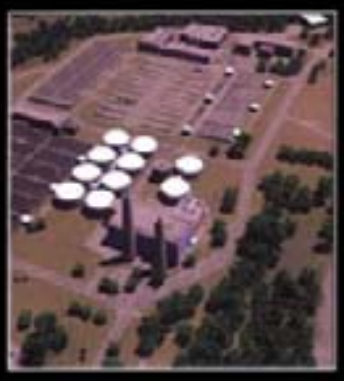
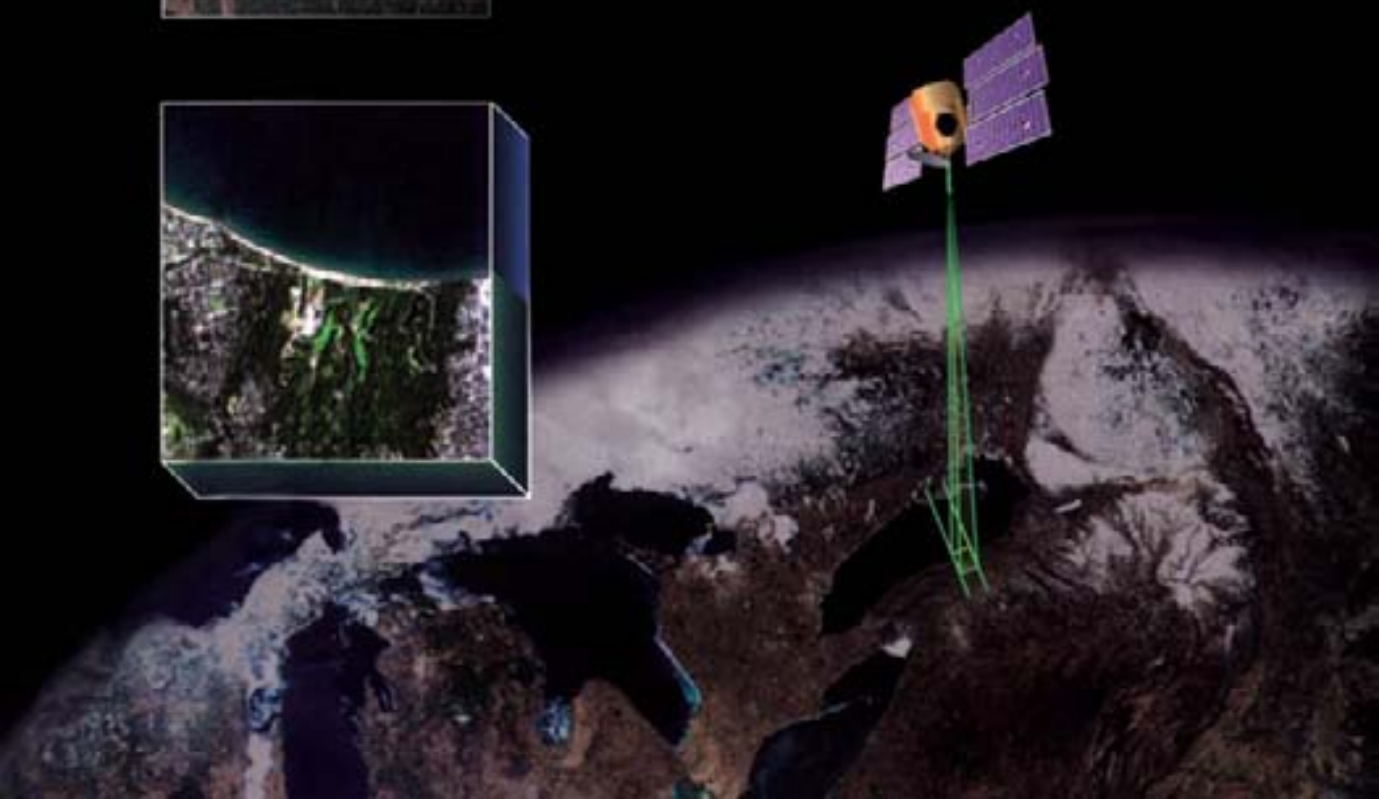
REMOTE SENSING

THE IMAGE CHAIN APPROACH

2ND EDITION

JOHN R. SCHOTT





REMOTE SENSING

2ND EDITION

This page intentionally left blank

REMOTE SENSING
THE IMAGE CHAIN APPROACH

2ND EDITION

JOHN R. SCHOTT
ROCHESTER INSTITUTE OF TECHNOLOGY

OXFORD
UNIVERSITY PRESS
2007

OXFORD
UNIVERSITY PRESS

Oxford University Press, Inc., publishes works that further
Oxford University's objective of excellence
in research, scholarship, and education.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

www.oup.com

Oxford is a registered trademark of Oxford University Press

Library of Congress Cataloging-in-Publication Data
Schott, John R. (John Robert), 1951–
Remote sensing : the image chain approach /
John R. Schott. — 2nd ed.
p. cm.
Includes index.
ISBN: 978-0-19-517817-3
1. Remote sensing. I. Title.
G70.4.S36 2006
621.36'78—dc22 2006022714

9 8 7 6 5 4 3 2 1
Printed in the United States of America
on acid-free paper

Dedicated with love and respect to
Pam, Stephen, Jesse, and Jack

This page intentionally left blank

PREFACE

I'm back at Wyldewood again almost exactly 10 years after I wrote the preface to the first edition. Much has changed. The reader familiar with the first edition will see this reflected in the expansion of this text from 10 to 14 chapters. Much of the expansion covers the increased emphasis by the remote sensing community on imaging spectroscopy. I've also tried to fill in many subtle points that may have been glossed over in the first edition - things I've learned or been taught by students or generous colleagues.

Much has not changed, so I'll let the preface for the first edition continue to tell the story with a few factual updates noted in italics.

J.R.S
Wyldewood Beach
Port Colbourne, Ontario
February 24, 2006

Preface to the First Edition:

Much of the remote sensing literature is written by and for the applications specialists who are users of remotely sensed data. The remote sensing "science" is often neglected or given only cursory treatment because of the need to stress the principles of the application area (e.g. geography, geology, forestry). Those books that more directly address remote sensing as a discipline have tended to heavily emphasize either the optics and physics of remote sensing or the digital image processing aspects.

This book treats remote sensing as a continuous process, including energy–matter interaction, radiation propagation, sensor characteristics and effects, image processing, data fusion, and data dissemination. The emphasis is on the tools and procedures required to extract information from remotely sensed data using the image chain approach.

This approach to remote sensing has evolved from over *two* decades of teaching remote sensing to undergraduate and graduate students and *three* decades of research and consulting on remote sensing problems for government and industry. That experience has often shown that individuals or organizations all too often focus on one aspect of the problem before considering the entire process. Usually this results in a great deal of time, effort, and expense to achieve only a small improvement, because all the effort was placed somewhere other than the weak link in the chain. As a result, the perspective on remote sensing presented here is to treat the process as a continuous flow and to study the underlying science to a level sufficient to understand the many constrictions that limit that flow of information to the eventual user.

Because the field of remote sensing is so large, I have chosen to limit the treatment to aerial and satellite imaging for earth observation. In addition, because the vast majority of remote sensing is done passively in the visible through the thermal infrared region, I have emphasized this area. Within this spectral region, the underlying science and techniques of quantitative radiometric image acquisition, image analysis, and *spectral* image processing are emphasized. The details of specific sensors and software packages are downplayed because of their ephemeral nature, and photo interpretation and photogrammetry are only briefly introduced because of their thorough treatment elsewhere.

In writing, I've always had two audiences in mind. The first is the traditional student. As a text, this book is aimed at graduate students in the physical or engineering sciences taking a first course in remote sensing. It would also be appropriate for advanced undergraduates or as a second course for students in applications disciplines. In several cases, where the mathematical principles may be beyond what students in a particular discipline are required to know (e.g., 2-D linear systems theory), I have attempted to use more extensive graphical and image examples to provide a conceptual understanding. I have assumed a working knowledge of university physics and calculus. In addition, parts of chapters *8* and *13* draw on linear systems theory, although these sections can be treated in a more descriptive fashion when necessary.

The second audience I had in mind when writing is the large number of scientists and engineers who were trained in the traditional disciplines and find themselves working in the remote sensing field. Having worked extensively with many of these scientists in government and industry, I wanted to compile a book that could be used to understand and begin to address many of the questions these individuals must first ask of the remote sensing field.

I hope both of these groups will find this a useful tool box for working the problems ahead.

J.R.S.
Wyldewood Beach
Port Colbourne, Ontario
December 29, 1995

ACKNOWLEDGMENTS

They say that memory is the first thing to go, and it must be so, because I'd forgotten how much effort by how many people goes into bringing a project of this scale to completion. Between the two editions, the cast of characters is truly enormous, and I have been very fortunate to have worked with such capable people. Rebecca Ruby, Nick Herber, and Marie Peters along with Margôt Delage on the first edition were responsible for turning my indecipherable hen scratches into the graphics included throughout this text. Nick Herber, Sharah Naas, and Marie Peters did the layout for the second edition, and Emmett Ientilucci did the first along with the bulk of the graphics for both covers. Once again, Nina Raqueño helped with many of the more technical issues associated with graphics. Donald Taylor, Adam Goodenough, and Paul Barnes made many of the analytical plots and associated computations. Scott Brown and Emmett Ientilucci produced many of the image examples, including the image processing. Rolando Raqueño and Tim Gallagher provided software and hardware support, respectively, to fill in any gaps the rest of us couldn't fill.

My friend and right arm for many years, Carolyn Kitchen, came out of retirement to convert my illegible rough drafts into processed words. Cindy Schultz provided the organization, communication, and infrastructure to keep everything moving and especially ensuring that I could steal the time to work on this project. As you all know, so very well, I could not have begun to do this without you, nor would it have been any fun, and I am deeply grateful.

I am also pleased to acknowledge the research staff and numerous students who have worked with me over the years in the Digital Imaging and Remote Sens-

ing Laboratory. Their inquiries, insights, energy, and enthusiasm for this field have kept me as intrigued with it today as when I was first captivated by it at the age of twenty. You have collectively given me a gift that few people ever experience, and I truly cherish it.

It is always gratifying when you reach out to busy professionals and receive their generous response. In trying to gather data and illustrations for this text, I was met over and over, not by obstacles, but by generous individuals who went the extra mile to help out. The individual organizations are referenced with each figure, but I want to specifically recognize and thank some of the individuals who were so generous. To Carroll Lucas of Autometric; Steve Stewart, Dan Lowe, and Robert Horvath of ERIM; Paul Lowman, Jim Irons, and Darrel Williams of NASA Goddard; Larry Maver of ITEK; Al Curran of Georgia Tech; Phil Slater and Bob Schowengerdt of the University of Arizona; Tom Chrien and Robert Crippen of NASA JPL; Tom Steck of the Army Night Vision Lab; Keith Johnson of KRC; Ron Beck and Jeffrey Eidenshink of the EROS Data Center; Phil Teillet of Canada Centre for Remote Sensing; Mike Duggin of SUNY CESF; Jim Barrow of GDE; and Lenny Lafier and Mike Richardson of Eastman Kodak, my respect and thanks.

I also thank the many individuals who, through their careful reviews, so often pulled my foot out of my mouth by catching the numerous blunders in the early versions of the text. Their insight and guidance added greatly to the final version. Much of the credit for the clarity and coherence goes to these individuals. What errors remain rest with me, for this crew did a fine job on a thankless task. I salute Bernie Brower, John Mason, Rulon Simmons, Jon Arney, Mark Fairchild, Jim Jakobowski, John Francis, Bill Philpot, Roger Easton, Carl Salvaggio, Kurt Thome, Carolyn Kitchen, Perti Jarvelin, Zoran Ninkov, Harry Gross, Lee Sanders, Phil Teillet, Sylvia Shen, Bill Philpot, Rulon Simmons, Joe Sirianni, Julia Barsi, Brian Markham, Darrell Williams, Roger Easton, Emmett Ientilucci, John Kerekes, Don Light, Dave Messinger, Harvey Rhody, Mike Richardson, Tony Vodacek, Don McKeown, and Dennis Helder.

Finally, I'd like to thank RIT for supporting me in this effort, and the editorial staff at Oxford University Press for encouraging me to start this project.

CONTENTS

REMOTE SENSING

2ND EDITION

This page intentionally left blank

CHAPTER 1
INTRODUCTION

The standard opening for any reference book is to define the subject: Remote sensing is the field of study associated with extracting information about an object without coming into physical contact with it. Most readers, having read that vague definition, look out the nearest window and remotely sense the weather conditions to help them decide whether to head out into the sunshine or muddle forward through this introduction. You have just engaged in a chain of events that represents remote sensing in its broadest context. You used your eyes to acquire data and your visual cognitive system to process the data, and you output a decision to read on and/or return to this later in the day. As we will see, this sequence of acquisition, processing, and output is characteristic of remote sensing systems. In most cases, we will find that the general definition of remote sensing given above is too broad for a manageable treatment. As a result, we will spend most of this chapter limiting the field to that segment I intend to address and providing a perspective for the approach taken in the succeeding chapters.

1.1 WHAT IS REMOTE SENSING (AS FAR AS WE'RE CONCERNED)?

The broad definition of remote sensing would encompass vision, astronomy, space probes, most of medical imaging, nondestructive testing, sonar, observing the earth from a distance, as well as many other areas. For our purposes in this text, we are going to restrict our discussion to earth observation from overhead. Within this restricted context, we could trace the origins of the field to prehistoric hunters or explorers who climbed the nearest hill to get the lay of the land. However, our

main interest will be in the principles behind overhead earth observation using aircraft and satellite (aerospace) remote sensing systems. Given this restriction, we recognize that we are talking about a field of study that for the most part is only a few decades old. This is an important perspective for the reader to keep in mind. In such a relatively young field, many of the principles are still being formulated, and in many areas the consistent structure and terminology we expect in more mature fields may be lacking. On the positive side, young fields, such as remote sensing, offer a myriad of opportunities for exploring unanswered (and often as yet unasked) questions. These questions address what is to be learned about the earth's surface and about the earth's land, water, and atmosphere. They can be further extended to include the condition of the water quality, the vegetation health, the pollutant levels, and how these conditions are changing with time. The tools used in addressing these questions will be our major concern.

1.2 WHY REMOTE SENSING?

As we continue to try to narrow our definition of remote sensing (at least as far as what we will cover in this book), it is important to keep in mind why we use remote sensing in the first place. The reason most often cited is that remote sensing literally and figuratively gives us a different way of looking at the world. This different view often adds significant amounts of incremental data that are useful to a host of applications. Some of these incremental data are due to the synoptic perspective provided by overhead images. Many more traditional approaches to earth observation (i.e., surface studies) can be limited by too much detail on too few samples or by only having data from a very restricted locale. This is the classic "can't see the forest for the trees" problem. The *synoptic* perspective offered by remote sensing lets us look at whole forests, regions, continents, or even the world and yet, at appropriate scales, can let us see not only the whole forest but the individual trees as well (cf. Figs. 1.1 through 1.4). This perspective lets us look for large-scale patterns, trends, and interactions and serves as an excellent means of guiding efforts to interpolate or extrapolate parametric values from extensively studied ground sites (ground truth sites). Coupled to this synoptic perspective is the opportunity to view the world in ways our visual system cannot. Even if we ignore sound waves, magnetic fields, nuclear radiation, etc., and restrict ourselves to remote sensing of electromagnetic (EM) radiation (as we will in this treatment), there is a great deal of the EM spectrum our eyes can't see. If we look at a transmission spectrum of the earth's atmosphere (cf. Fig. 1.5), we see a number of transmission bands or windows through which we could peek if we could find appropriate sensing technology. As we will see in later chapters, sensing through these windows lets us look at vegetation stress, surface temperature, atmospheric water content, and a host of other parameters that our visual system couldn't begin to see.

Another view offered by remote sensing that is hard to achieve in other ways is a temporal perspective over large areas. This lets us look at changes over time spans of minutes to decades (cf. Figs. 1.6 and 1.7). Not only can this be used to study changes in the condition of the earth's surface and the atmosphere, but it can

Figure 1.1 Photographic image of the Earth acquired by Skylab astronauts. See color plate 1.1.

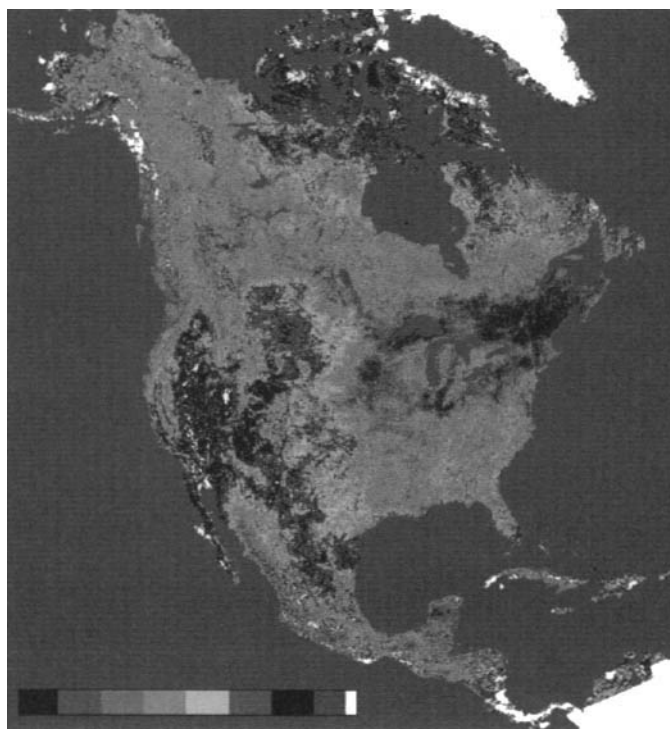


Figure 1.2 Map of the normalized difference vegetative index of North America derived from AVHRR data. See color plate 1.2.

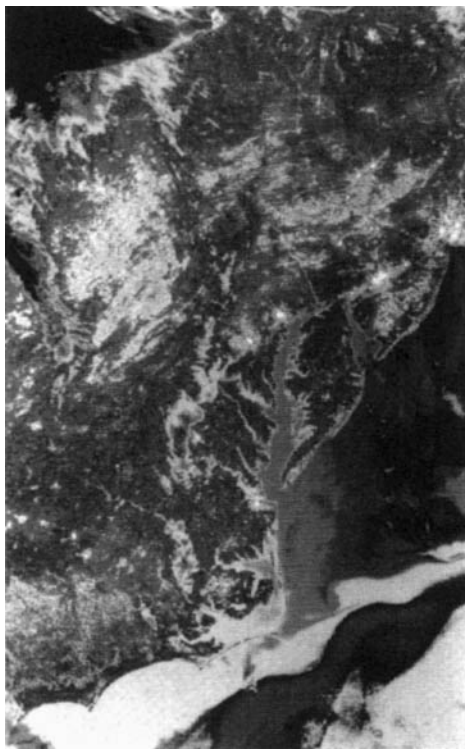


Figure 1.3 Image of eastern U.S. produced from the thermal channel of the Heat Capacity Mapping Mission (HCMM) radiometer. Note the warm urban areas and the warmer water in the Gulf Stream. See color plate 1.3.

