaust flow field starts. There may be more than one engine flow, with a significant amount of mixing either within or outside of the nozzle. For turbojets, there is only a single flow. Turbofans have a hot core flow and a warm fan flow, which may mix inside the nozzle for some engine types. On some engines, the fan gases are ducted separately all the way to the exit plane, and there the mixing begins. Figure 6.9 depicts three turbofan engine nozzle configurations. In the first engine type, mixing occurs only outside the nozzle, and the cycle deck output can be used. In the second and third engine types there is partial or full mixing within the nozzle. If the engine has an afterburner, that internal turbulent combustion may cause the core and fan flows to mix completely, resulting in a single flow.

Formal computation of IR radiation in an external plume proceeds through a number of steps, starting with computation of the nozzle exit plane gas conditions, using engine cycle information, and nozzle mixing analyses. The exit plane conditions are used, along with the ambient flight conditions, to compute the external plume flow field and radiation.

Some of the same assumptions used to simplify a formal treatment of plume flow fields can be employed in simpler plume radiation estimation techniques. The exit plane is often assumed to be nonuniform, but axisymmetric about the nozzle centerline, because most nozzles in current use are circular. Many aircraft are now being designed with two-dimensional (noncircular) nozzles, for which more sophisticated treatments must be employed.

Mixing in jet engine exhausts is affected by the core and fan flow parameters, the nozzle geometry, and structural elements, such as mixing vanes. A reasonable simplifying assumption is to neglect the effects of the nozzle walls on the mixing.

Nonuniformity at the engine exit can be traced to several sources. The first is the radial flow variation at the exit of the last turbine stage. This variation is a result of the spanwise work distribution in the turbine chosen by the turbine designer. It varies with engine operating conditions but has generally been measured as part of the engine development program. Another source of engine exit variation is the wake from obstructions in the flow such as struts.

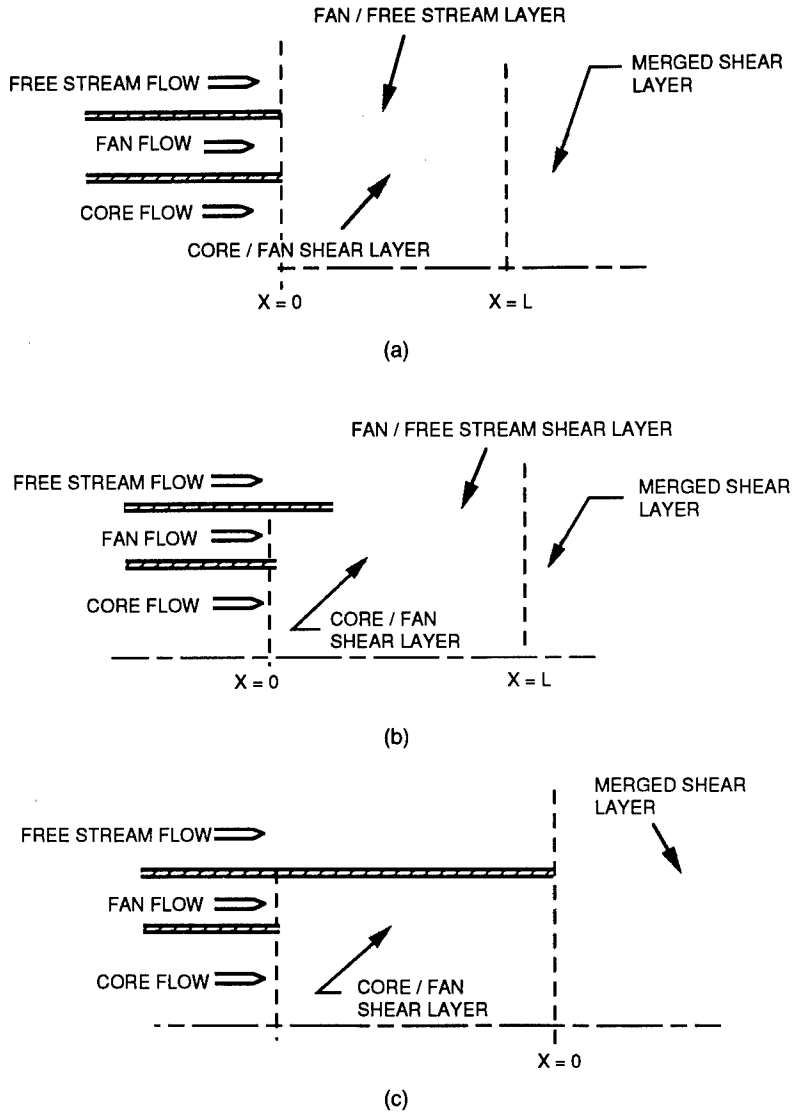


Fig. 6.9 Diagrams of mixing between core, fan, and free stream flows for three nozzle configurations: (a) type 1 engine, (b) type 2 engine, and (c) type 3 engine.