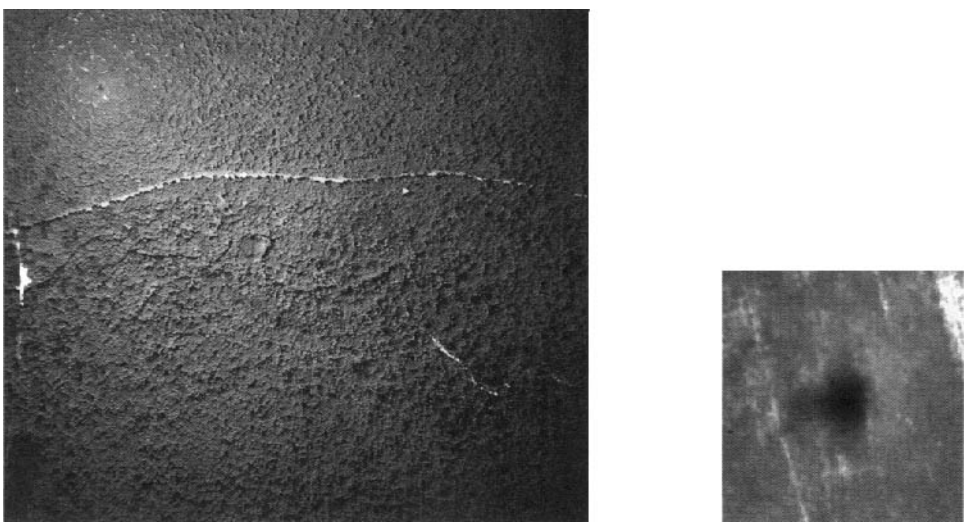


Figure 1.4 (a) Color-infrared aerial photograph of a forested area. Note the "hot spot" to the top left where the line of sight is along the sun-sensor-target line such that the scene elements obscure their own shadow. (b) A zoom of the center of the "hot spot" in a subsequent frame over open fields where the plane's shadow is more defined. See color plate 1.4.

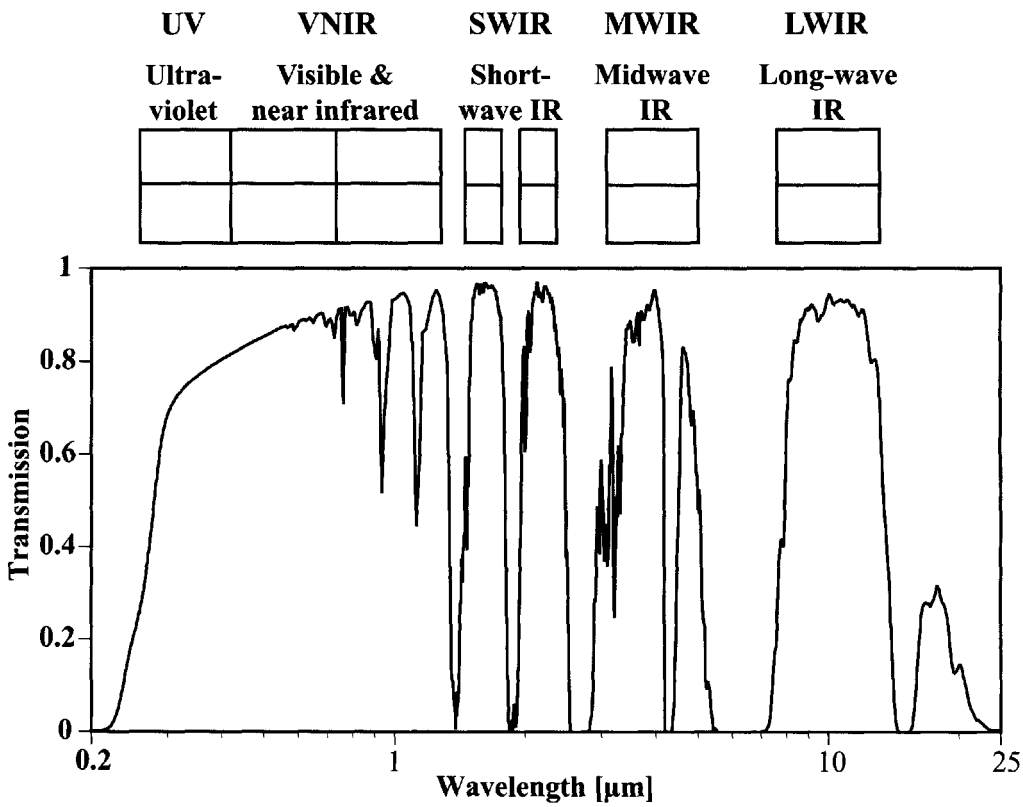


Figure 1.5 Atmospheric transmission spectra showing windows available for Earth observations.

also be used to effectively "see through clouds" in atmospheric windows where this would normally be impossible. This is accomplished by taking a sequence of images (e.g., on five consecutive days) and cutting and pasting together an image that represents a cloud-free composite, as shown in Figure 1.8.

1.3 WHAT KINDS OF REMOTE SENSING?

Remote sensing of EM radiation is often described in terms of what spectral windows one is peeking through. We are going to restrict ourselves primarily to remote sensing through the windows between 0.4 and 15 μm. Figure 1.9 shows how different the earth can appear when viewed simultaneously in four different spectral regions. As we will see, it is these differences that facilitate many of the information extraction techniques we will discuss in later chapters.

Much of remote sensing is done in the visible and near infrared (VNIR) in the daytime. However, nighttime imaging in the VNIR window can yield information on the location, extent, and activity level of population centers, as illustrated in Figure 1.10. By selecting a window in the thermal infrared region dominated by self-emission due to the temperature of objects, we can even "see" in the dark, making 24-hour remote sensing a possibility. The thermal infrared windows let us image a

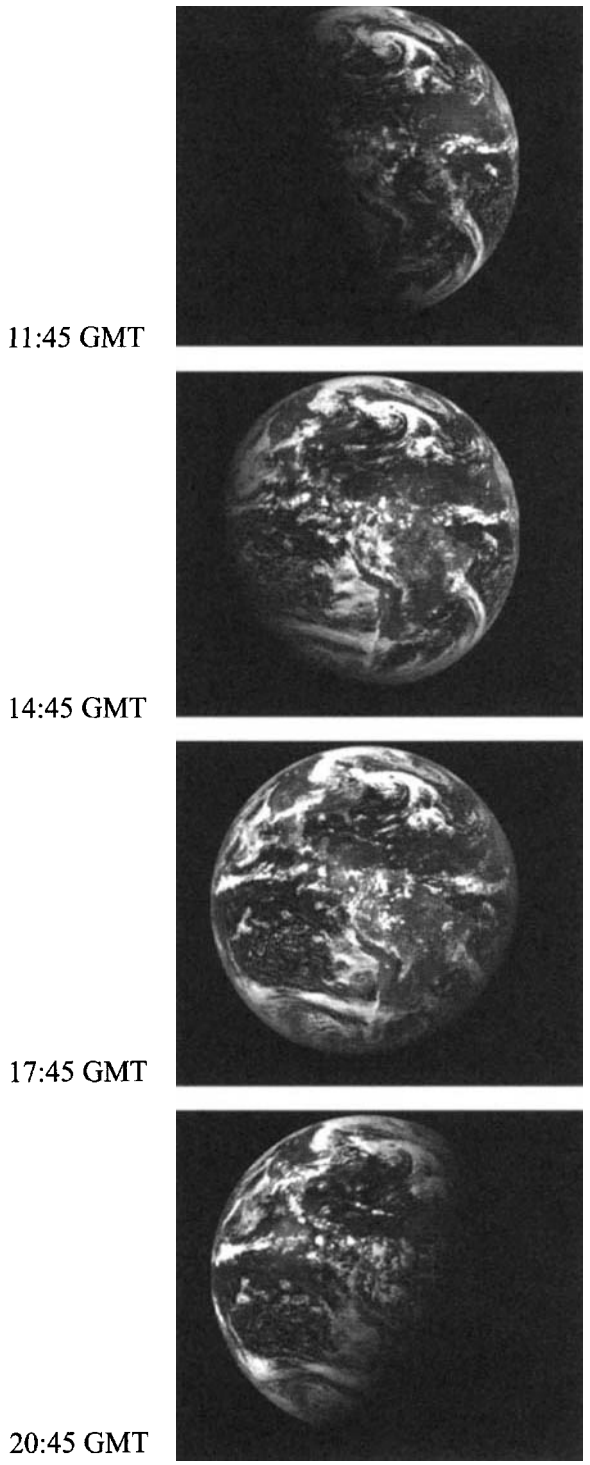


Figure 1.6 Time sequence of GOES images showing cloud dynamics over North and Central America. These images are acquired every 30 minutes. The images selected are more widely spaced over the course of the day to emphasize the movement of the clouds.

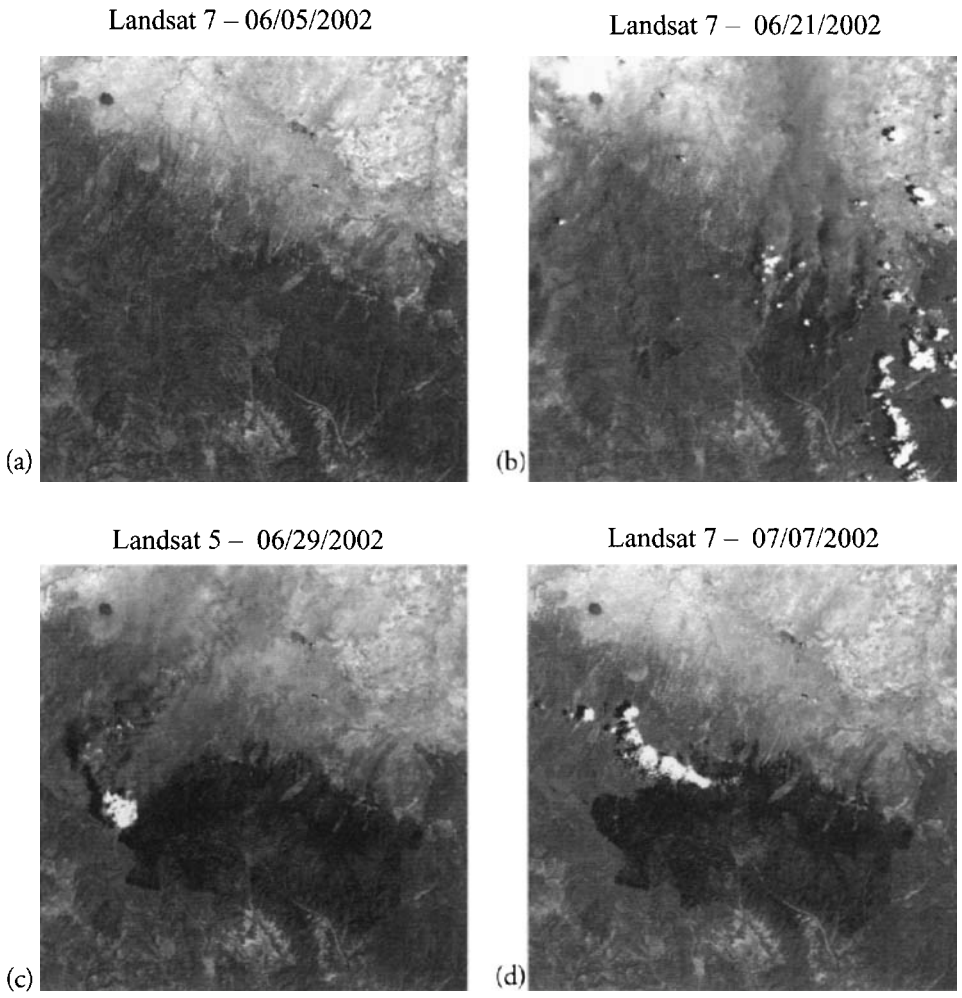


Figure 1.7 Landsat 5 and 7 change sequence showing (a) prefire condition, (b) the Rodeo fire in Arizona (burning for 3 days at this point) and the Chediski fire to the west (burning for 1 day), (c) the combined fires still burning and (d) the extent of the burn scar. The red region shows the burn scar in the band 7, 4, 2 band combination shown in Color Plate 1.7.

great deal more detail at night, ranging from terrain features to detailed inspection of facilities, forest fires, etc. (cf. Fig. 1.11). By choosing to restrict ourselves to passive remote sensing, we will limit our discussion to sensors that collect energy that is either emitted directly by the objects viewed (e.g., thermal self-emission) or reflected from natural sources (typically the sun). We are imposing this limit on the topics covered merely to reduce the volume of material.

There are many very useful active remote sensing systems whose treatment is simply beyond our scope. These active systems employ an active source that illuminates the scene. In some cases, the sensed energy is reflected or scattered from the source as in the case with the synthetic aperture radar (SAR) image shown in Figure 1.12. Radar images the reflected microwave energy emitted by the sensor itself. By selecting suitably long wavelengths, it is possible to use radar to image

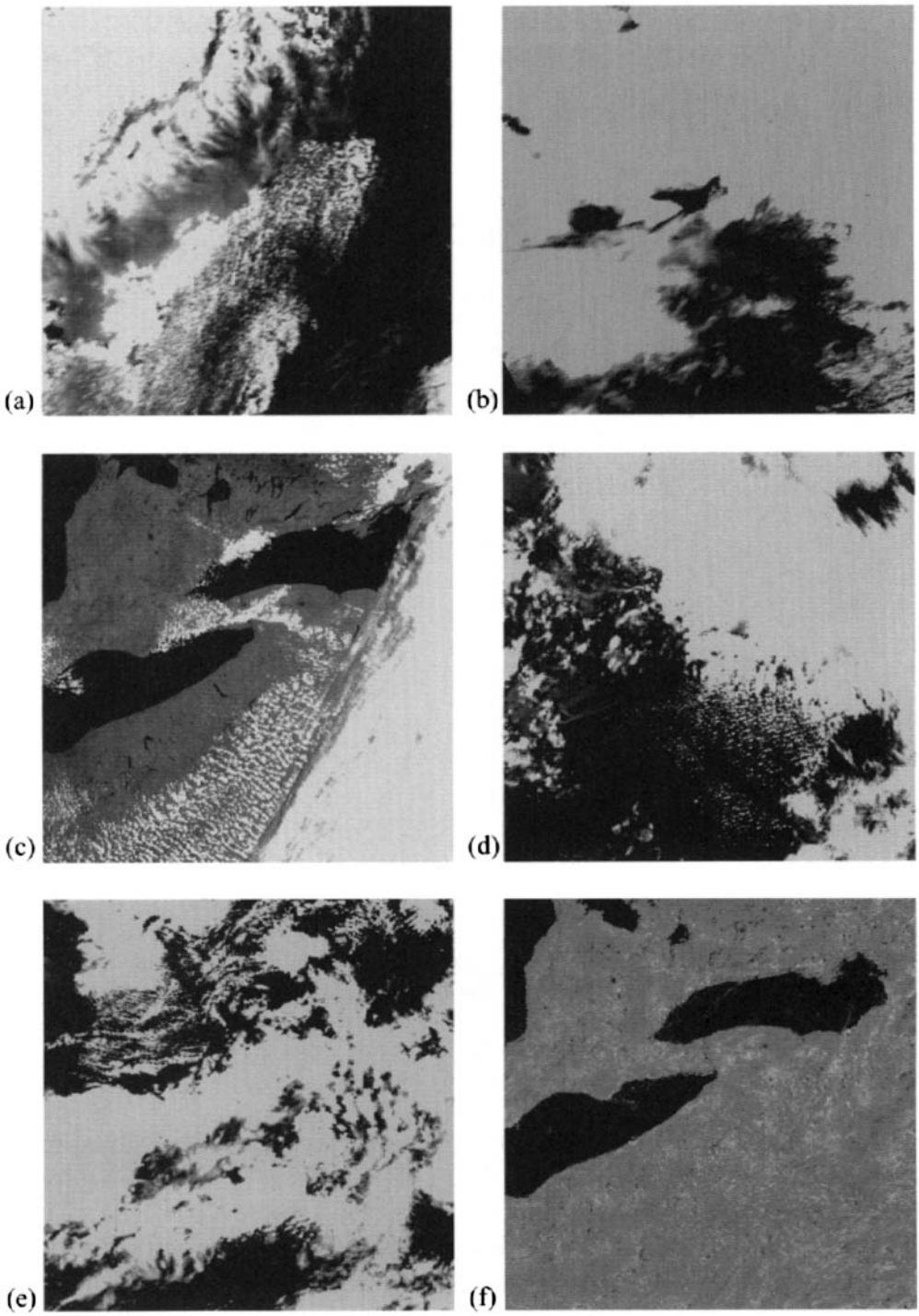


Figure 1.8 Five-day sequence of AVHRR images of the eastern Great Lakes and a cloud-free composite (f) made by combining the five-day sequence. See color plate 1.8.

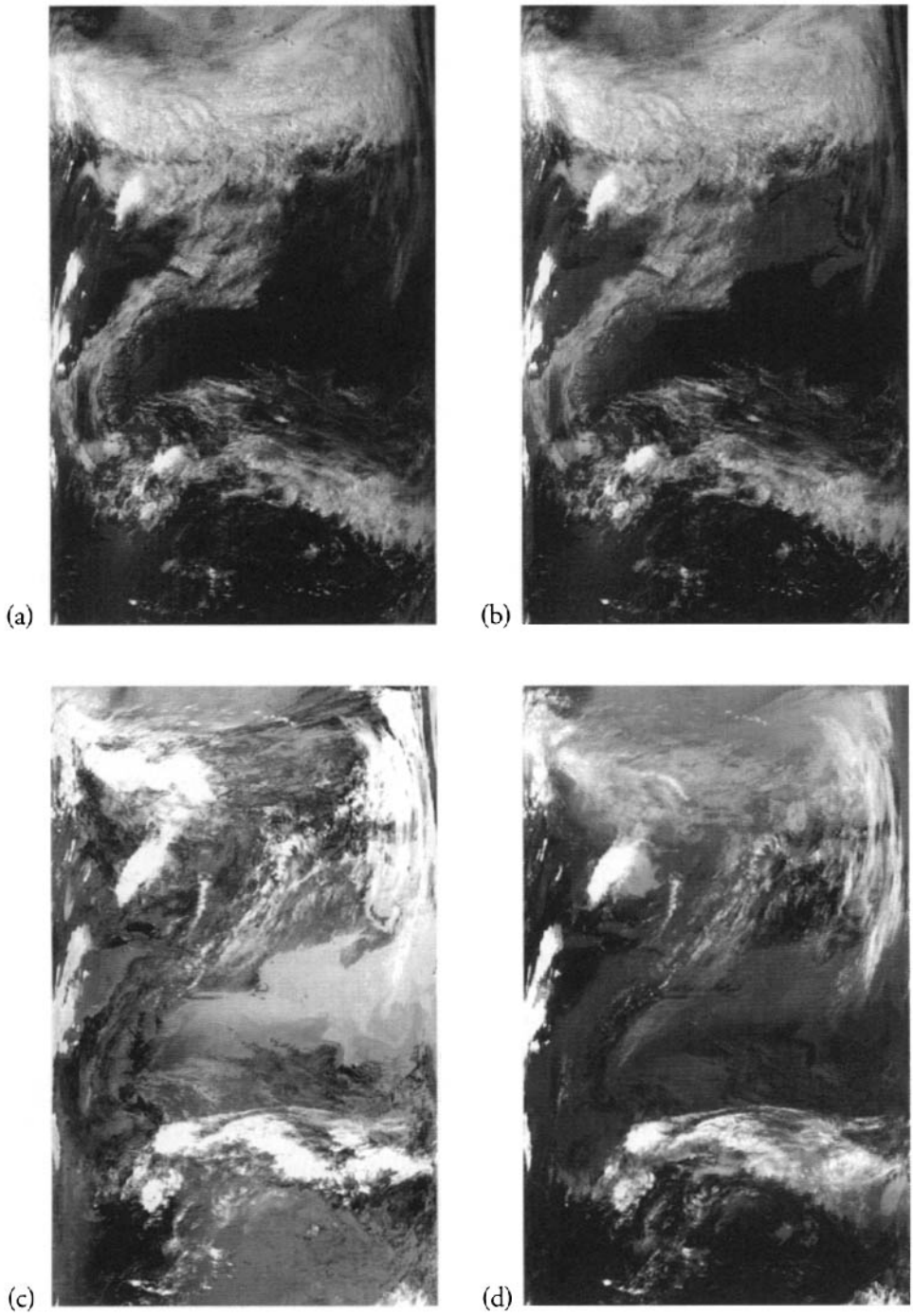


Figure 1.9 Simultaneously acquired AVHRR images showing the east coast of North America as viewed through (a) red, (b) near infrared, (c) midwave infrared, and (d) long-wave infrared windows. (Note that the MWIR and LWIR images are inverted, so white is cold to preserve the white shades for the clouds.)

Figure 1.10 Nighttime visible image of the U.S. Street lighting clearly demonstrates major population centers and traffic corridors, but no terrain or feature detail is available. This image is a composite made up from Defense Meteorological Satellite Program (DMSP) images.

through clouds, providing an all-weather capability not available with visible or IR systems.

In other cases, the active source energy is absorbed by the scene elements and reradiated in other wavelength regions. This is the case, for example, in laser-induced fluorescence imaging. The surface is irradiated by a laser that stimulates some of the materials to fluoresce. The fluorescent energy is emitted at longer wavelengths and sensed by the imaging system. The amount of fluorescence is a function of material type and condition. Many of the principles discussed in this volume are applicable to active remote sensing systems. However, covering the specifics of these systems would make this volume prohibitively long. The reader should consider Elachi (1987) and/or Henderson and Lewis (1998) for a thorough introductory treatment of radar principles and Measures (1984) for a treatment of active remote sensing with laser systems.

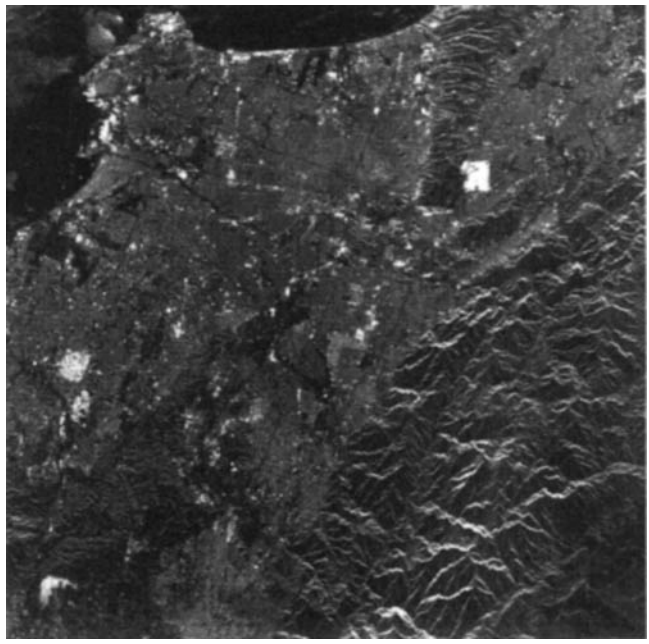
So far, we have restricted our scope to passive aerospace remote sensing of the earth with emphasis on the 0.4–15 μm region. Within that scope, we are going to place a strong emphasis on quantitative analysis, particularly digital image processing for target identification and radiometric analysis for condition assessment. The more classical analysis referred to as *photo* or *image interpretation* will be briefly discussed in Chapter 2. This is the process of extraction of information

(a)

(b)

Figure 1.11 Nighttime infrared images of (a) a portion of an HCMM image of the Appalachian Mountains in central Pennsylvania showing terrain features and (b) an airborne infrared image of a government office complex in Albany, NY, acquired in the winter as part of a heat loss study. Note in particular how the service floors in the office towers are bright (hot). The egg-shaped object includes a large auditorium whose heated shape shows through the roof structure.

(a)



(b)

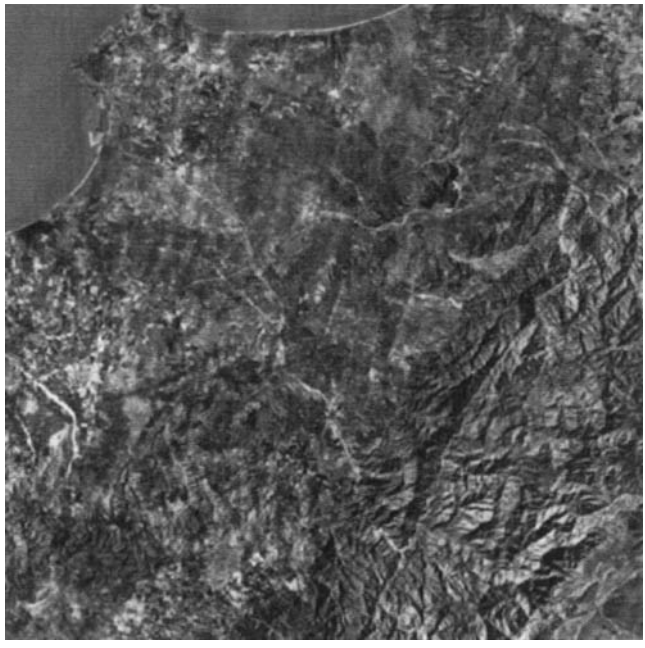


Figure 1.12 (a) Synthetic aperture radar (SAR) image of Los Angeles and (b) a portion of a Landsat image of the same region shown for reference.

by a human analyst viewing the photographic or electro-optical image to determine the location and condition of features. It is based largely on spatial patterns and is a critical part of most remote sensing activities. Expertise in this field is, however, largely a function of the application area of interest (e.g., geology, forestry). More detailed treatment of photo interpretation approaches and tools can be found in Lillesand and Chipman (2003), Philipson (1997), and Liang et al. (1951).

In contrast to photo interpretation, which relies heavily on the skills of an individual analyst, quantitative analysis attempts to develop analytical tools often based on brightness levels in one or more spectral bands. From an operational standpoint, a major goal of quantitative analysis is to reduce the burden on human analysts by performing tasks that are difficult or tedious for an image interpreter. This is largely accomplished using two approaches. The first approach uses quantitative measurement of radiance levels that are very difficult to differentiate visually. The second approach uses computer-based algorithms to combine many sources of data (e.g., spectral bands, texture metrics, spectral feature vectors, etc.) that can overwhelm a human analyst. From a scientific perspective, a major goal of quantitative analysis is to develop functional relationships between remotely sensed data and parameters of interest (e.g., land cover or material type, location, extent, concentration, orientation, condition, change in condition).

An additional advantage of quantitative analysis over image interpretation is the ability to perform error analysis. Functional relationships are quantitative such that errors can be assigned to estimates of parameter values and confidence levels to classification parameters.

Another often overlooked goal of quantitative analysis is to make the remote sensing data as useful as possible to the interpreter. This can take the form of improving the visual appearance of an image or a data set for human analysis or transforming a maze of quantitative analytical results to a simplified form (e.g., crop yield is expected to be X metric tons with an error of $\pm Y$ metric tons). From this standpoint, the whole field of scientific visualization becomes closely coupled to the output side of the remote sensing process.

In looking at quantitative analysis from the scientific perspective, it is easy to delineate some of the capabilities, as well as clear limitations of the remote sensing approach. In an idealized functional form, quantitative analysis would yield an objective, reproducible, quantitative, error-free functional relationship for each parameter of interest of the form

$$Y = f(X_1, X_2, \cdots X_N) \tag{1.1}$$

where Y is one of several parameters of interest (e.g., material type, concentration of pollutant), and $X_1, X_2 ... X_N$ are the remotely sensed independent variables needed to characterize the Y parameter(s). These variables ($X_1, X_2 ... X_N$) would ideally be in the form of reproducible material characteristics such as spectral reflectance, temperature, orientation, emissivity, scattering function, or polarization response. However, in many cases the variables might be expressed in terms of observed values such as digital counts in each spectral band in an image. The tools used in trying to define the inputs to Eq. (1.1) (i.e., X_1 values) and the nature and form

of the functional relationships (f) with the parameters (Y) of interest will occupy most of the following chapters. However, the critical assumption of remote sensing (for both image interpretation and quantitative analysis) should be addressed here, that is, that a relationship exists between the parameter of interest and some combination of EM signals from the earth or its atmosphere. For example, the concentration of suspended solids in surface waters normally impacts the reflectivity of the water. Therefore, we could hope to use remote sensing to observe and maybe quantify the concentration of suspended solids. On the other hand, the concentration of D_2O (heavy water) in a body of water has no obvious EM manifestations and could not be observed using the remote sensing methods we are considering. There is no functional relationship of the form of Eq. (1.1), and EM remote sensing is not an available option for addressing this latter problem. On the other hand, there are times when a parameter may be indirectly measured, or at least inferred. For example, if a factory is releasing a colorless chemical into the surface waters, there may be no direct functional relationship of the form of Eq. (1.1). On the other hand, if the chemical is toxic to phytoplankton (algae), which do have an optical manifestation, then it may be possible to infer the presence and even the concentration of the chemical (though other means would clearly be required for corroboration). Part of the remote sensing process is the search (both theoretical and experimental) for these EM manifestations and the definition, development, and application of tools to observe and measure them.

In considering whether a material or condition will have an observable optical manifestation, it is useful to look at the properties of matter that control the optical interactions. A full treatment of this topic is well beyond our scope and includes topics from quantum mechanics, physical chemistry, and materials science, as well as treatment of the physical and geometric optics that incorporate the surface, bulk, and geometric structure of materials. We will limit ourselves to a brief consideration of the *spectral absorption* properties of materials. A more complete treatment of this topic can be found in Pieters and Englert (1993).

Recall from elementary quantum theory that photon absorption by a material is a discrete process that typically takes place when the photon promotes an electron to a higher energy level (electron transition), induces an energy change in the molecular binding (rotational and vibration transitions), or a combination of these (overtones). In general, the electron transitions are rather energetic and require high-energy (often ultraviolet) photons. Furthermore, recall that in solids the electronic orbitals are highly perturbed (dispersed) by the binding forces. As a result, absorption features are highly broadened from the single absorption line feature we would think of in a simple atom in gaseous form (Bohr atom). This leads to one of the first fundamental distinctions in spectral features. Gases tend to have very narrow absorption features (cf. Fig. 1.13), and solids and liquids tend to have much broader spectral features. The very broad electron transitions in the UV lead to the trend in many solids to have decreasing reflectance as the wavelength decreases [cf. Fig. 1.13(a)]. At longer wavelengths (NIR-SWIR), the bending and stretching overtones introduce many absorption features. This continues with overtone and fundamental vibrational and rotational absorption features in the MWIR and LWIR.

Figure 1.13(a) shows that rocks and minerals will have potentially differential reflectance features across a broad spectral range, with the SWIR being particularly interesting. Figure 1.13(c) shows that while different gases have features across the full spectrum, the "fingerprint" region for gases occurs at longer wavelengths, particularly in the LWIR (cf. Sec. 3.4 for a more complete treatment of atmospheric gases). Vegetation as shown in Figure 1.13(b) has a wealth of features in the VNIR that may be used both to identify and characterize its condition. Water is one of the few materials regularly seen in liquid form. Figure 1.13(d) shows that most of the variability in water is in the visible and just into the NIR because the strong absorption at longer wavelengths masks variability due to the constituents in the water.

This brief treatment is intended only to point out some of the most fundamental optical properties of material and how the variability between and within a material type may manifest to remote sensing systems.

Our emphasis here will be on characterizing the properties of materials that affect the EM radiation emitted, reflected, or scattered, the processes that govern

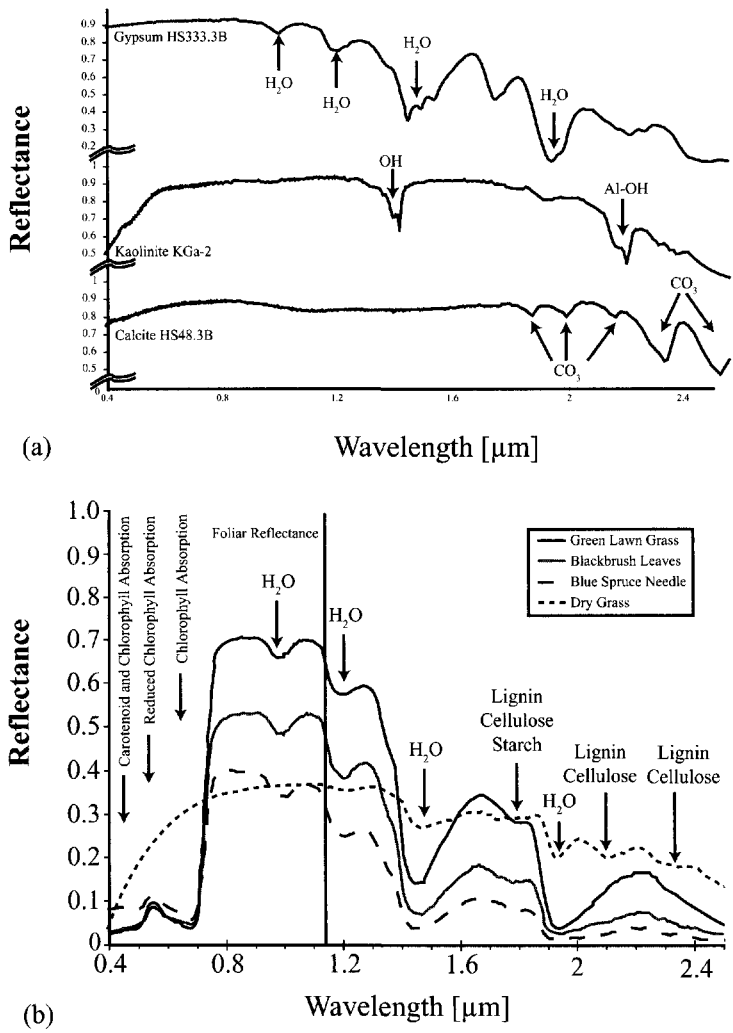


Figure 1.13 Spectral proterties of materials.

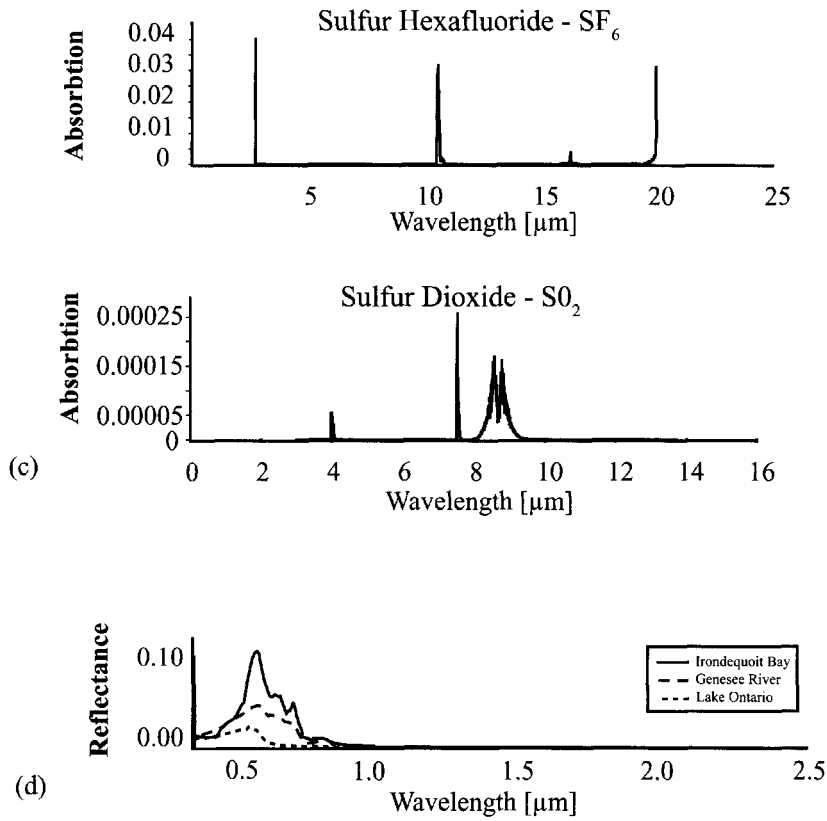


Figure 1.13 continued. Spectral proterties of materials.

the propagation of EM radiation to the sensing platform, and the types of sensors that can be used to collect and record the relevant EM energy levels. Then we will look at tools that can be used to derive the inputs to functions of the form of Eq. (1.1) and methods to improve the quality of both the analytical and visual data available to image analysts. This emphasis on theoretical principles and analytical tools stems from the assumption that, in most cases, the remote sensing scientist will be working closely with experts trained in the application areas of interest. In such teams, the remote sensing scientist must learn enough of each application area to understand how remote sensing can be of use and must teach the corresponding application experts enough remote sensing fundamentals so they can help define the problems and expectations in a common format. Our emphasis here is on the tools the remote sensing scientist needs and the principles with which the applications expert needs to be familiar. In the context used throughout this treatment, a tool is an instrument, algorithm, or often a combination of instruments and algorithms used to construct a solution to a problem. Tools in this context are most often thought of as general and reusable (e.g., an algorithm for atmospheric compensation or edge detection). The tools would then be used to help build a solution to an application-specific problem. Readers uncomfortable with their backgrounds in the relevant physics related to electromagnetic radiation or basic optics should consult a university physics text such as Halliday et al. (2005). For the more ambi-

tious reader, a complete treatment of some of the optical principles drawn on here is contained in Hecht (2002).

Applications of remote sensing will be treated only to the extent that they illustrate how certain analytical tools can be used. With remote sensing extensively used in fields including meteorology, oceanography, forestry, agriculture, archeology, reconnaissance, geology, geography, range management, hydrology, and atmospheric and soils science, we could not begin to adequately address the domain-specific remote sensing issues. Many of these applications areas are addressed in the *Manual of Remote Sensing* [cf. Estes and Thorley (1983), Rencz (1998), and Ustin (2004)], as well as in numerous remote sensing texts that approach the field from the applications perspective, e.g., Barrett and Curtis (1976) and Jensen (2000).

1.4 THE IMAGE CHAIN APPROACH

In order to take advantage of the synoptic perspective that remote sensing offers, we will emphasize imaging systems throughout our treatment. An *imaging system* is a system where the input is a scene and the output is an image. A simple imaging system would involve image capture, image processing, and image display. This is not intended to exclude point radiometers (e.g., those used in atmospheric sounders) or other nonimaging devices, but merely to indicate that most of our treatment of remote sensing will be interwoven with elements of imaging science. We will often find it convenient to think of the remote sensing process as a chain of events or steps that lead to a final output. This output will often be an image, but it might as easily be a map or a table of figures or a recommendation to a decision maker (e.g., harvest the trees in an infested forest before the mortality rates increase the fire hazard and reduce the timber value). From a conceptual or philosophic point of view, we will refer to the study or characterization of the sequence of steps or chain of events that lead to that final output as the *image chain approach.*

The events along the way may be very different and apparently unrelated. For example, some events will be natural processes, such as sunlight striking the ground. Some will be the result of simple choices or procedures, such as what type of sensor to use or what time of day (TOD) to acquire the image. Some of the events may be very intense activities directed at the imaging process, such as calibration of the image data or digital image processing. Finally, some events involve the inclusion of external data (e.g., ground truth, previous maps of the area, or inclusion of an expert's knowledge of a target). All of these events can be linked together by the fact that they impact the final output.

This perspective of viewing remote sensing as a series of events is called the image chain approach. Each event or step can be thought of as a link in the chain. The image chain approach is based on the premise that if we study and understand the chain of events associated with a particular output image or product, we will be better able to understand what we have (i.e., what the product tells us or means), what we don't have (i.e., the limitations or confidence levels associated with the output product), and where those limitations were introduced (i.e., where are the

weak links in the image chain). The image chain approach is often useful when we are trying to understand the fidelity of the image or information output from the remote sensing process or designing remote sensing systems, or procedures to ensure fidelity. Recognizing that each link in the chain results from the relationship to previous links, we begin to recognize that the chain, and hence the image fidelity (or output product), is only as strong as the weakest link.

The image chain approach treats the entire remote sensing process as an interrelated system. It generally includes an imaging system, and the components or transfer relationships in the imaging system will be referred to as links in the imaging chain. The *image chain* combines the links in the imaging chain with any nonimaging components or transfer relationships. The image chain describes the components and transfer mechanisms that characterize the entire remote sensing process as a system. The systems nature of the image chain approach lets us also think about the image chain as being composed of many interwoven strands. These subchains or subsystems, which we will often refer to as the *strands* of the image chain, are often separated out and treated independently for convenience and then recombined to form the complete image chain.