Other observed nonuniformities have not been traced to specific causes but may result from a poor combustor pattern factor (uniformity).

As exhaust gases flow through a nozzle, they isentropically expand and contract because of the varying cross section down the nozzle. For some engines, the exit-plane pressure is not equal (or pressure matched) to ambient pressure (particularly true with fixed-nozzle engines at high power). Most modern augmented aircraft are equipped with fully variable nozzles that expand the nozzle flow to ambient pressure. Such pressure-matched flow has weak shock structure, so that fairly accurate results can be obtained by manually computing pressure-matched exit conditions from the true exit conditions.

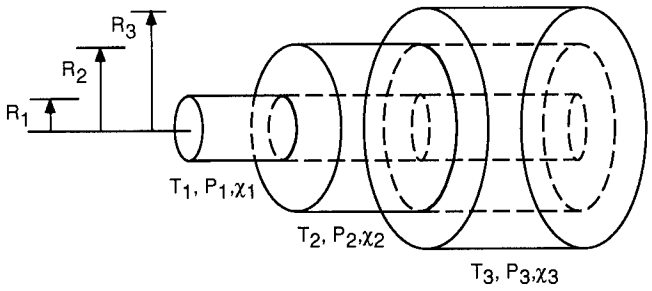


Fig. 6.10 Cross-sectional view of plume showing specification of plume properties for three cylindrical regions.

6.8.1 Plume Flow Properties

Calculations of the IR radiation from a plume require a characterization of the temperature, pressure, and molecular concentrations of the flow field. Axisymmetric flows can be conveniently represented for radiance computations by a set of concentric, cylindrical shells. The flow properties must therefore be known as functions of the axial and radial (z,r) coordinates within each cylindrical shell.

As shown in Fig. 6.10, the plume properties are assumed to be constant within each annular region between cylinders. The plume properties, for the purpose of calculating emission, are specified by temperature T, pressure P, species mole fractions X_j, and the plume physical dimensions of cylinder radius R, and axial location Z. In a detailed treatment, the cross-sectional properties of the plume may be specified at many axial locations downstream of the exit plane.

To determine the spatial distribution of radiation in a plume, the radiation transport must be calculated for lines of sight through the plume.

6.8.2 Plume Radiance and Transmittance Calculation

If plume properties are defined along a line of sight, the radiative transport for that line of sight can be determined. The form of the radiative transport equation can be taken to be[9]

$$L(\lambda) = \int_0^s L_0(\lambda,T)\frac{d\tau(\lambda,s)}{ds}\,ds\ , \tag{6.30}$$

where

$L(\lambda)$ = spectral radiance for line of sight from 0 to S
$L_0(\lambda,T)$ = Planck function spectral radiance at wavelength λ and temperature T(s)
$\tau(\lambda,s)$ = transmittance to the observer at wavelength λ and distance s along line of sight
s = distance along line of sight (increasing in the direction of light travel).

Equation (6.30) assumes that no scattering occurs within the plume. The plume properties are assumed to be constant within spatial regions along the line of sight, allowing the integral to be evaluated piecewise, and thus giving a summation:

$$L(\lambda) = \sum_{i=1}^{i=i\max} L_0(\lambda,T_i)[\tau_i(\lambda) - \tau_{i-1}(\lambda)] \ , \tag{6.31}$$

where i denotes the i'th spatial region along the line of sight and $\tau_i(\lambda)$ is the total transmittance from the outside edge of the plume facing the observer through the i'th spatial region. The transmittance from the observer to the edge of the plume $\tau_{i\max}(\lambda)$ is defined as equal to unity to calculate the source emission. Note that for a single uniform "slab" of gas, we would have

$$L(\lambda) = L_0(\lambda,T)[1 - \tau(\lambda)] \ .$$

The calculation of the spectral transmittance along a line of sight, τ, can be quite complex. When transmittance is computed at high spectral resolution, so that the gas absorption is nearly constant at each spectral point, it follows a simple exponential known as Beer's law. Such resolutions are typically around 0.1 cm^{-1}. However, because most passive infrared systems operate with spectral passbands of the order of 1000 cm^{-1} wide, an analyst would need to compute about 10,000 spectral points.

At coarser spectral resolutions (typically 5 to 20 cm^{-1}), the gaseous absorption is averaged. Because the exponential of an average value does not equal the average of an exponential, Beer's law leads to inaccuracies at low and moderate resolutions. The handbook by Ludwig et al.[9] provides an excellent introduction to this problem.