Our approach will be to try to define (in most cases quantitatively) the links in the image chain. This approach serves several purposes. First, and most fundamentally, by fully characterizing the image chain from end to end, we ensure an understanding of the process. Besides the academic value and the warm fuzzy feeling associated with understanding the process, the image chain approach lets us analyze the process and address questions related to what we have, what information we can extract from the data, and (often as helpful) what information is not available from the data. The image chain may be modeled quantitatively. Such a model can then be used as the basis for reverse or systems engineering to help in analyzing image data. If we know all the links along the chain, then it is often possible to work backwards to extract critical information from a particular link (e.g., backing out atmospheric effects to compute surface reflectance values). A quantitative analysis of the image chain also helps the analyst identify weak links in the system. This points to where one can most effectively expend efforts to improve the chain. This might take the form of drawing on other data to reinforce a weak link or identifying where one might want to try estimation, modeling, or iterative techniques to strengthen a link in the chain. The image chain approach can be particularly useful in many aspects of image analysis because it points to where a process can be improved. This is particularly useful for system design and upgrade studies, but can also be useful as pointed out above in focusing our image analysis efforts. By focusing our attention on the weak links, we can often achieve significant end-to-end improvements in the overall system. On the other hand, as the chain analogy suggests, we could spend a great deal of time, money, and effort trying to improve the system by working on a link that was already strong and achieve little or no net improvement in the overall system because we had ignored weaker links (cf. Fig. 1.14).

Throughout this treatment we will be concerned with how the image chain impacts image fidelity. We have carefully chosen to use the term *fidelity* rather than *quality*. Image quality is usually associated with the visual appearance of images

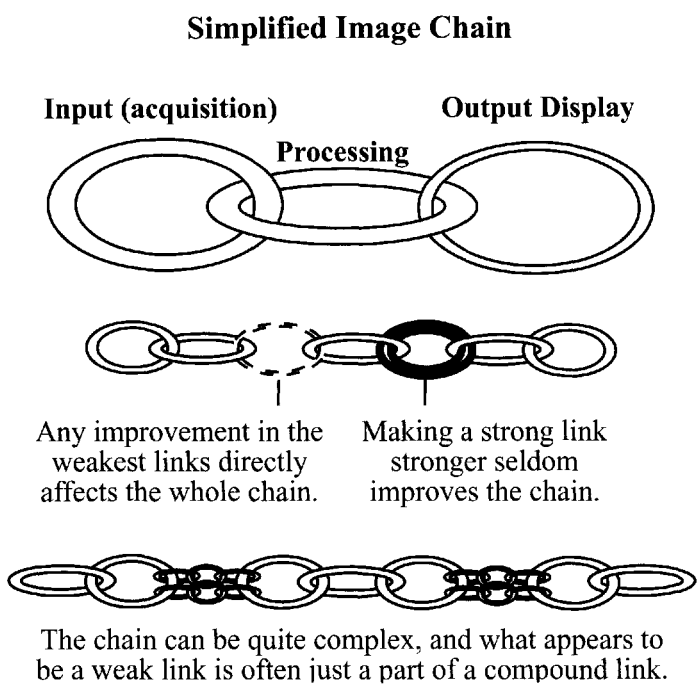


Figure 1.14 Image chain analogy. The chain is only as strong as its weakest link.

and, as a result, will be greatly influenced by visual and psychophysical response functions. Our concern here emphasizes ways to make measurements about the earth using remotely sensed data. As a result, our concern focuses on how well the remote sensing system reproduces the characteristics of interest. We will use the term *fidelity* to refer to how well the image chain reproduces these characteristics.

In general, there is good correlation between fidelity and perceived image quality. An image with high fidelity will, in general, be perceived as being of high quality. However, it is not uncommon for images with poorer fidelity to be perceived to be of higher quality. This is because the visual processing system applies different criteria (or at least different weights to similar criteria) in evaluating quality (cf. Chap. 13). By emphasizing the importance of fidelity, we seek to ensure the integrity of the final image for measurement purposes. In addition, the perceived image quality will usually be maintained. However, if a system is being designed solely or predominantly for visual image quality, the fidelity criteria emphasized here should be adjusted for their importance to perceived image quality.

The image chain approach can often be segmented into different aspects or parameters for convenience. For example, we sometimes will find it convenient just to look at certain parameters of a system such as spatial resolution, radiometric fidelity (including signal-to-noise issues), spectral resolution, or temporal fidelity. We can think of each of these parameters as having its own image chain, with the appropriate characteristics of the final output (e.g., spatial image fidelity) a function of the interplay along the links in the chain. We must, of course, keep in mind that these various chains are seldom independent of each other, and the final system

is the result of the interplay and collective strength of all of the strands in the chain (cf. Fig. 1.15).

In most cases, we will look separately at different strands in the image chain (e.g., spatial resolution) and track some measure of fidelity through the links in the chain. The spatial fidelity of the image will be measured based on how well the image reproduces the radiance variations in a scene. As discussed in Chapter 13, we will use the *modulation transfer function* (*MTF*) to measure spatial image fidelity, with an ideal system having an MTF of one at all possible spatial frequencies. The radiometric fidelity is usually characterized in two ways. The first method uses the *noise* level defined as the signal variation about what should be a constant level (i.e., when the sensor is sensing a uniform surface). An ideal system would have zero noise at all spatial frequencies. The second is the difference between a measured radiometric value (e.g., radiance or reflectance) and the actual value. An ideal system would be perfectly calibrated such that there is no error between measured and actual values. These and other measures of fidelity will be introduced in the following chapters. In Chapter 13 we will look at how these measures can be used to evaluate the performance of individual links in the image chain, as well as end-to-end performance.

We will often use a "systems theory" approach to modeling the imaging chain. This approach is treated in considerable detail in Gaskill (1978) and Dainty and Shaw (1974). We will, for the most part, make a number of simplifying assumptions. In particular, we will usually assume that the systems are linear and shift (space) invariant. The linearity assumption requires that a change in an input variable will result in a proportional change in an output parameter. The space invariance assumption requires that an input at any spatial location will produce the same output function (e.g., the image of a point source will produce a bright-

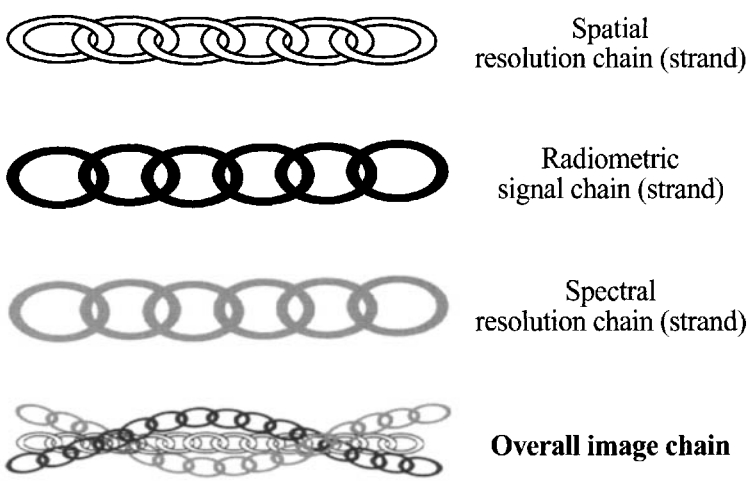


Figure 1.15 Aspects of the image chain. The overall remote sensing system can often be divided into several chains for convenience, but the final system performance is a function of the interplay of all the chains.

ness distribution with the same shape whether imaged at point A or point B). Most systems violate our simplifying assumptions to some extent, and in some cases a more rigorous treatment will be required. However, experience has shown that most reasonably well-behaved imaging systems can be approximated with good success using the linear-space invariant approximations. Thus, unless otherwise noted, we will assume that a linear space invariant system model is being assumed for our imaging chain.

While we don't believe it is useful to push any analogy too far, we think the reader may find the image chain a useful and rewarding perspective from which to study remote sensing. As a result, this book is organized around the image chain approach. Chapter 2 contains mostly historical and conceptual background material and terminology. Chapters 3 and 4 treat the early links in the image chain by tracing EM flux from its origins to the sensor. The emphasis in these chapters is on the strand of the image chain associated with the radiometric signal. Chapters 5 and 6 deal with spatial, spectral, and radiometric links in the image chain associated with imaging sensors. Chapter 7 deals with reverse engineering back along the radiometric links in the image chain from the sensor to the ground to compute target-specific values such as reflectance and temperature. The links in the chain associated with monochrome, multispectral, spectroscopic, and model-based image processing are covered in Chapters 8–11, respectively. By the end of Chapter 12 we've introduced the basic principles associated with the entire image chain as it applies to remote sensing systems, including image fusion, data fusion in Geographic Information Systems (GIS), and output products. In Chapter 13, we treat various aspects of the image chain to see how we could identify and potentially correct limitations in the chain. Finally, in Chapter 14 we discuss how synthetic image generation can be used in conjunction with the image chain approach as a tool to understand and analyze remotely sensed images and imaging systems.

1.5 REFERENCES

Barrett, E.C., & Curtis, L.F. (1976). *Introduction to Environmental Remote Sensing.* Chapman & Hall, London.

Dainty, J.C., & Shaw, R. (1974). *Image Science.* Academic, NY.

Elachi, C. (1987). *Introduction to the Physics and Techniques of Remote Sensing.* Wiley, NY.

Estes, J.E., & Thorley, G.A., eds. (1983). *Interpretation and Applications.* In Vol. 2 of *Manual of Remote Sensing*, R.N. Colwell, ed. American Society of Photogrammetry, Falls Church, VA.

Gaskill, J.D. (1978). *Linear Systems, Fourier Transforms, and Optics.* Wiley, NY.

Halliday, D., Resnick, R., & Walker, J. (2005). *Fundamentals of Physics.* 7th. ed., Wiley, NY.

Hecht, E. (2002). *Optics.* 4th. ed., Addison-Wesley, Reading, MA.

Henderson, F.M., & Lewis, A.J., ed. (1998). *Manual of Remote Sensing, Vol. 2, Principles and Applications of Imaging Radar.* Wiley, NY.

Jensen, J.R. (2000). *Remote Sensing of the Environment: An Earth Resources Perspective.* Prentice Hall, NJ.

Liang, T., Costello, R.B., Fallon, G.J., Hodge, R.J., Ladenheim, H.C., Lueder, D.R., & Mallard, J.D. (1951). Airphoto Analysis: Cornell University Land Form Series. Vols. 1–6, NRL 257-001. Washington DC.

Lillesand, T.M., & Chipman, J. W., (2003). *Remote Sensing and Image Interpretation.* Wiley, NY.

Measures, R.M. (1984). *Laser Remote Sensing: Fundamentals and Applications.* Wiley, NY.

Philipson, W.R., ed. (1997). *Manual of Photographic Interpretation.* American Society of Photogrammetry and Remote Sensing, Bethesda, MD.

Pieters, C.M., & Englert, P.A.J. (1993). *Remote Geochemical Analysis Elemental and Mineralogical Composition.* Cambridge University Press, Cambridge.

Rencz, A., ed. (1998). *Manual of Remote Sensing, Vol. 3, Remote Sensing for the Earth Sciences.* Wiley, N.Y.

Ustin, S., ed. (2004). *Manual of Remote Sensing*,3rd Ed., Vol. 4, Remote Sensing for Natural Resource Management and Environmental Monitoring, Wiley, NY.

CHAPTER 2
HISTORICAL PERSPECTIVE AND PHOTO MENSURATION

Observing the earth and recording those observations as images is a form of remote sensing that has been taking place for nearly a century and a half. The development of the first commonly used form of photography is often attributed to L. J. M. Daguerre, who from 1835 to 1839 perfected the daguerreotype. The first known photographs from an overhead platform were taken about 20 years later by a French portrait photographer, Gaspard Felix Tournachon (a.k.a. Nadar). In 1858 Nadar took photographs from a balloon equipped with a darkroom to process the wet plates used in the collodion process he employed. In addition to their novelty value, Nadar's aerial photographs soon found use in surveying and map making [cf. Stroebel and Zakia (1993)]. For more than 100 years, camera photography remained the only tool for remote sensing from overhead platforms. The platforms changed to include kites, pigeons, and airplanes. Eventually, 101 years after Nadar's first flight, the first photographs of earth were made from space. These were acquired in August 1959 by Explorer 6 in a nonorbital space flight. The cameras and photographic processes also improved considerably over that period to include the first color photographs from an unmanned Mercury-Atlas flight (MA-4) in 1960.

The modern era of remote sensing began that same year, 1960, with the launch of the Television Infrared Observation Satellite (TIROS-1). This satellite and its successors carried vidicon cameras into space to begin the systematic process of monitoring our weather and global environment from space [cf. Simonett (1983)]. The modern era is characterized by the availability of satellite platforms, electro-optical (EO) sensor systems, and quantitative analytical tools for processing both

photographic and electro-optical images. The major thrust of the following chapters will be on sensors and analytical methods that relate to this more recent era of quantitative radiometric photographic and electro-optical imaging, for both aircraft and satellite systems. In this chapter, we want to lay the foundation for the rest of the book by looking at how this new era of remote sensing evolved, and then by delving into some aspects of the field that are supportive of our major thrust. Section 2.1 deals with photo interpretation and Section 2.2 with *photogrammetry* (the science of taking spatial measurements from photographs). Both topics are critical to understanding the origins of contemporary remote sensing. The interested reader should consider Estes and Thorley (1983), Wolf and DeWitt (2000), and Slama (1980) for a more complete technical treatment of these topics.

2.1 PHOTO INTERPRETATION

Since its inception, overhead photography has been (to use today's vernacular) a dual-use technology (i.e., it has commercial and military applications). Over the decades following Nadar's 1858 images, advances in the civilian use of aerial photography in the mapping and surveying sciences were adopted by the military. Similarly, technological advances in cameras and film for reconnaissance were rapidly employed in surveying, land cover mapping, and resource assessment. The field of *photo interpretation* evolved from the joint civilian and military need to learn how to extract more information from aerial images, particularly black-and-white images. In the early days, black-and-white images were all that was available. As early color film became readily available, its spatial resolution, so critical to visual analysis, could not match the high resolution of the black-and-white films preferred by photo interpreters for most applications.

Estes et al. (1983) describe visual cues that have been identified as part of the psychovisual trigger mechanisms used by photo interpreters in extracting information from overhead images. These cues are interesting to consider since visual analysis of overhead images is still a major aspect of essentially all remote sensing applications. Furthermore, as we look at the more quantitative, machine-oriented methods of analysis discussed in later chapters, it is important to recognize what humans can do well and how they do it. This may, in some cases, help us to mimic what our visual system does with machine-based methods. It is also important to recognize what the human visual and cognitive system does not do well to ensure that, where possible, we develop machine-based methods to cover these limitations.

Figure 2.1 lists some of the cues that Estes et al. (1983) have identified as useful in photo interpretation, along with visual examples of each cue. As we review these cues, it is useful to imagine how you would use them individually or in combination to address a particular problem. Recognize that the utility of a cue depends on the area of application and the analyst's level of knowledge regarding the application area. It is also useful to think of how one might teach a machine (computer) to reach the same conclusion as a human observer. We will begin by

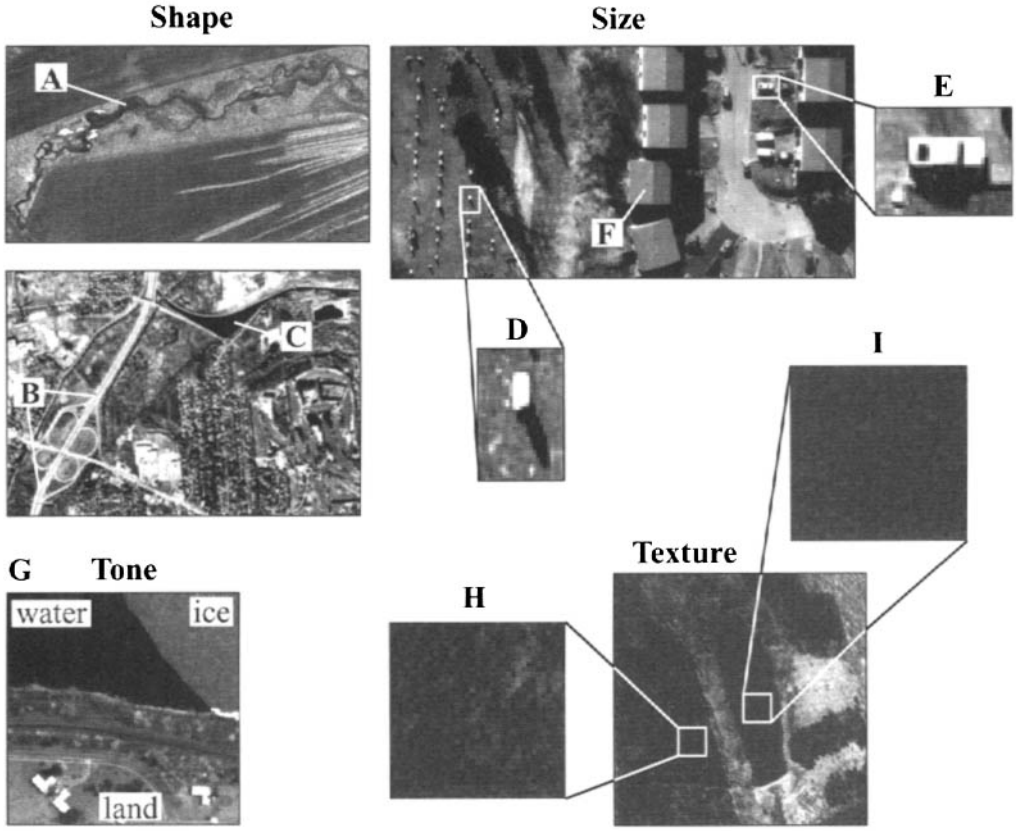


Figure 2.1 Cues used in photo interpretation.

briefly looking at the visual cues listed in Figure 2.1, recognizing that these are not exhaustive but merely indicative of keys that might be used by a photo interpreter.

Shape. Shape is simply the geometric outline of an object. Often this simple outline carries a surprising amount of information about the nature or function of the object. Even with no additional information, we can tell that objects B and C are manmade, and A is unlikely to be manmade. Thus, we see how shape can provide important information, even with extremely little detail in the image.

Size. Size may refer to the area of an object or to a single dimension such as the length of a road or airport runway. As soon as we see an image, our eye-brain cognitive system automatically assigns a scale. This is not a conscience logical process, nor is the scale factor quantitative. We simply recognize one or two objects (e.g., car, house, lake, continent), and by knowing their size, we proportionately dimension other objects in the image and recognize the extent or coverage of the entire image. In most cases of earth observation, this is an easy process because easily recognized objects are readily found. However, its importance is perhaps best recognized by examining an image of the moon or some landscape with which you are completely unfamiliar. In the example, the same squarish shape shown at D could be a tombstone, a car (E), or a house (F).

Tone. Tone is the brightness level in a monochrome image or the combination of brightnesses (i.e., color), in a color image. In the image at G, tone easily lets us differentiate water from ice. The visual system cannot easily quantify brightness, but as we discuss in Section 12.1, it relies heavily on tonal differences as cues. In multispectral images (color), the human's ability to quantify becomes even more restricted compared to the potential information content. This is one area where machines are much more adept than humans and is the basis for much of the multispectral scene classification described in Section 9.2.

Texture. Texture describes the structure of the variation in brightness within an object. Forest canopies and water in certain spectral bands have the same mean brightness but very different texture, which allows them to be easily distinguished (e.g., H is forest, I is water).

Pattern. Patterns are shapes with identifiable geometric or periodic attributes. The patterns associated with a drive-in movie theater and a baseball diamond are shown as J and K. These are manmade patterns. However, both natural and manmade patterns are common and exist at all scales (e.g., corn rows and drainage patterns). It is important to realize that the process of recognizing the handful of edges associated with objects J and K as a drive-in and a ball diamond involves processing a great deal of cultural knowledge about your environment. This same recognition process would be much more difficult for someone who had never seen these structures before. Thus, the extraction of information from pattern data requires some form of prior learning process. This turns out to be true for both human and machine observers.

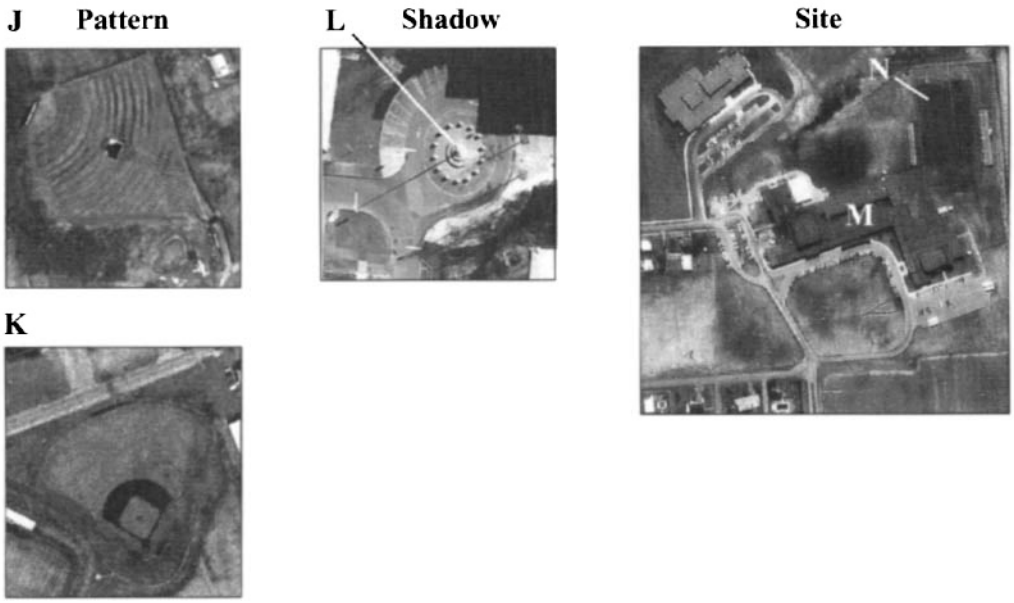


Figure 2.1 continued. Cues used in photo interpretation.

Shadow. Shadows are often thought of in a negative sense by interpreters because it can be difficult to see objects in deep shadow. On the other hand, shadows can provide insight into the height profile of objects. The flagpole is much easier to identify by its shadow at L than in the actual image.

Site. Site refers to the geographic location of a target or the location of one feature relative to another. Some authors differentiate these concepts with the geographic location being referred to as *site* and the relative location of features as *association.* Site information can help us determine that trees are more likely to be conifers because we know we are well up on a mountainside. It can also help us determine that if the building at M is co-located with the pattern at N, then it is probably a high school. Use of site as a cue requires a great deal of location-specific geographic knowledge and/or cultural knowledge. For the building at M to become a high school, we first recognize it as a building; second, we see it co-located with a pattern we recognize as a playing field (football) surrounded by an oval that we associate with a track. This combination of interrelated features, coupled with our cultural and geographic knowledge of construction patterns in North America, lets us conclude that the building is a high school. Implicit in this conclusion is a series of hypothesis tests, additional search procedures, and cues of which we may not even consciously be aware. For example, seeing the playing field, we look for indications that this is a sports complex (i.e., stadium-style seating and large parking lots). We reject this hypothesis and suggest a school. Grammar schools are ruled out since they normally don't have football fields and tracks. A college is a possibility but would more likely be associated with a cluster of buildings. We reject that hypothesis as unlikely and conclude that there is a high probability it is a high school.

The ability of the human analyst to combine the types of cues listed above with a host of other data acquired over a lifetime makes human analysis a critical component in most remote sensing tasks. We currently have no machine-based equivalent with anywhere near the capabilities necessary to allow us to off-load this task to machines. On the other hand, human analysts require training and experience before they become proficient in many aspects of image analysis. They are also expensive, are subject to fatigue, and have a difficult time when quantitative analysis is required or where many image inputs must be combined simultaneously. As a result, we have moved into an age of machine-assisted image analysis where we are attempting to move more of the burden of image analysis to machines. This lets human analysts do only those tasks that the machines cannot.

Our ability to analyze and reason based on spatial patterns was what drove most of early remote sensing. The human visual-cognitive system has evolved to perform this task. Applying it to overhead images is a fairly straightforward transition requiring only adaptation for the perspective change. Indeed, much early remote sensing simply involved physically providing the analyst with a synoptic perspective. In reconnaissance, this meant that the pilot or an observer took notes or marked up a map to record visual observations of interest. This procedure is still used today in some applications. The forest service, for example, might have an observer in a light aircraft mark up a map showing forest type and condition prior to a spray program to control a gypsy moth infestation. These aerial sketch maps,

as they are called, are one of the most direct forms of remote sensing. Because of the inherent potential for placement error, lack of detailed analysis, and ease of missing a critical phenomena, direct visual observation was, and is, normally augmented or replaced by photographic records. This allows for more detailed analysis, measurement, and mapping and in most cases provides higher resolution through the use of increasingly sophisticated film, camera, and mechanical systems. In many cases, the photographic process is augmented by the visual system in terms of pointing the camera at the right target or locating the aircraft over the target. For much of the photography from space on manned missions, this is still the most common approach.

The analysis of aerial photos evolved along two separate but closely coupled paths. The civilian applications to mapping and surveying continued to grow and become more sophisticated and quantitative as the field of photogrammetry evolved. In parallel with a growing civilian interest, particularly in resource inventory and mapping (e.g., forest cover, soil type and condition, geologic patterns), the First and Second World Wars saw a tremendous push by the military for improved collection systems for reconnaissance applications.

Improvements in film and camera systems were quickly adapted for civilian applications. These applications continue to push the evolution of imaging technology in many areas. A good example of this evolution is the history of color infrared film, or camouflage detection (CD) film, as it was originally known. As photography evolved, it became possible to produce color images by making multilayered photographic emulsions where each layer, through a combination of inherent spectral sensitivity and filtering, was effectively sensitive to a different spectral region (e.g., red, green, and blue). When the film is developed, different color dyes are chemically generated in the different layers to produce a full-color image. This basic process is illustrated in Figure 2.2 and treated in greater detail by James (1977). The location and concentration of the dye are functions of the exposure at any given point. The resultant color images are inherently registered, and by selecting the dyes appropriately, an image that approximates true color can be produced.

When camouflage nets or paints were imaged with panchromatic or color film, they looked enough like natural vegetation to make detection difficult. However, the camouflage of that era typically did not have a high reflectance in the near infrared. Vegetation does (cf. Fig. 2.3), and camouflage detection film was developed to take advantage of this difference in spectral reflectance. The sensitivities of the emulsion layers are shown in Figure 2.4(a). The *color IR film*, as it is commonly called, is normally flown with an external yellow (minus blue) filter to negate the blue sensitivity of the film. In order to display the images, dyes are coupled to the emulsion layers as shown in Figure 2.4(b). Objects such as vegetation with high infrared reflectance values appear red in the resultant image. Camouflage, which was designed to look like vegetation only in the visible, does not have a high infrared return. Therefore, it has more neutral tones on CD film, making it much easier to detect.

Over time, camouflage became more sophisticated, and materials were developed that matched vegetation over a broader spectrum. Figure 2.5 shows an

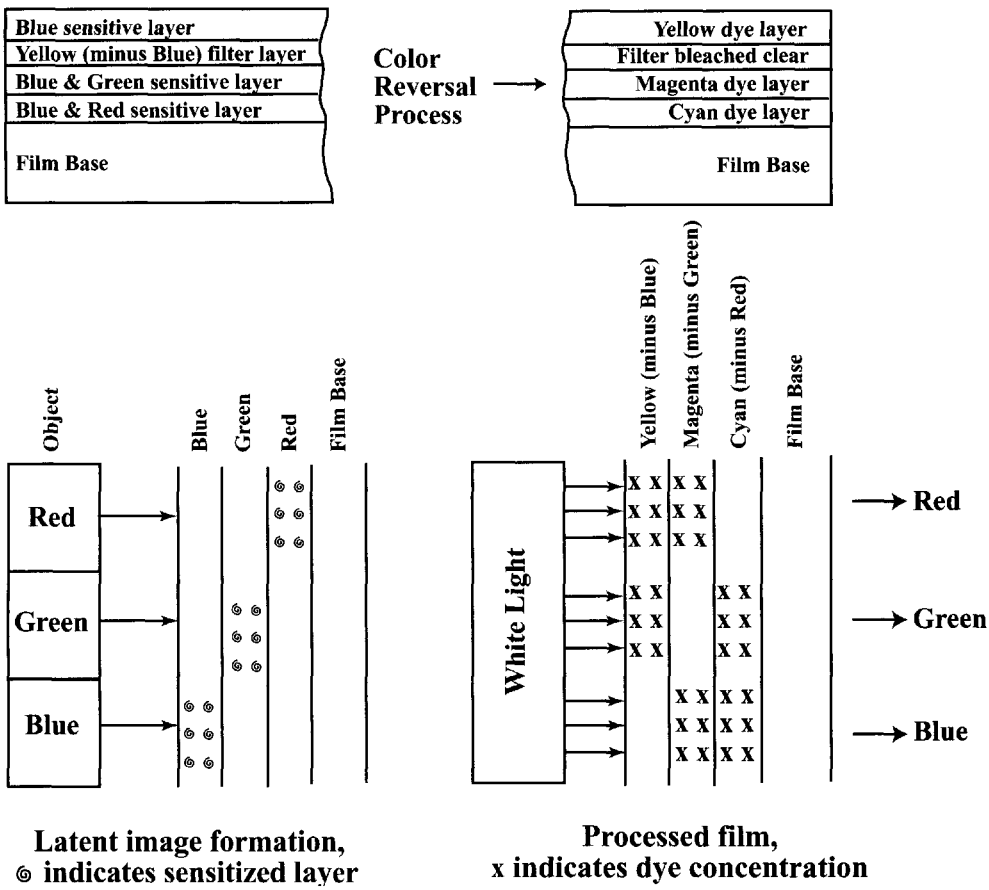


Figure 2.2 Color film concepts.

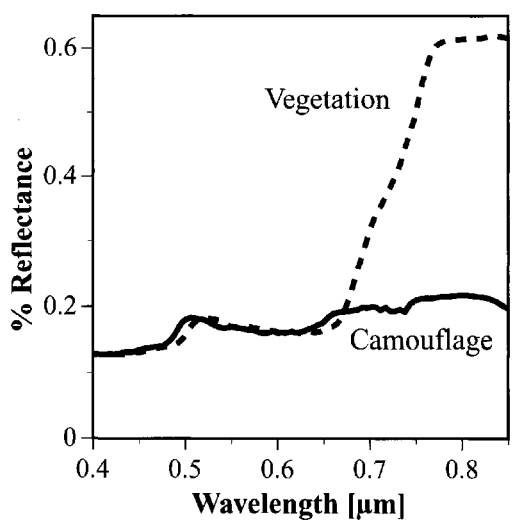


Figure 2.3 Spectral reflectance curves for vegetation and early camouflage material.

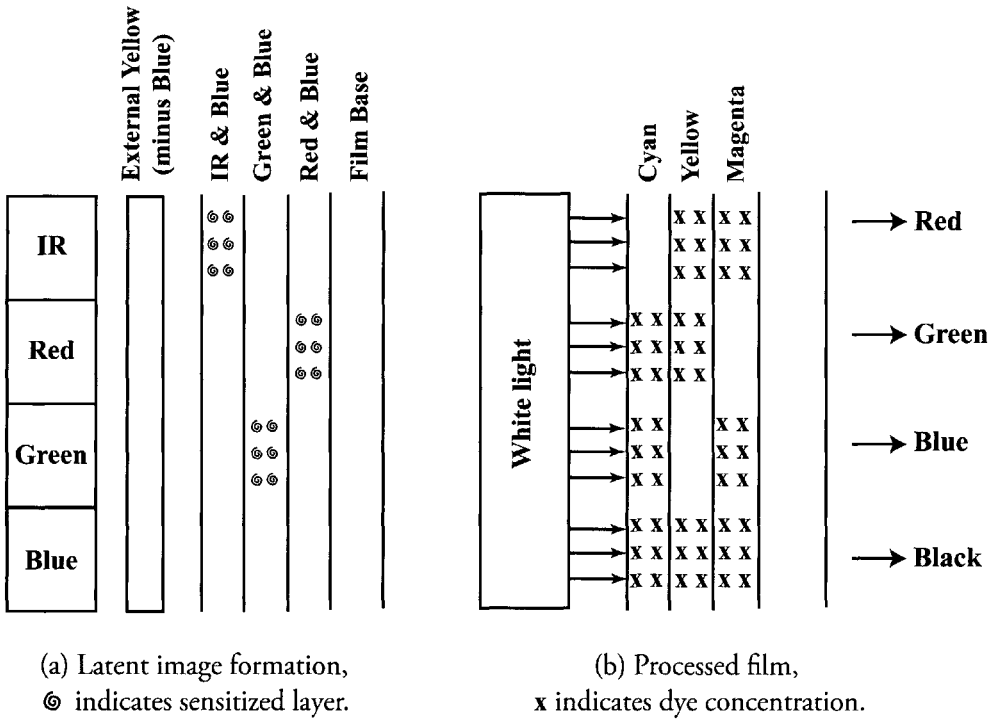


Figure 2.4 Spectral sensitivity of color infrared film and dye coupling concepts (spectral sensitivity curves are shown in Fig. 5.5).

example of a color infrared image of camouflage. In the meantime, CD film was quickly discovered by the rest of the remote sensing community to have other uses. It is widely used in nearly all photographic studies of vegetation condition, because variations in vegetation density and health manifest as changes in color tones in the color IR images (cf. Sec. 5.1). Figure 2.6 shows an example of how color infrared images can be used to delineate vegetation condition.

The interest in imaging in multiple spectral bands grew as scientists discovered that changes in the condition or makeup of soils, water, rock formations, and vegetation often manifested as subtle changes in the reflectance properties in certain spectral bands or by the relative changes between multiple spectral bands. In order to observe these spectral reflectance phenomena, referred to as *spectral signatures*, two approaches were pursued. In some cases, specialized multilayer films were developed. Sprechter et al. (1973) describe a two-layer color film designed for water penetration to aid in the study of water depth and bottom condition in shallow waters. However, this approach is limited to a few spectral bands, and the cost and technical complexity of developing specialized films limit this option to high-volume applications. An alternative, and much more commonly used, approach was found in the form of multilens or multicamera systems. Here spectral filters are used in combination with black-and-white film types to allow the simultaneous acquisition of many images of the same scene in different spectral bands. Figure 2.7 shows an ITEK six-lens system and part of an image set acquired with

Figure 2.5 B/W-infrared image of camouflage objects. See color plate 2.5.

Figure 2.6 B/W-infrared photo showing vegetation condition. A gypsy moth infestation has defoliated most of the trees in the image. "Cosmetic" spray programs along the roadways have prevented severe defoliation in these areas. See color plate 2.6.

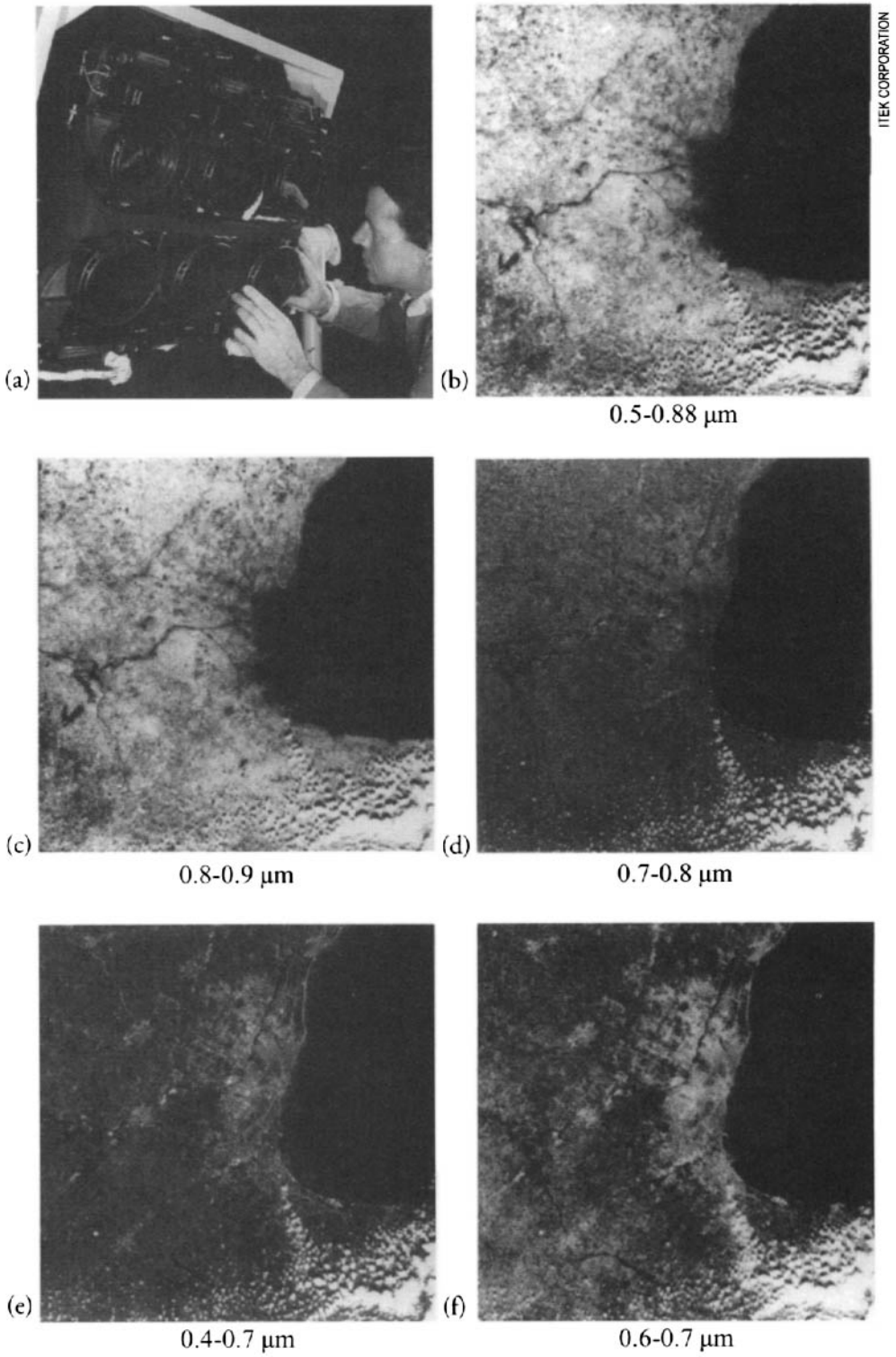


Figure 2.7 Multilens (camera) systems used in collecting multiple frames of black-and-white film through different spectral filters.

the camera system. The advantages of this approach over multilayer films are the large number of spectral bands that can be acquired and the ability to control the spectral region sampled through careful selection of film-filter combinations. The disadvantages of this approach are the increased size and cost of the collection system coupled with the problem that the spectral band images are no longer inherently registered as they are with color film. In the early multiband photographic systems, this problem was solved by projecting the images in groups of three onto a common plane using various forms of optical recombination. The images were then shifted and rotated until they were in visual alignment. By projecting each of three images through a red, green, or blue filter, the spectral information in the resulting color image could be interpreted. With today's technology, it would be more common to digitize all the images into a computer compatible form (cf. Sec. 5.1) and use digital image registration and resampling techniques (cf. Sec. 12.3).