One of the most important plume emission regions results from CO_2, extending roughly from 4 to 5 μm (2000 to 2500 cm^{-1}). CO_2 is abundant in most exhausts, and it is a strong absorber. Nonafterburning turbine engine exhausts are typically at temperatures of 500 to 800 K at the engine exit, thus the Planck function peaks near the CO_2 region. (Hotter exhausts emit strongly at shorter wavelengths, but are still intense in the 4- to 5-μm region.) Fortunately, that CO_2 band has such a dense line structure at plume temperatures, that its absorption does not vary much within 10-cm^{-1} spectral bins. Beer's law can thus be used with moderate validity.

Optically active species and their associated bands include: CO_2 (2.7, 4.3, 15.0 μm), H_2O (2.7, 6.3 μm), and CO (4.65 μm). Absorption coefficients for these and other molecules can be obtained from Refs. 9, 10, and 11. The CO_2 4.3-μm band model parameters of Bernstein et al.[11] are particularly useful for calculating exhaust plume radiation in the MWIR.

6.8.3 Estimating Plume Radiation in the MWIR

Calculating the radiation from exhaust plumes is complex because their shapes and volumes are not fixed for all flight conditions and they tend to be optically translucent. Aircraft are typically endowed with weak exhaust plumes, having relatively low temperature gases (several hundred versus several thousands

of degrees Celsius) and/or low concentrations of optically active molecular species. Radiation from such plumes are primarily confined to the rotational band systems associated with the exhausted products of combustion, including CO_2, H_2O, and sometimes CO. The fact that these same gases occur in the surrounding atmosphere in varying concentrations adds considerable complexity to the process of radiation transport, and thus to IR signature calculations. (Radiation from hot gas can be absorbed by colder gas, which itself radiates weakly.)

Systems that operate by sensing radiation at wavelengths around 3 to 5 μm are often specifically designed to detect radiation from hot CO_2, exploiting this spectral regime's absence of water vapor absorption bands. Although absorption by atmospheric water vapor is thus avoided, atmospheric CO_2 is sufficient to significantly reduce the amount of radiation near the band center. The hotter the exhaust gas, the more radiation emanates from wavelengths far from the band center (the wings of the band). Absorption by cooler intervening gas selectively removes the radiation from the band center, causing a decrease in intensity at these wavelengths. This well-known phenomena of band reduction can be better understood by examining the curves in Fig. 6.11, which show the absorption coefficient (a measure of the ability of the gas to absorb radiation) for CO_2 gas at different temperatures.

IR technology that exploits radiation at longer wavelengths, from 8 to 12 μm, exploits a spectral region that is devoid of strong CO_2 or H_2O bands. In fact, unless an aircraft is employing rocket-propelled or afterburner-assisted flight, plume radiation can generally be neglected as a significant contributor to the total long-wavelength IR signature.

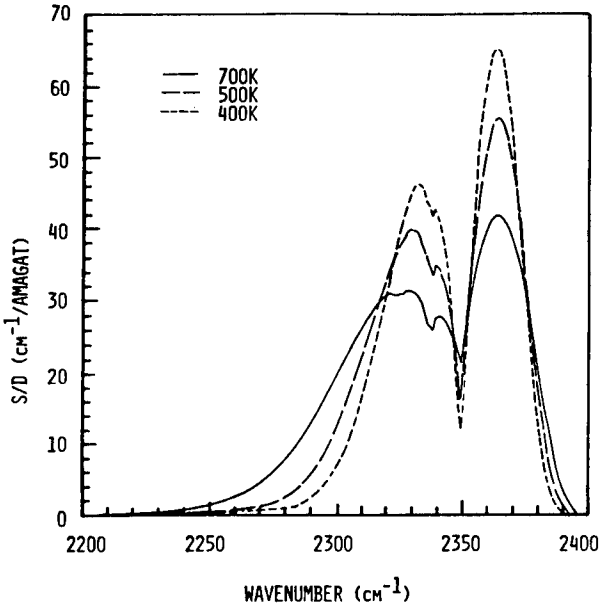


Fig. 6.11 CO_2 4.3-μm absorption coefficients at different band temperatures. These values have been averaged over 5-cm^{-1} spectral bins, so that no line structure is apparent.