The growth of photographic science and technology, coupled with the synoptic perspective provided by remote sensing, allowed us to look at the earth's surface in new ways. These new perspectives, and hitherto unseen spectral and spatial patterns, revealed to scientists a host of new applications. These applications include archeology, meteorology, marine science, water resources assessment, land use analysis, geology, soils science, civil engineering, agriculture, forestry, range management, and wildlife biology [cf. Estes and Thorley (1983)]. The application of photographic and electro-optical remote sensing to these various fields not only spurred the rapid advance in image acquisition techniques but also stimulated the evolving field of quantitative imagery analysis.

2.2 QUANTITATIVE ANALYSIS OF AIR PHOTOS

The quantitative analysis of remotely sensed data is normally divided into at least two parts. The first, *photogrammetry*, is concerned with using photographic images to make measurements of the size, height, and location of objects or land forms. As such, it includes the science of mapping the topography of the earth's surface and of locating and measuring the dimensions of objects on the surface. The second part is based on radiometric analysis of the images. It is this radiometric analysis that will occupy the bulk of our attention in later chapters. However, in this chapter we want to lay the foundation for the latter treatment by introducing some basic photogrammetry concepts, as well as the concept of using a camera as a radiometer.

2.2.1 Photogrammetry

To keep our treatment relatively simple, we are going to restrict our consideration of photogrammetry to measurements from conventional, vertical (nadir) viewing photographic systems. However, the treatment is easily extended to more complex EO systems. Furthermore, we are only going to introduce the most rudimentary photogrammetric concepts that are necessary for general remote sensing purposes. A more complete treatment can be found in Wolf and DeWitt (2000) and Slama

(1980). Our interest in photogrammetry parallels that of early photogrammetists, who wanted to know (a) how to convert a distance measurement on a photo to a distance on the ground, (b) how photo coordinates could be related to ground or map coordinates, (c) how height could be determined from photo coordinates, and (d) how air photos could be used to make topographic maps.

To begin, we need to introduce the terminology and a set of parametric conventions to be used in our discussion of photogrammetry. These are illustrated in Figure 2.8. For simplicity, we will assume that we are only dealing with vertical images, that all measurements are from positive photographic prints, and that no geometric distortions exist due to imperfections in the camera system or spatial instabilities in the film. Note: All these restrictions can be relaxed, but the treatment is beyond our scope. The camera location when the image is acquired is referred to as the *exposure station* (L). The camera focal length is f. The positive image is considered to be projected back through the exposure station and located in a plane a distance f below the exposure station (i.e., it is a simple mirror image of the negative at the exact scale such as would be obtained by contact printing). The *datum* refers to a reference plane to which points on the terrain are projected for mapping or measurement purposes [e.g., mean sea level (MSL) is a common datum plane]. By convention, the plus x dimension in the image is along the flight path, with the origin at the optical axis, and the plus y axis is 90° counterclockwise from the plus x axis. Lowercase letters are used to define image locations on the positive print (primed on the negative), with the uppercase letter representing the corresponding point in the scene (primed uppercase values represent the projection of the point to the datum plane). The points on the optical axis are generally denoted by the letter o and referred to as *principal points*; therefore, the *optical axis* would pass through the points o', L, o, O, and O'. An arbitrary scene coordinate system is then defined by projecting the image's x and y axes onto the datum plane, with the plus z axis being straight up.

An analysis based on the proportionality of sides of similar triangles in Figure 2.9 shows how image measurements are related to scene or map measurements by the local image scale(s), i.e.,

$$s = \frac{f}{H - h} \tag{2.1}$$

where s is the image *scale*, f is the *focal length*, H is flying height above datum, and h is the height of the measured point above datum. Clearly, the scale of the image will vary with elevation, and this variation will be greatly reduced in high-altitude or satellite images. Figure 2.9 also makes it clear that measurements (distances, angles) made on the image must first be corrected for elevation variation before they are meaningful. In many cases, this requires an ability to convert image coordinates into ground coordinates.

The equations of transformation can again be easily derived using the proportionality feature of sides of similar triangles based on the geometric construction shown in Figure 2.10. The resulting transformation equations are

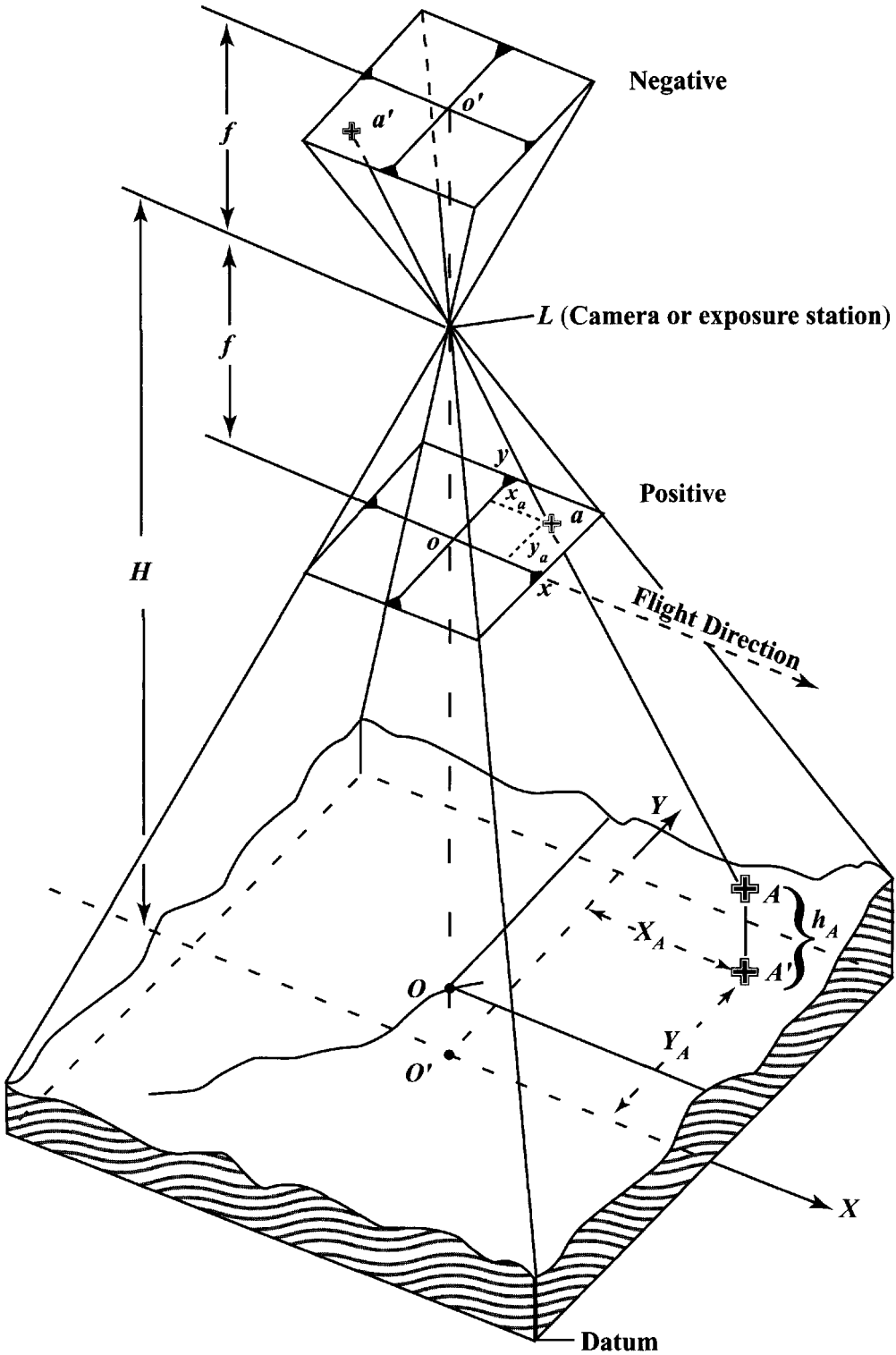


Figure 2.8 Geometric orientation parameters used in photogrammetry.

Scale at height h_{AB} is $s_{AB} = \frac{\overline{ab}}{\overline{AB}} = \frac{\overline{La}}{\overline{LA}} = \frac{\overline{Lo}}{\overline{LE}} = \frac{f}{H - h_{AB}}$

Scale at height h_{CD} is $s_{CD} = \frac{\overline{cd}}{\overline{CD}} = \frac{\overline{Lc}}{\overline{LC}} = \frac{\overline{Lo}}{\overline{LG}} = \frac{f}{H - h_{CD}}$

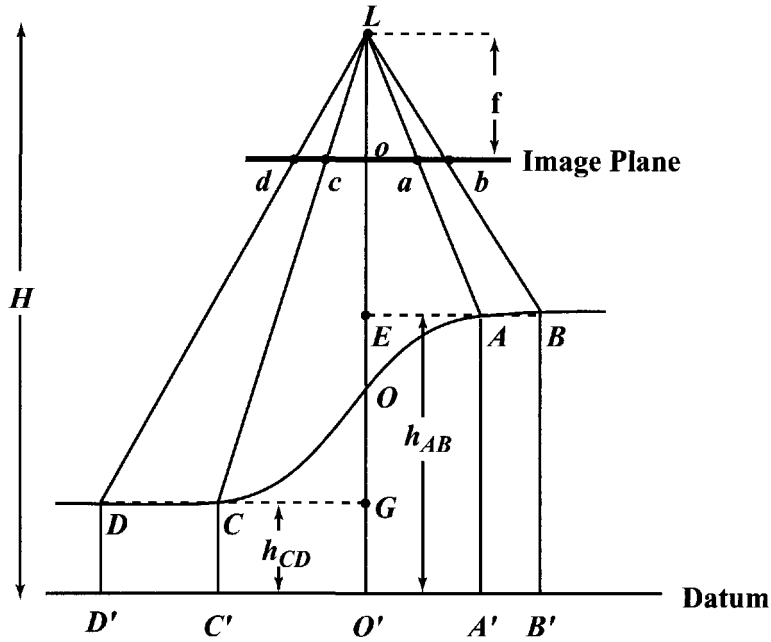


Figure 2.9 Scale determination.

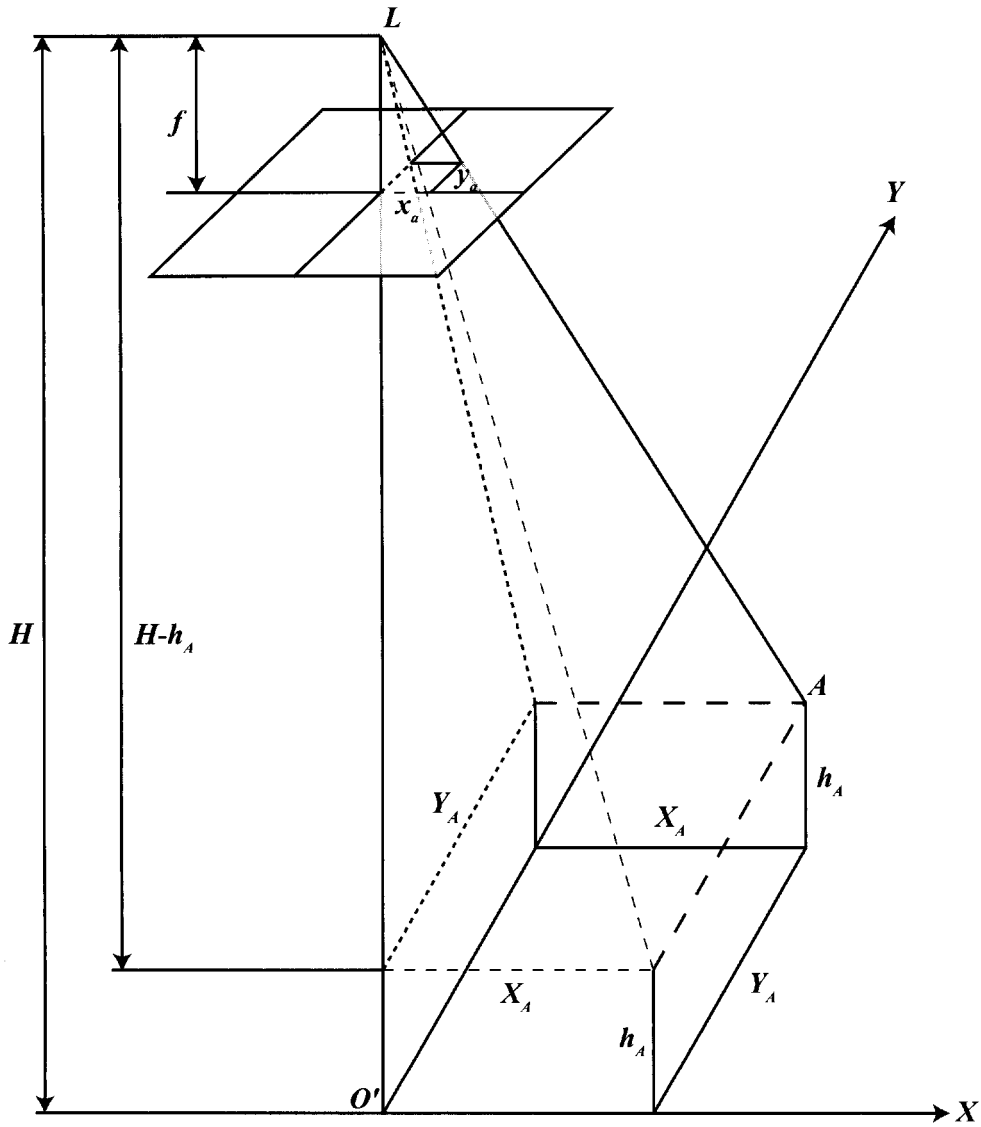
$$X_A = \frac{H - h_A}{f} x_a$$

and

$$Y_A = \frac{H - h_A}{f} y_a \tag{2.2}$$

where X_A and Y_A are the geometric coordinates in the ground coordinate system for point a in the image with image coordinates x_a, y_a. Using Eq. (2.2) we could transform any data in the image to a ground coordinate plane and make quantitative dimensional and angular measurements in the ground coordinate plane. In addition, if the image contained two or more points whose location and height were known in a global ground coordinate system, a simple geometric transform from the arbitrary X, Y space to a global reference system can easily be derived. The only drawback to this process is that Eqs. (2.1) and (2.2) require the elevation of the target (h_A).

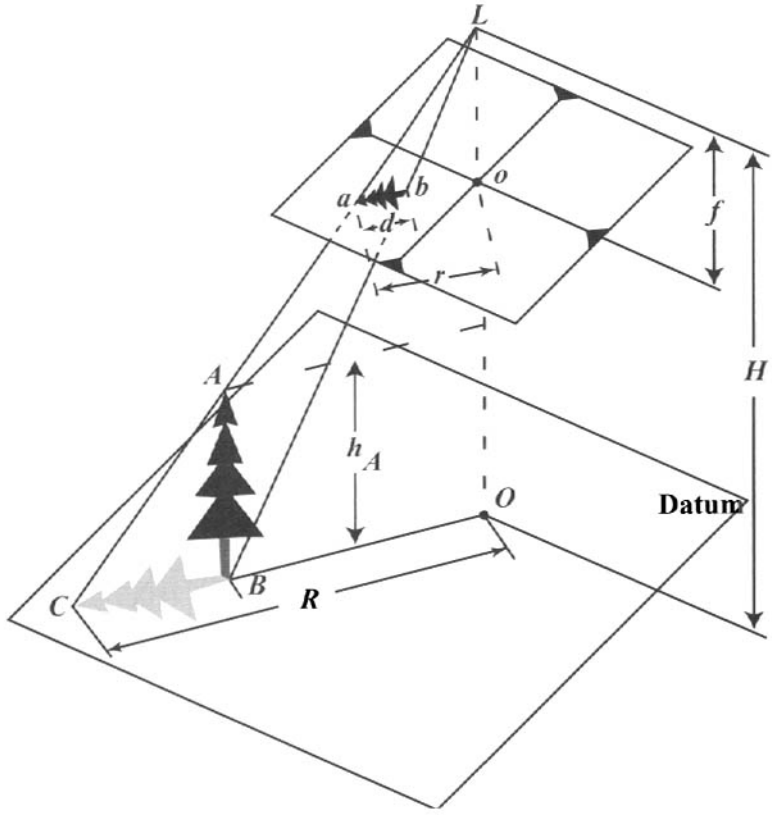
There are two common methods used in photogrammetry for determining elevation. The first is normally used for determining the height of vertical targets such as towers and buildings relative to the local ground plane. This method uses the *relief displacement* that results when the top of an object viewed off nadir is



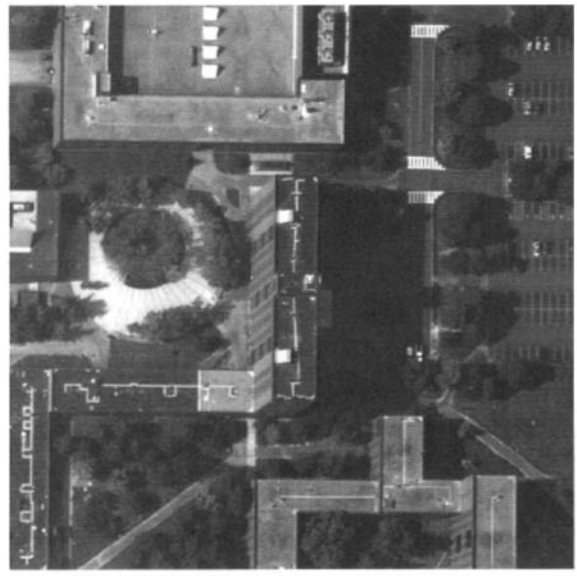
$$\frac{x_a}{X_A} = \frac{f}{H - h_A}$$ From similar triangle - - - -

$$\frac{yA}{Y_A} = \frac{f}{H - h_A}$$ From similar triangle .....

Figure 2.10 Geometric construction for derivation of ground coordinate transforms.



(a) Geometric construction used to derive an expression for the target height based on relief displacement.



(b) Portion of an image showing radial relief displacement.

Figure 2.11 Relief displacement concepts.

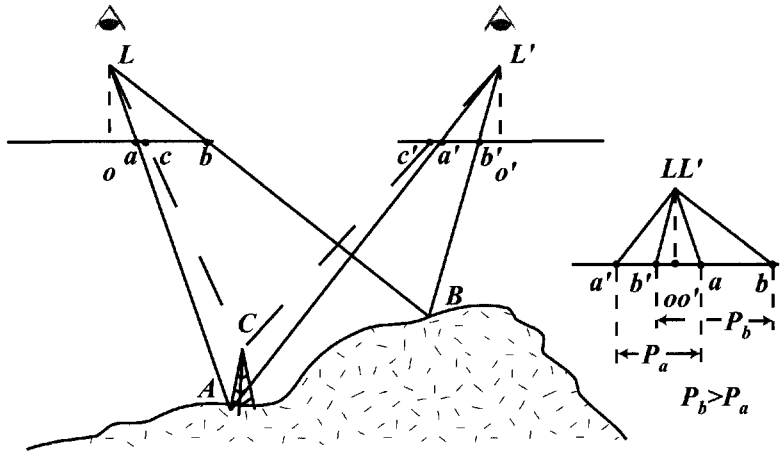


Figure 2.12 x axis displacement (parallax) in stereo pair images is a function of object height.

imaged farther away from the principal point than the bottom of the object (i.e., the distance d as shown in Fig. 2.11). Note that for a photographic system, the relief displacement is radially outward from the center of the format and increases with object height and radial distance. Once again, using cascaded similar triangles, we can derive an expression for the height of a target as

$$h_A = \frac{Hd}{r} \tag{2.3}$$

where h_A the height of the target, d is the relief displacement as measured on the photo, H is the flying height above the base of the target, and r is the radial distance from the principal point to the top of the target. Note that many EO systems that don't look fore and aft of nadir generate images with relief displacement only outward from the center line of the image along a line perpendicular to the flight line. In those cases where the side of an object can be seen due to relief displacement, the height can be determined from two simple measurements if the flying height above local datum is known. Regrettably, this method is not applicable to objects near the center of the format or for finding the height of sloped surfaces such as terrain.