In the MWIR, the plume may represent the strongest signature component (depending on aspect and the range to the observer) and therefore should be dealt with carefully. In general, the total radiation from a volume of hot gas depends on its transmissivity τ, which in turn depends on the gas temperature T, the concentration of emitting species χ_i, and their distribution throughout the volume of the gas. The transmissivity of a gas along a line of sight can be written using Beer's law as

$$\tau(T,v) = \exp\left[-k(T,v)\chi P\left(\frac{273}{T}\right)D \right] , \tag{6.32}$$

where P is the gas pressure; $k(T,v)$ is the absorption coefficient at temperature T and wavenumber v; and D is the distance through the gas along the optical path. We define the *optical depth* to be

$$u(v) = k(T,v)\chi P\left(\frac{273}{T}\right)D . \tag{6.33}$$

For optical depth greater than 1, the gas absorbs strongly, and radiates approximately as a blackbody.

Exhaust plumes can generally be divided into regions within which both temperature and species concentrations can be assumed for simplicity to be constant. The bulk of the radiation from jet engine exhaust plumes comes from the vicinity of the exit plane in a region called the inviscid core. As much as three-quarters of the total plume radiation may be generated within a length of plume shorter than the structural length of the aircraft. Cooler plume regions emit more weakly and are more easily absorbed by the cool atmospheric gas. It is thus important for IR calculations to estimate the conditions, size, and shape of the hottest exhaust gas.

For a circular axisymmetric nozzle, the inviscid core is roughly cone shaped (see Fig. 6.12). The base of the cone is the area of the nozzle exit (with radius R_0). Its length L has been empirically determined[12] to be dependent on exhaust Mach number M_e, an eddy mixing coefficient κ, and the density differential between the ambient free stream ρ_0 and the exhaust gas ρ_j:

$$\Delta\bar{\rho}_e = \rho_j/\rho_0 ,$$

so that

$$L = R_0 X/\kappa(\Delta\bar{\rho}_e)^{1/2} , \tag{6.34}$$

where

$$\kappa = 0.08(1 - 0.16\, M_e)(\Delta\bar{\rho}_e)^{-0.22} , \tag{6.35}$$

and X is a characteristic axial scale length empirically found[12] to be equal to 0.70. (Note that assuming pressure matched conditions, $\Delta\bar{\rho}_e \approx 1.0$.) For an

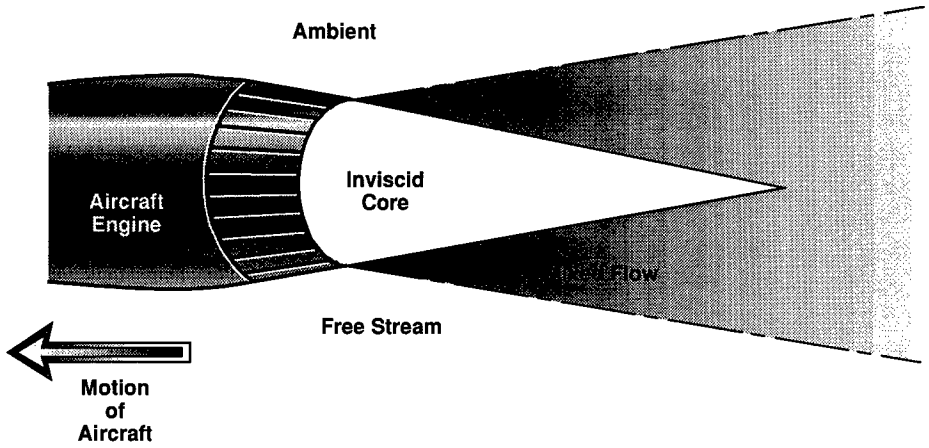


Fig. 6.12 Simplified aircraft jet engine exhaust plume.

exit Mach number equal to 1.4, for example, $\kappa = 0.062$ and the length of the inviscid core is 11.3 times the nozzle radius.

Estimating the radiant intensity of an exhaust gas at MWIR wavelengths entails calculating an average plume radiance and an effective, unobstructed plume core projected area. Because the plume is translucent, its radiance is spatially variable even assuming an exhaust gas whose inviscid core is at uniform temperature, pressure, and species concentration. Depending on the core geometry (which depends on the nozzle shape), different lines of sight penetrate different column lengths of optically active gas.

The degree of plume translucence along any line of sight is determined by the optical depth of the emitting gas, and is spatially variable over the plume. The average optical depth, over a plume whose properties are constant within a volume, can be estimated by solving for the average line of sight path length $\bar{x}$ through a plume. For a plume seen broadside,

$$\bar{x} = \frac{\iint_{A_p} x \, dy \, dz}{\iint_{A_p} dy \, dz} = \frac{\text{volume}}{\text{area}} ,$$

when A_p is the projected area of the plume.