For mapping height throughout a scene, images of the scene from two different perspectives are required. Such stereo pairs have long formed the basis for three-dimensional viewing using optical systems designed to effectively place one eye at one camera station and the other at a second. As seen in Figure 2.12, height variations result in differential displacement distances along the x direction for objects in the scene, which the visual system interprets as height changes in the same way our normal stereoscopic vision does. By simply forcing one eye to see one image and the other a second, *stereo* perception can be achieved. Figure 2.13 is a stereo pair that can be viewed using a simple pocket stereo viewer. One can take advantage of this phenomenon by measuring the x axis displacement of scene ele-

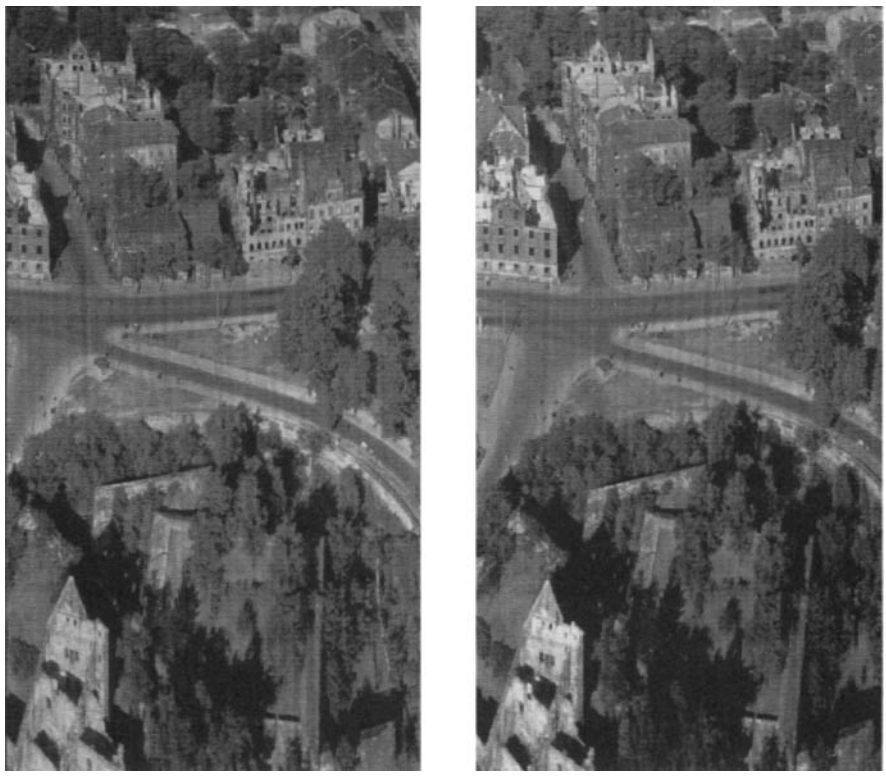


Figure 2.13 Black and white stereo pair reproduced from early color film. This stereo pair was used to assess bomb damage during World War II.

ments in a stereo pair. This displacement is known as *parallax* (p). The geometric analysis is shown in Figure 2.14. First, the location of the first images principle point is found in image 2 (and vice versa) to determine the local line of flight to be used as a common x axis. Next, the point of interest (a, a') is located in both images, with the prime used to designate objects in image 2. We define the *air base* (B) to be the distance between the exposure stations (L and L'). We identify two similar triangles by projecting a, a', and A onto the X, Z plane to form LA_xL' and shifting $L'o'$ to coincide with Lo to form a'_xLa_x. Then we have

$$\frac{p_a}{f} = \frac{B}{H - h_A} \tag{2.4}$$

where $p_a = x_a - x'_a$ is the parallax (note: x'_a would be negative in the example shown). Rearranging yields an equation for height determination for an arbitrary point A:

$$h_A = H - \frac{Bf}{p_a} \tag{2.5}$$

Equation (2.4) can be rearranged and substituted into Eq. (2.2) to also yield expressions for the ground coordinates in terms of parallax, i.e.,

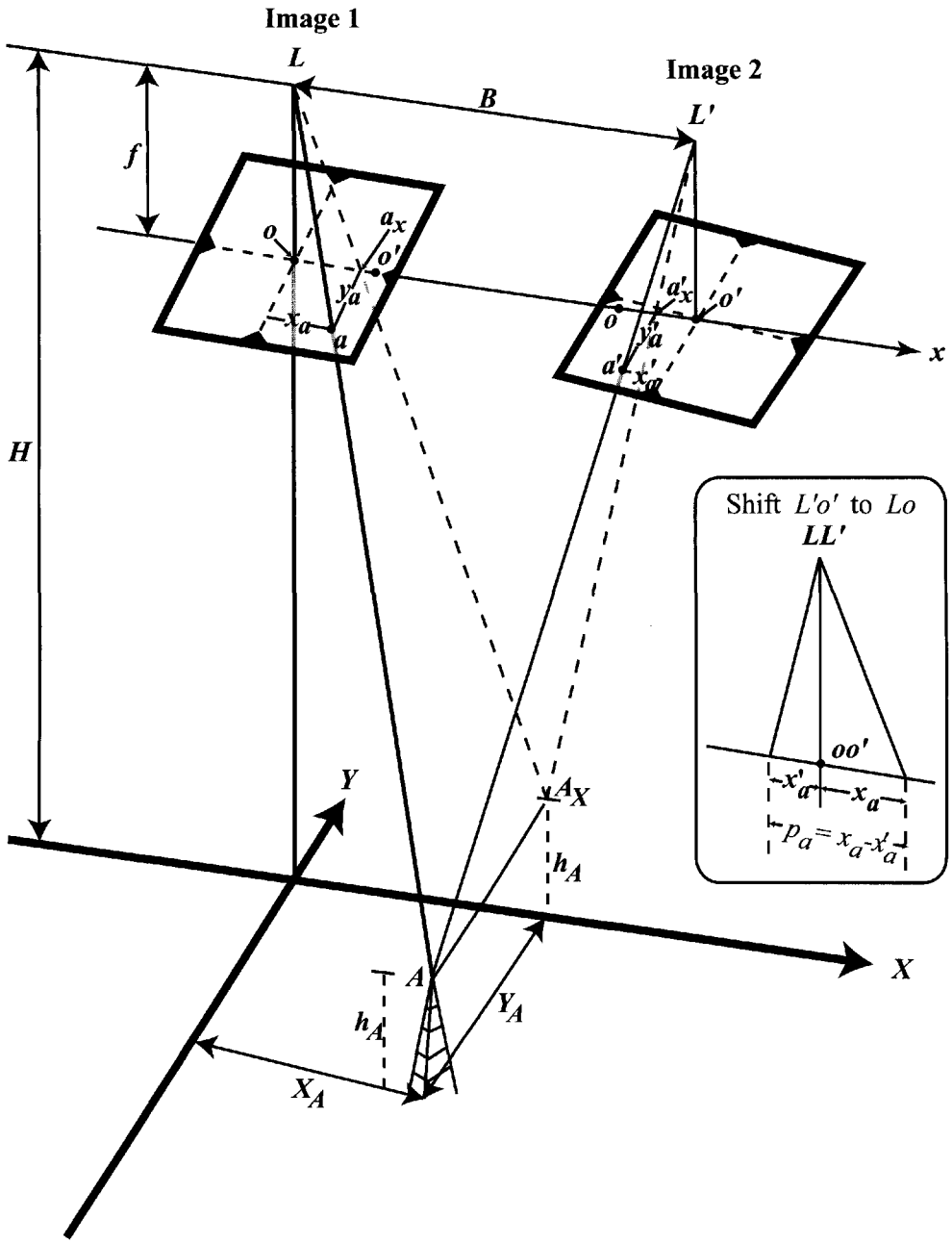


Figure 2.14 Geometric construction for derivation of the parallax equation for height determination.

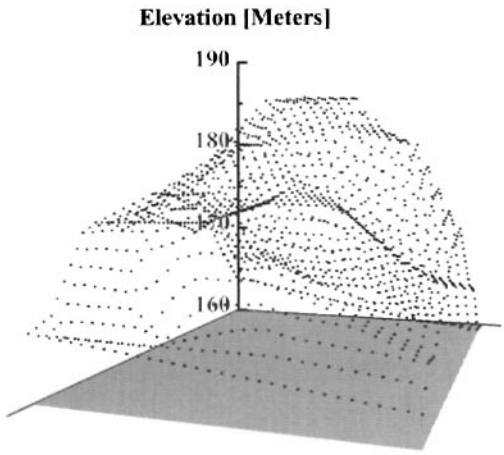
$$X_A = \frac{Bx_a}{p_a} \tag{2.6}$$

$$Y_A = \frac{By_a}{p_a} \tag{2.7}$$

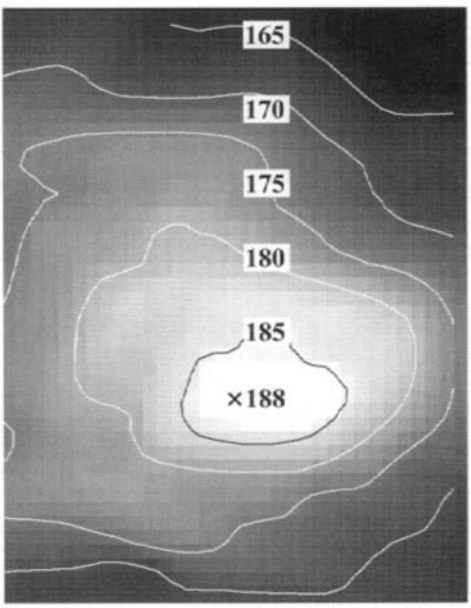
The parallax measurements can be made by direct measurements from the photos with a ruler or, more commonly, with any of a host of devices designed to simplify the job of mapping height throughout an image. Many of these employ the human visual system's stereoscopic perception to allow height determination even over uniform terrain where no distinct features exist for reference. These devices, called stereo plotters, can be used to compile topographic maps or *digital elevation models* (*DEM*) of the scene. A DEM is a representation of the terrain in a grid form where every grid center represents an *x,y* location in the datum plane having a height associated with it, as shown in Figure 2.15. The elevation, slope, and orientation data that can be derived from stereo image pairs are extremely useful in all the obvious engineering and hydrological applications of remote sensing, as well as in many more subtle radiometric applications that will be introduced in later chapters.

In traditional photogrammetry, precisely locating the camera station (*L*) and orienting the optical axis often require a great deal of effort and difficult-to-obtain ancillary data. Today, the combination of the *Global Positioning System* (*GPS*) and *Inertial Navigation Systems* (*INS*) makes this process much more straightforward. The GPS consists of 24 or more satellites in earth orbit that have precise knowledge of their own location. Each sends out a unique digital code tied to a very accurate atomic clock. Each GPS receiver picks up the signal from the satellite and compares the received code to an internally generated code that it produces at the same time. Due to the time delay in getting the radio wave from orbit to the GPS location, the codes will be slightly out of sync. Digital logic computes how much the codes need to be shifted to realign with each other. The shift is proportional to the time delay. This time delay, coupled with the speed of light, yields the distance from the satellite to the GPS receiver. As illustrated in Figure 2.16, each satellite distance generates a sphere of valid locations, and three such spheres reduce the mathematically valid locations to two points, one of which can be logically rejected (e.g., it is deep in space absurdly far from a possible location). In practice, a fourth satellite is used in the GPS process to account for the inaccuracies of the clocks on the GPS receivers. Using four or more satellites generates multiple solutions for the receiver's location. If the location errors are attributed to a shift in the receiver's time code, a correction (or unshifting) can be computed that moves the solutions to a single valid point. The output from this process is a precise *X*, *Y*, *H* location of the receiver (i.e., the camera location). The downside to GPS is that it generally cannot provide good orientation information or rapid updates of location for rapidly or irregularly moving objects. INS systems consisting of three axis accelerometers and gyroscopes can provide detailed information about the displacement and orientation of a sensor platform. By combining GPS and INS data, we precisely know both the sensor location (X, Y, H) and the orientation of the optical axis (roll, pitch, and yaw) to high accuracies at essentially any point in time.

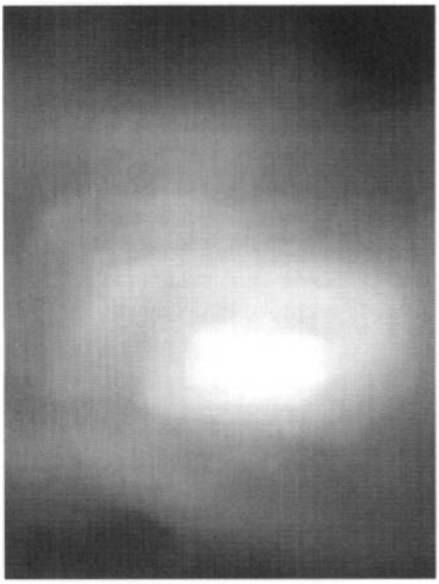
Before leaving this discussion of photogrammetry, we should point out the rapidly expanding use of alternate more direct methods of terrain mapping, including Interferometric Synthetic Apertures Radar (IFSAR) and terrain mapping *Laser Detection and Ranging* (*LADAR*). Terrain mapping LADAR systems take advantage of the GPS-INS technology described above to precisely locate the airborne or



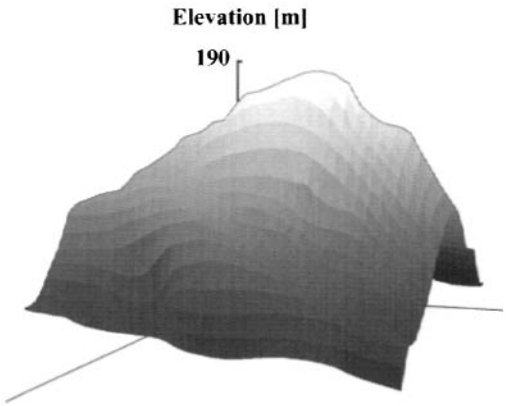
(a) Plot of elevation heights (points or posts).



(b) Contour plot superimposed on gray-level coding of elevation.



(c) Gray-level coding of elevation.



(d) Combination of (a) and (c) to allow surface visualization.

Figure 2.15 Digital elevation model concept.

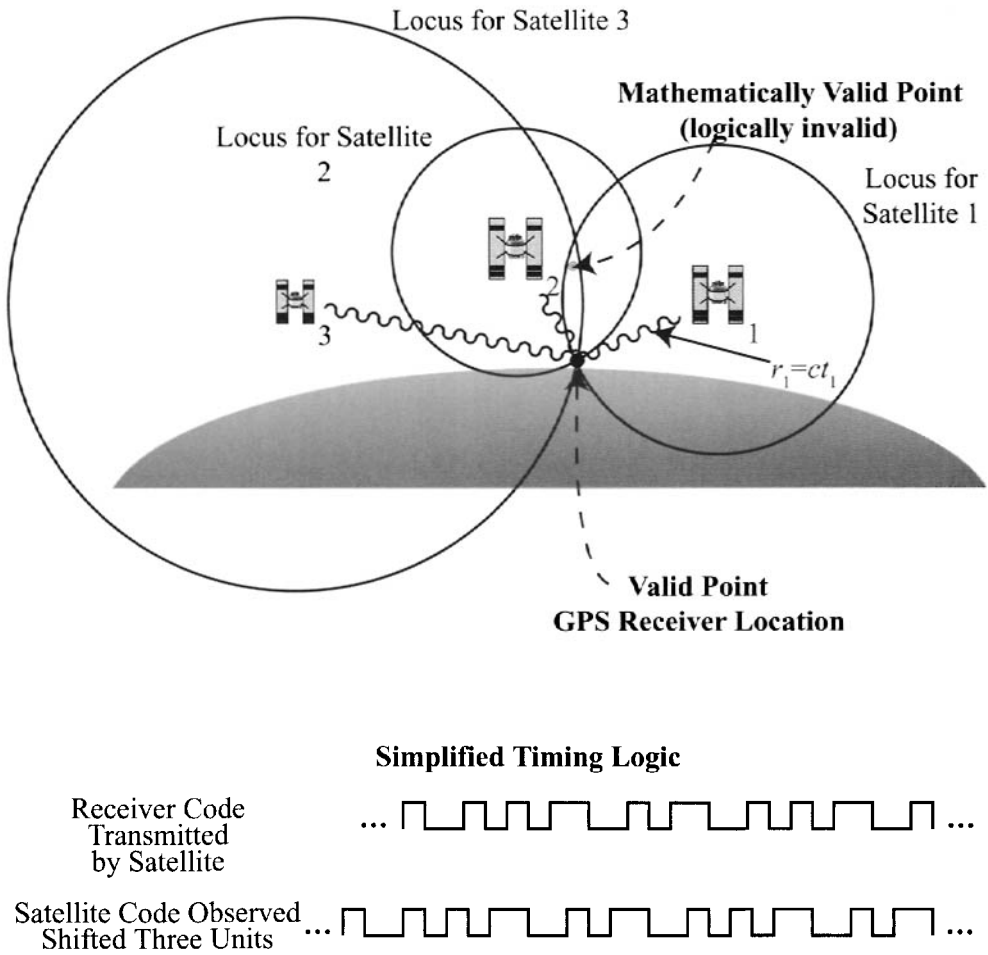
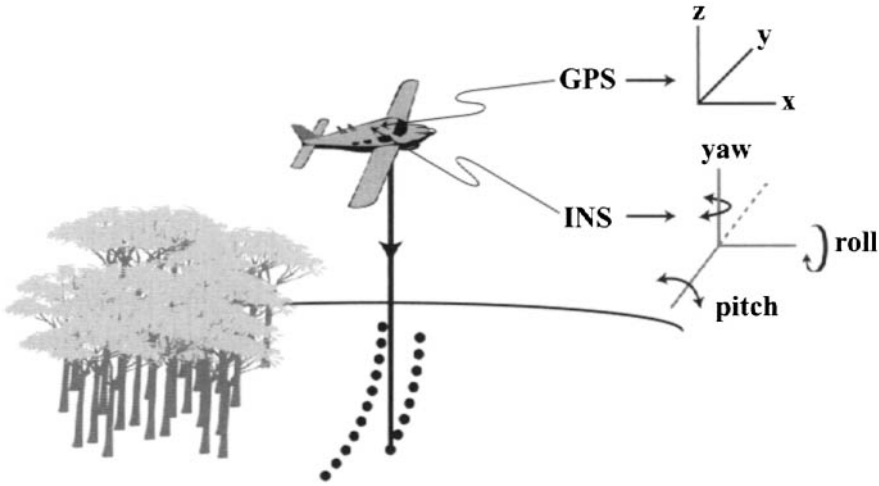
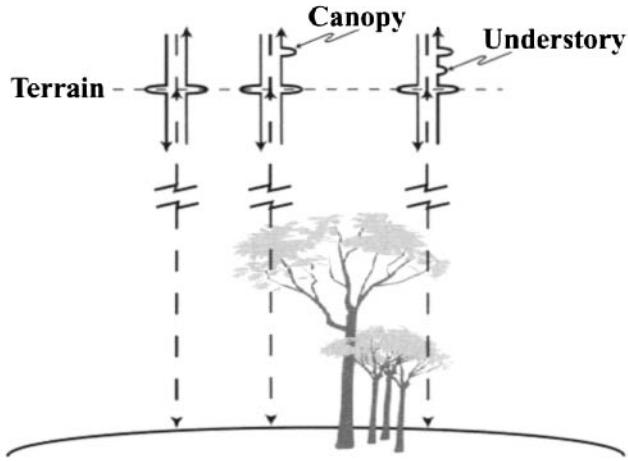


Figure 2.16 Simplified illustration of GPS concept showing three satellites and a drawing of how the timing logic can be used to decode signals into time and distance. In practice, a fourth satellite is needed to remove errors due to inaccuracies in the receiver's clock.