For a cone viewed broadside, the solution is

$$\bar{x} = \frac{\text{volume}}{\text{area}} = \frac{(\pi R_0^2 L)/3}{R_0 L} = \frac{\pi R_0}{3} .$$

For example, if $R_0 = 0.625$ m, then $L = 7.06$ m and $\bar{x} = 0.654$ m. This average

pathlength can be used to estimate an average optical depth for a homogeneous cone shaped plume, as well as an average pathlength (APL) transmittance τ^{APL}:

$$\bar{u} = k_{CO_2}\chi_{CO_2}p\left(\frac{273}{T}\right)\bar{x} \ , \tag{6.36}$$

$$\tau^{\mathrm{APL}} = \exp(-\bar{u}) \ . \tag{6.37}$$

Note that calculating the average transmittance $\bar{\tau}$, and therefore a true average radiance, entails solving

$$\bar{\tau} = \frac{\int_A\int \tau(y,z)\, dy\, dz}{\int_A\int dy\, dz} = \frac{\int_A\int \exp[-u(y,z)]\, dy\, dz}{A_p} \ ,$$

a much more difficult proposition, especially for off-beam perspectives of the plume. Instead, we can consider the plume radiance for two limiting cases: an optically thin and thick gas.

In the optically thin case, u is everywhere small, and

$$\tau_{\mathrm{thin}}^{\mathrm{APL}} = e^{-\bar{u}} \approx 1 - \bar{u} = 1 - \left[k_{CO_2}\chi_{CO_2}p\left(\frac{273}{T}\right)\bar{x}\right] \ , \tag{6.38}$$

or

$$\tau_{\mathrm{thin}}^{\mathrm{APL}} = 1 - K\bar{x} \ , \tag{6.39}$$

where

$$K = k_{CO_2}\chi_{CO_2}p\left(\frac{273}{T}\right) \ . \tag{6.40}$$

Because we are considering a uniform region of gas, we can write

$$L_{\mathrm{thin}}^{\mathrm{APL}} = (1 - \tau)L_0 \approx K\bar{x}L_0 \ . \tag{6.41}$$

The total radiant intensity I from the plume is then

$$I_{\mathrm{thin}}^{\mathrm{APL}} = \int L dA \approx A_p L_{\mathrm{thin}}^{\mathrm{APL}} \tag{6.42}$$

$$I_{\mathrm{thin}}^{\mathrm{APL}} = A_p K\bar{x}L_0 \ . \tag{6.43}$$

However, $A_p\bar{x}$ is an effective volume V of the plume, so

$$I_{\text{thin}}^{\text{APL}} = KL_0V \ . \tag{6.44}$$

Thus, in the optically thin case, the radiant intensity of the plume is proportional to the volume. Although a broadside perspective was initially assumed, the optically thin limit is also true for the more general case of any observer direction that provides an unobstructed view of the plume.

In the optically thick limit, the transmissivity is near zero, so $\bar{L} \approx L_0$. In this case, the radiant intensity is proportional to the plume projected area. In most normal circumstances, the plume will be neither completely optically thin or thick, so the value of the radiance will be intermediate to the two extremes. When the observer's perspective changes from its direct beam view, the average path length, which was used to determine a representative optical depth and the gas transmittance, will also change. For near broadside views, the off-beam average optical depth $\bar{x}'$ can be taken to be

$$\bar{x}' = \bar{x}/\sin\alpha \ ,$$

where α is the aspect angle measured from nose-on.

To calculate an exhaust plume's radiant intensity, its projected area must be determined. Estimating the projected area of the inviscid core of exhaust gases exiting nonaxisymmetric, rectilinearly shaped nozzles can be accomplished by treating the plume as if it were comprised of individual flat surfaces. In that case, calculating the projected area of plume surfaces proceeds in the same manner as for the airframe. Calculating the plume projected area for curvilinear exhaust nozzles may be more difficult. For example, a circular nozzle produces a conically shaped inviscid core. The projection of a cone seen broadside is a simple triangle. Viewing the tail of the aircraft presents a circular plume. For aspect angles between broadside (90 deg) and tail (180 deg), the projected shape may be represented by a combination of ellipse and triangle. The projection of a cone shaped inviscid core may present one of four different views, depending on the aspect angle α or the equivalent off-broadside angle ($\phi = \alpha - 90$):

1. direct beam view ($\alpha = 90$ deg or $\phi = 0$ deg): triangular cross-sectional area
2. off-beam view ($0 \text{ deg} < \phi < \phi_1$): full ellipse area plus small extended triangle
3. near-tail view ($\phi_1 \leq \phi < 90$ deg): full ellipse area
4. tail view ($\phi = 90$ deg): full circular area,

where the value of ϕ_1 is determined by the ratio of nozzle radius R_0 and core length L. These possibilities can be represented by a triangular surface (the inviscid core) attached orthogonally to a circular surface (the nozzle), shown in Fig. 6.13.

The constant ϕ_1 is the aspect angle at which the triangle's projected length, $L\cos\phi$, is less than the semiminor axis of the ellipse, or

$$L\cos\phi_1 = R_0\sin\phi_1 \ , \tag{6.45}$$

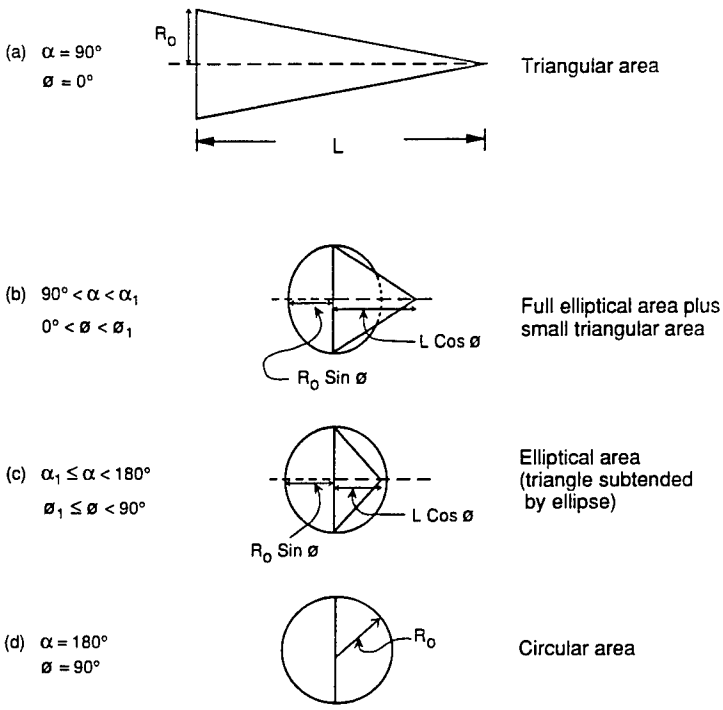


Fig. 6.13 Four different perspectives for simplified model of a conical plume each of which demands specific approach to calculating plume area.

$$\frac{L}{R_0} = \frac{\sin\phi_1}{\cos\phi_1} = \tan\phi_1 \ , \tag{6.46}$$

$$\phi_1 = \tan^{-1}\left(\frac{L}{R_0}\right) \ . \tag{6.47}$$

For aspect angles between 0 deg and ϕ_1, the projected area of the inviscid core is comprised of the area of the full ellipse and the small portion of the plume that lies beyond the boundaries of the ellipse. In this case the approximate projected area is

$$A_p = \pi R_0^2 \sin\phi + \frac{R_0}{L\cos\phi}(L\cos\phi - R_0\sin\phi)^2 \ . \tag{6.48}$$

For a nozzle whose radius is 62.5 cm and a plume whose inviscid core length is 7.06 m, $\phi_1 = 85$ deg. An observer at 110-deg aspect ($\phi = 20$ deg) sees a plume core projected area of 4.3 m^2. Because the plume is translucent and conical, the effective projected area may be somewhat less, because, at the edges of the cone, the column length of radiating gas is much less than at the center. Thus, the value of the radiant intensity calculated for the entire core may be somewhat inflated.

More important, radiation from the inviscid core must pass through a region of mixed exhaust and ambient gases (see Fig. 6.12). This region diverges into the ambient free stream with distance downstream of the nozzle. The warm gas contains exhaust products that are very effective in absorbing the radiation from the core, particularly at wavelengths near the band center. Temperatures in this region vary radially, being hot near the core, and cold (ambient) at the outer edge. We approximate this region by assigning uniform warm gas to the inner half of the diverging sheath, as shown in Fig. 6.14. For the example discussed in the following, we take the pathlength of this warm intervening absorbing gas to be equal to the difference between the nozzle radius and half the average core pathlength ($\bar{x}/2$). We must calculate both the transmittance and emission from this shroud of warm gas surrounding the inviscid core. Thus, calculating a plume's radiant intensity entails first calculating the optical depth, transmittance, and radiance for two separate gas regions. These calculations are most simply performed using values of absorption coefficient for small spectral regions or sub-bands that span the molecular band. The resulting radiances are then spectrally integrated to obtain the average source radiance for each gas.

Tables 6.9 and 6.10 show the results of radiance and transmittance calculations for a plume observed from 110-deg aspect (20 deg tailward of beam). Table 6.9 gives the spectral sub-band results of radiance and transmittance for the sheath of warm gas surrounding the hotter inviscid core. The core radiance and transmittance values are shown in Table 6.10, which also lists the spectral sub-band radiance transmitted through the warm sheath. Addition of the sub-band radiance values from hot core and warm sheath represent a total average source radiance for the plume. The absorption coefficients for this computation were taken from Fig. 6.11.

The product of the average source radiance for each gas, times its projected area, yields the radiant intensity from that region. For the warm sheath surrounding the inviscid core, the projected area is of a concentric cylinder whose

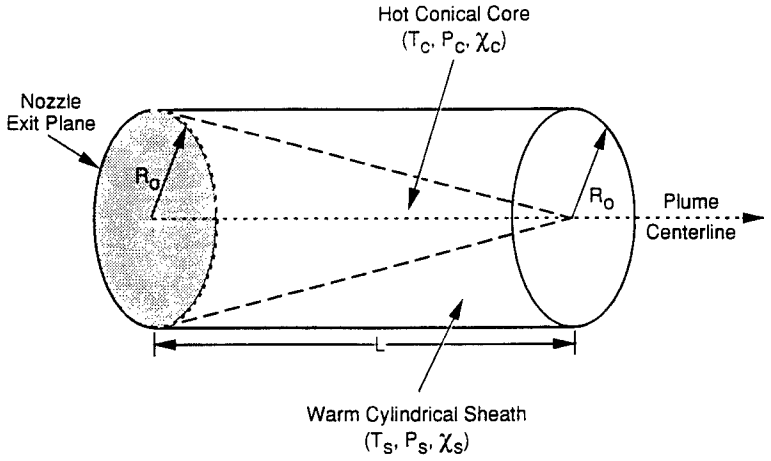


Fig. 6.14 Geometry of plume approximation showing conically shaped inviscid hot core of length L, with properties T_C, P_C, and χ_C, embedded in a cylindrical sheath of warm gas, whose properties are defined by T_S, P_S, and χ_S.

Table 6.9 Transmittance and Radiance for the Warm Sheath of Mixed Gases Surrounding the Inviscid Core, at 520 K, 0.02 Mole Fraction of CO_2, and 0.89 Atmospheres of Pressure, Observed at 110-deg Aspect with Average Column Length of 31.69 cm

Spectral Sub-Band (cm^{-1})	Average Absorption Coefficient at STP (cm^{-1})	Warm Gas Blackbody Radiance (W/sr m^2) $L_0^W(T,\upsilon)$	Warm Gas Transmittance τ_W	Warm Gas Emitted Radiance $(1 - \tau_W)L_0^W$
2216–2255	0.0	10.71	1.0	0.0
2256–2285	1.2	7.570	0.70	2.2
2236–2315	13.0	7.245	0.021	7.1
2316–2345	40.0	6.932	7.2(−6)	6.9
2346–2375	54.0	6.628	1.1(−7)	6.6
2376–2405	1.2	6.335	0.70	1.90
Total (spectrally summed)				24.7

Table 6.10 Transmittance and Radiance for the Inviscid Core Gas at 760 K, 0.04 Mole Fraction of CO_2, and 0.89 Atmosphere Pressure, Observed at 110-deg Aspect, with Average Column Length of 65.45 cm

Spectral Sub-Band (cm^{-1})	Average Absorption Coefficient at STP (cm^{-1})	Hot Gas Blackbody Radiance (W/sr m^2) $L_0^H(T,\upsilon)$	Hot Gas Transmittance τ_H	Hot Gas Emitted Radiance $(1 - \tau_H)L_0^H$	Warm Gas Transmittance τ_W	Hot Transmitted Radiance $(1 - \tau_H)\tau_W L_0^H$
2216–2255	0.8	76.5	0.51	37.3	1.0	37.3
2256–2285	4.5	55.7	0.023	54.4	0.70	38.1
2236–2315	20.0	54.7	5.4(−8)	54.7	0.021	1.1
2316–2345	31.3	53.7	4.2(−12)	53.7	7.2(−6)	0.0
2346–2375	41.0	52.7	1.2(−15)	52.7	1.1(−7)	0.0
2376–2405	4.0	51.6	0.035	49.8	0.70	34.9
Total (spectrally summed)						111.4

radius is equal to that of the nozzle (R_0). In this case, the projected area of such a cylinder is 8.7 m^2 and the total radiant intensity from the entire plume is about 700 W/sr.

6.9 SUMMARY

Employing the estimation techniques compiled in previous sections, it is possible to calculate an approximate intensity for the flying wedge, which includes the most important contributing phenomena. This approximation includes object radiation resulting from thermal energy from both hardbody and exhaust

gases. Environmental sources of reflected radiation are accounted for including the sun, the earth, the atmosphere, and clouds.

The integrated source radiant intensity of the flying wedge, flying in straight and level flight, at 1000 m, at Mach number 0.5, under a thick, but not total, overcast (the object is illuminated by the sun) is found by simply summing each of the contributing sources of radiation. It is instructive to understanding the signature phenomenology to decompose the total source radiant intensity into the contributions from each source from each surface. As illustrated in previous sections, the relative importance of different radiation mechanisms depends on the viewing aspect.

The relative contribution from each source of radiation for the flying wedge is revealed by summing the individual amounts due to each mechanism. This is shown for each aspect and bandpass in Table 6.11. Thus, for the observer with a view slightly aft and below the object's beam, the total source radiant intensity is, at MWIR wavelengths,

$$I_{\text{beam}} = 770 \text{ W/sr} ,$$

and at LWIR wavelengths,

$$I_{\text{beam}} = 1940 \text{ W/sr} .$$

For the observer with a nose view, the total source radiant intensity is, at MWIR wavelengths,

$$I_{\text{nose}} = 510 \text{ W/sr} ,$$

and at LWIR wavelengths,

$$I_{\text{nose}} = 1850 \text{ W/sr} .$$

When a solar glint is observed in the MWIR, or when a hot, optically thick exhaust plume is in view, these values depart significantly from those obtained by assuming the object behaves as a simple blackbody. The real difference can be seen in the mechanisms that produce the radiation from each surface, and in their variation with observed aspect.

Table 6.11 Radiant Intensity in Watts per Steradian Contributed by Each Source, for Each Wavelength and Case

Case and Bandpass	Thermal Emission	Earthshine Reflections	Skyshine Reflections	Solar Reflections	Plume Emission	Total
Ventral beam—MWIR	63.1	2.3	2.5	0.5	700.0	770
Nose—MWIR	55.8	0.0	3.1	451	0.0	510
Ventral beam—LWIR	1760	67.0	118	0.0	0.0	1940
Nose—LWIR	1560	4.4	153	135	0.0	1850

The atmospheric transmission losses from the object to the observer, particularly important when viewing an exhaust plume, can be accommodated by estimating an in-band transmittance, or by using atmospheric models, such as the Air Force LOWTRAN series. Intersurface reflections can be estimated by treating surface pairs in a manner similar to the calculation for surface point source illumination.

Other aspects are calculated in a similar fashion to the approach presented here. More complex targets typically require a significant increase in the number of surfaces that must be modeled. In this event, the advantage of using numerical techniques is clear.

References

1. C. R. Wylie, *Advanced Engineering Mathematics,* McGraw-Hill, New York (1975).
2. A. Berk, L. S. Bernstein and D. C. Robertson, *MODTRAN: A Moderate Resolution Model for LOWTRAN 7,* Spectral Sciences, Inc., Burlington, MA, Final Report No. GL-TR-89-0122 (1989).
3. G. F. Byrne, J. E. Begg, P. M. Fleming, and F. X. Durin, "Remotely sensed land cover temperature and soil water status—a brief review," *Remote Sensing of the Environment* **8,** 291–305 (1979).
4. P. Beckman and A. Spizzichino, *The Scattering of Electromagnetic Waves from Rough Surfaces,* Macmillan, New York (1963).
5. K. E. Torrance and E. M. Sparrow, "Theory for off-specular reflection from roughened surfaces," *Journal of the Optical Society of America A* **57,** 1105–1114 (1967).
6. T. S. Trowbridge and K. P. Reitz, "Average irregularity representation of a rough surface for ray reflection," *Journal of the Optical Society of America A* **65,** 531–536 (1975).
7. F. E. Nicodemus, J. C. Richmond, J. J. Hsia, I. W. Ginsberg, and T. Limperis, *Geometrical Considerations and Nomenclature for Reflectance,* National Bureau of Standards Monograph 160 (Oct. 1977).
8. J. P. Holman, *Heat Transfer,* pp. 194–203, McGraw-Hill, New York (1981).
9. C. B. Ludwig, W. Malkmus, and J. A. L. Thompson, *Handbook of Infrared Radiation from Combustion Gases,* NASA SP-3080 (1973).
10. S. J. Young, "Band model parameters for the 2.7 μm bands of H_2O and for the 2.7 μm bands of H_2O and CO_2 in the 100–3000°K temperature range," The Aerospace Corporation, El Segundo, CA (July 1975).
11. L. S. Bernstein, D. C. Robertson, and J. A. Conant, "Band model parameters for the 4.3 μm CO_2 band from 200 to 3000°K—II. Prediction, comparison to experiment, and application to plume emission—absorption calculations," *Journal of Quantitative Spectroscopy & Radiative Transfer* **23**(2), 169 (1980).
12. P. O. Witze, "Centerline velocity decay of compressible free jets," *AIAA Journal* **12,** 417–418 (1974).

Index