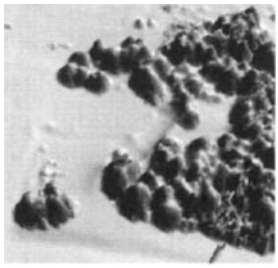
satellite based laser source and receiver (cf. Fig. 2.17). The instrument sends out a series of laser pulses that are swept across the terrain. The time from pulse generation to detection provides the round trip transit time to the ground and back and, therefore, the range to the ground from which the terrain elevation can be computed. Current systems can generate pulse densities allowing terrain mapping on posts of fractions of a meter on a center and with elevation variations of centimeters. By just looking at the leading pulse edge from these systems, we end up mapping not the surface of the earth but the tops of whatever may be above the terrain (e.g., tree canopies and buildings). In some cases, this may be desired; in other cases, spatial filters must be employed to attempt to remove these features from the final maps. Advanced systems allow the user to time sample the return signal at a high enough rate to characterize the full shape of the returned laser pulse. This allows the mapping, for example, in tree canopies of the height of the canopy and the location of the ground. More detailed analysis can even be performed to study the density of



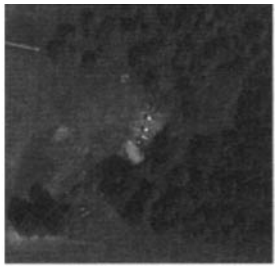
(a) The terrain sampling processes showing point sampling with scanned laser pulses from which range is calculated.



(b) A simplified illustration of how the structure of the return pulse can be analyzed to provide information on canopy height and makeup.



(c) LADAR-derived surface elevation map of tree line with height mapped as shades of gray.



(d) Aerial image of the same area shown in (c).

Figure 2.17 Illustration of the use of LADAR to map terrain elevation.

the canopy and the characteristics of sub canopy foliage. Figure 2.17(c) shows the high level of detail that can be obtained using advanced LADAR surface mapping systems. A more thorough treatment of modern photogrammetric methods, including digital approaches to stereoscopic image analysis and the use of LADAR and IFSAR for terrain elevation mapping, can be found in McGlone (2004).

2.2.2 Camera as a Radiometer

As the field of remote sensing evolved in the middle of the 20th century, interest grew in trying to quantify some of the information that appeared to be available in the air photos. For example, an image analyst could see that one water body was more turbid than another, but not how much different, or that one soil condition was different than another, but not what caused the difference or what the magnitude of the difference might be. Laboratory and field studies showed that many questions about the condition of objects were related to the reflectance of the object. Since the optical density recorded on the film varied with the radiometric reflectance of the object, the logical question became how to quantify the relationship between the two, or how to employ the camera as a radiometer. Piech and Walker (1971, 1974) demonstrated that water turbidity could be related quantitatively to film density and that soil texture and moisture could be studied based on the optical density values recorded on color film transparencies. The details of this type of analysis are developed in Chapters 3-7. For the present, we simply postulate that the required functional relationships exist, i.e.,

$$r = f'(cond.) \tag{2.8}$$

$$L = f''(r) \tag{2.9}$$

$$H = f'''(L) \tag{2.10}$$

$$D = f''''(H) \tag{2.11}$$

$$D = f(cond.) \tag{2.12}$$

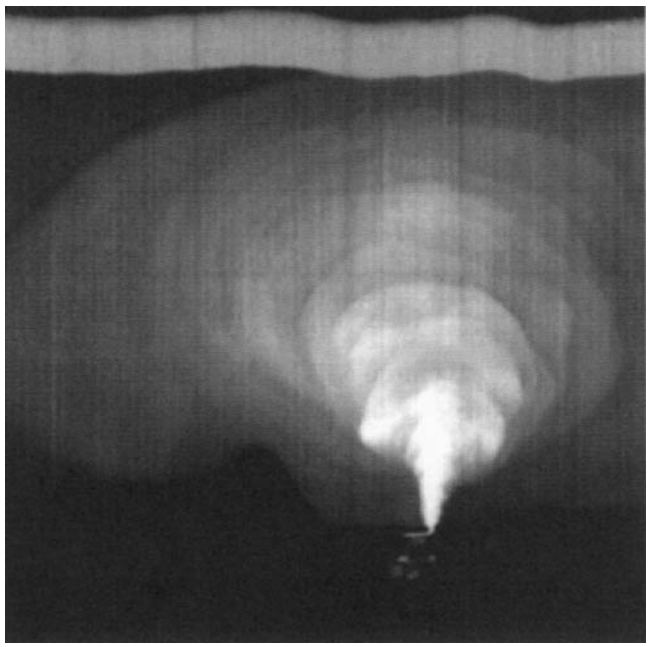
where *Cond.* is a parameter representing the type or condition of the target (e.g., turbidity, soil moisture, vegetation health), r is the reflectance of the target (possibly in several spectral regions), L is the radiance reaching the sensor, H is the exposure on the film, D is the *density* (- log of the transmission) recorded on a photographic transparency, and f, f', etc., represent the functional relationships between the indicated parameters.

If we can derive an expression for each of the primed functions in Eqs. (2.8)–(2.11), or for the aggregate function f in Eq. (2.12), then it is possible to begin to think of the camera as a radiometer. In early applications, this was accomplished empirically by making many field observations of the parameter of interest (cond.) and film density readings at image locations corresponding to where the "ground truth" observations were acquired. *Ground truth* is a term used to describe any surface measurements used to assist in the analysis of remotely sensed data. In this

case, the ground truth would consist of an extensive set of measurements to quantify the condition parameter (e.g., soil moisture). A curve fit between the two data sets would generate an approximation for the functional form of f in Eq. (2.12). The resulting relationships were invariably nonlinear, required extensive ground truth to develop, required near simultaneous ground truth and overflight data, were not robust, and had errors associated with not knowing the nature of the cascaded functions (f', f'', etc.) [cf. Lillesand et al. (1973)]. Nevertheless, early successes with this "empirical" approach led to efforts to derive theoretical expressions for the various functions and to calibration methods aimed at improving the robustness, accuracy, and utility of quantitative radiometric image analysis. The bulk of our attention in the rest of this volume will be directed at understanding the tools that evolved in response to the need to use the camera as a radiometer.

2.3 EVOLUTION OF EO SYSTEMS

From the late 1950s onward, *electro-optical* (*EO*) systems have competed with and/or complemented photographic systems in most applications. EO systems range from the simple vidicon and solid-state video cameras (familiar to us through the consumer market) to very exotic systems with either complex opto-mechanical systems for image formation or large arrays of tiny solid state detectors. Film cameras have a distinct advantage in that they provide a low-cost, high spatial resolution solution to many problems. However, as imaging acquisition requirements became more demanding, film solutions were not always available. For example, film spectral response could not be pushed much beyond 1 μm. Imaging in the thermal infrared spectral region required, and became a driving force behind, the development of EO sensors. Another force behind the evolution of EO systems was the growing interest in using spectral reflectance variations to help identify and characterize earth surface features. As discussed in Section 2.1, to do this simultaneously in many spectral bands is very difficult with film. These problems can be overcome with EO systems, although only at the cost of geometric integrity and spatial resolution (particularly in early systems). This ability to "see" things the camera couldn't by using a wider range of the electromagnetic spectrum and also to see in many relatively narrow spectral regions led to the development of a host of airborne EO systems. We will consider these systems in more detail in Chapter 6. The utility of these systems can be seen in the thermal infrared image of a cooling water discharge from a power plant shown in Figure 2.18. The brightness of this image is proportional to temperature. Images from EO IR systems such as this were used to map the magnitude and extent of the environmental impact of cooling waters on the receiving waters and to monitor compliance with environmental regulations. These images let us study the world in ways that had never before been possible. Figure 2.19, for example, shows several images from an airborne EO system representing different regions of the EM spectrum. Note both the dramatic and the more subtle changes between materials in different wavelength regions. In Chapter 9 we will discuss how these variations in the observed spectral brightness can be used to map land cover type and the condition of selected targets (e.g., crop stress, water quality).



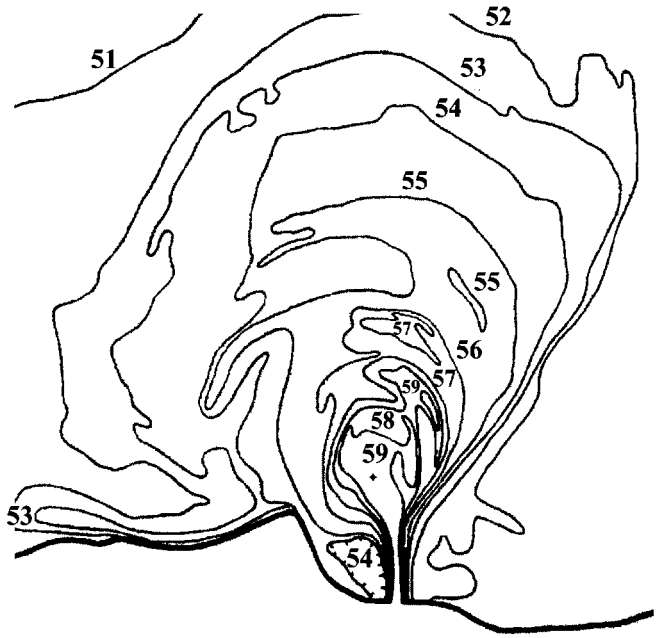


Figure 2.18 Thermal infrared image of a power plant cooling water discharge (white is hot) and surface temperature map derived from the image (°F).

2.4 SPACE-BASED EO SYSTEMS

Early photographs from space clearly demonstrated the value of the synoptic perspective from space. This interest in space-based systems helped to spur the evolution of EO systems because of the operational problems associated with the recovery of film from space. The inherent electronic nature of the EO images solved the problem, since the images could be transmitted to ground receiving stations.

In order to understand the performance of some of the space-based sensors described in Chapter 6, we need to understand a little about satellite orbits. Simple scale effects dictate that most systems for earth observation use *low earth orbits (LEO)* (450–1200 km) to maintain acceptable spatial resolutions. Furthermore, to facilitate comparable image scales anywhere on the earth, near circular orbits are usually preferred. Orbit selection has to take into account the satellite motion, earth rotation, and the relative sun-earth-satellite orientation angles. In many cases, we desire an imaging system that will periodically pass over nearly all points on the earth (i.e., it would periodically image most of the earth when in nadir viewing mode).

To accomplish this, near polar orbits are used to take advantage of the earth rotation as shown in Figure 2.20. To facilitate analysis of the satellite images, it is often desirable to have the sun-earth-sensor angle reasonably constant over time. This makes images taken on multiple overpasses more comparable, since we don't have to correct for illumination angle variations. This can be accomplished to a good approximation using a sun synchronous orbit. In a sun synchronous orbit, the satellite crosses the equator at the same local solar time with each pass, and, since it is traveling relatively fast compared to the earth's rotation, it will pass over points north and south of the equator at approximately constant solar time. The only variations in sun-object-sensor angle will be due to seasonal variations in the solar angle. Repetitive images of the same area at the same time of year will have nearly equal sun-object-sensor angles. All of these factors combine to dictate sun-synchronous, near-polar, low earth orbits for most sensors. These are typically several hundred kilometers above the earth, have orbital periods of about 90–100 minutes, and are inclined 5–10 degrees from the poles (inclination angles of 95–100°). The Landsat satellites, for example, were placed in approximately 705 km high sun synchronous orbits crossing the equator at approximately 10:00 a.m. in the descending node. The 98.2° inclination angle, coupled with the 90-minute orbital period, 185-km swath width, and the earth rotation, results in complete global coverage (except at the poles) every 16 days. This type of orbit is commonly used for earth observation and high-resolution meteorological sensors (e.g., SPOT, Landsat, NOAA, TERRA).

Another common type of circular orbit is the *geosynchronous earth orbit (GEO)*, which has a 24-hour period and is used to locate a satellite roughly over a fixed longitude. A satellite in GEO revolves around the earth with a period equal to the earth's rotational period. The ground track will tend to be roughly a figure 8, as shown in Figure 2.21, with the latitude limits a function of inclination. If the inclination is near zero, the satellite will effectively remain stationary over a point

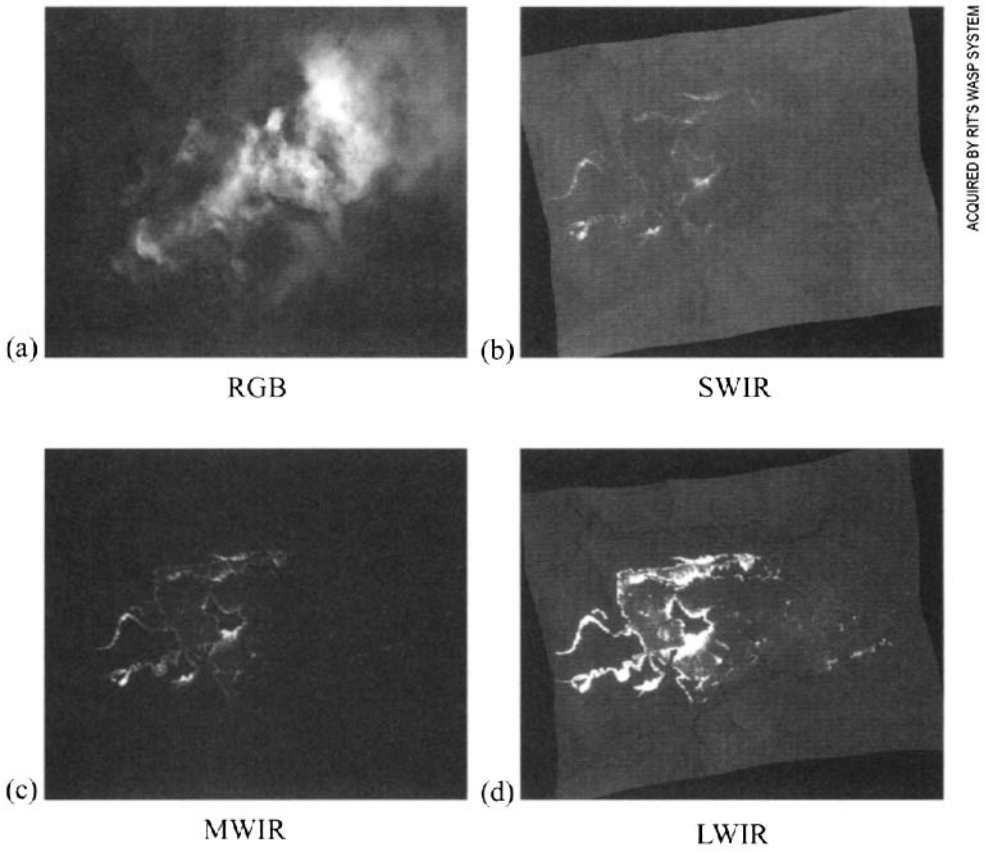


Figure 2.19 Multispectral images of an active fire taken simultaneously in various spectral bands. Note the visible (RGB) image shows smoke but could not be used to map the fire line, the SWIR images shows the active burn area, and the MWIR and LWIR images show both the active fire and smoldering or recently burned areas.

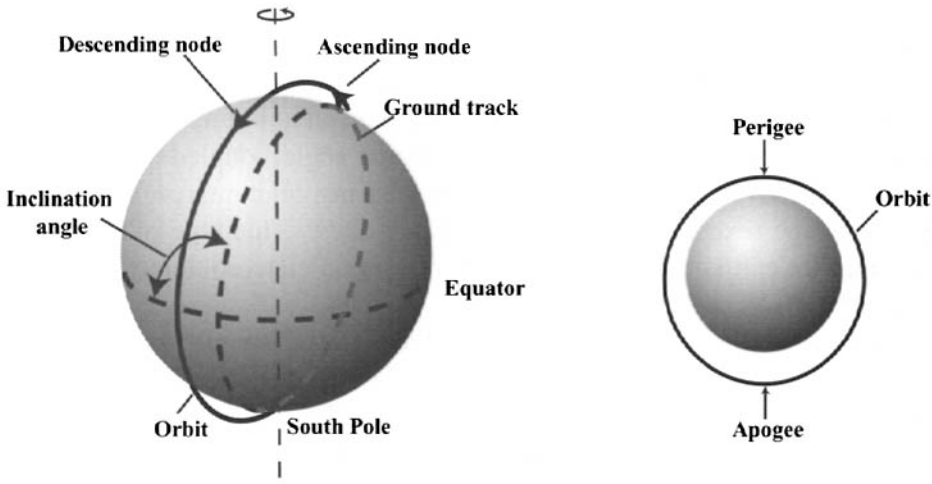


Figure 2.20 Some low earth orbit (LEO) concepts.

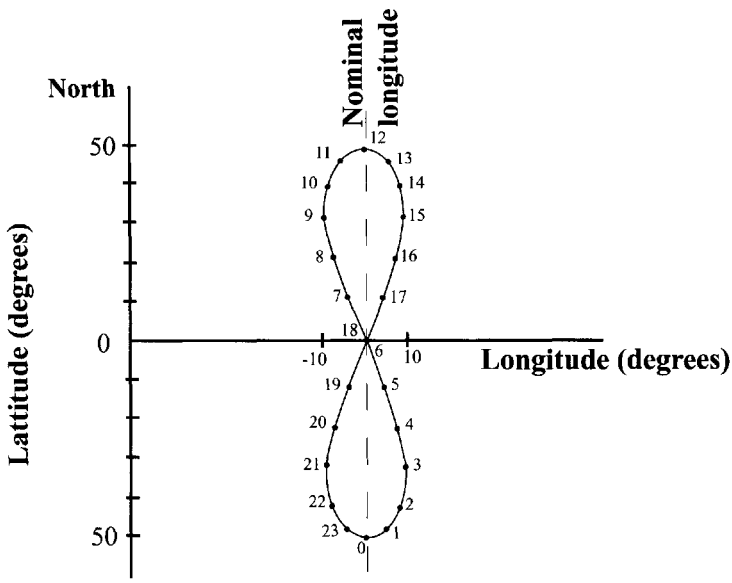


Figure 2.21 Ground track for a satellite in geosynchronous orbit.

on the equator in what is called a *geostationary orbit.* These geostationary orbits are convenient when rapid repetitive coverage of an area is required. The moving images of cloud patterns so commonly seen on televised weather reports are acquired from geostationary satellites with sensors that acquire images every half hour (cf. Fig. 1.6). Regrettably, in order to maintain a satellite in GEO, it must be at an altitude of approximately $36 \cdot 10^3$ km (roughly 40 times higher than the LEO systems discussed above). As a result, it is very difficult to operate high-resolution systems from GEO. On the other hand, it is a very attractive orbit for communications systems. A communications satellite in geostationary orbit will always be in line-of-sight communication with the same points on the ground. A constellation made up of three geostationary satellites can be in line of sight of each other and any point on the ground. The Landsat 4 and 5 satellites used the *tracking data relay satellite system* (*TDRSS*) to send images from essentially any point in LEO to a single ground receiving station by first relaying the images up to GEO and then down to the receiving station (cf. Fig. 2.22).

Satellite imaging of the earth developed hand in hand with the international space program. Early photos were recovered from V-2 launches in the late 1940s and early 1950s. Following the first manmade satellite launch of Sputnik 1 on October 4, 1957, photographs and video images were acquired by U.S. Explorer and Mercury programs and the Soviet Union's LUNA series. Just 3 years after Sputnik 1's debut (August 1960), the U.S. initiated its space-based reconnaissance program acquiring high-resolution photographic images from space (cf. McDonald, 1995). Figure 2.23 shows a photograph from the CORONA satellite program of a military target. This program of returning photographic images to earth using reentry vehicles called *film return buckets* continued into the 1970s. That same year (April 1960), the first U.S. orbiting EO system for regular monitoring of earth resources

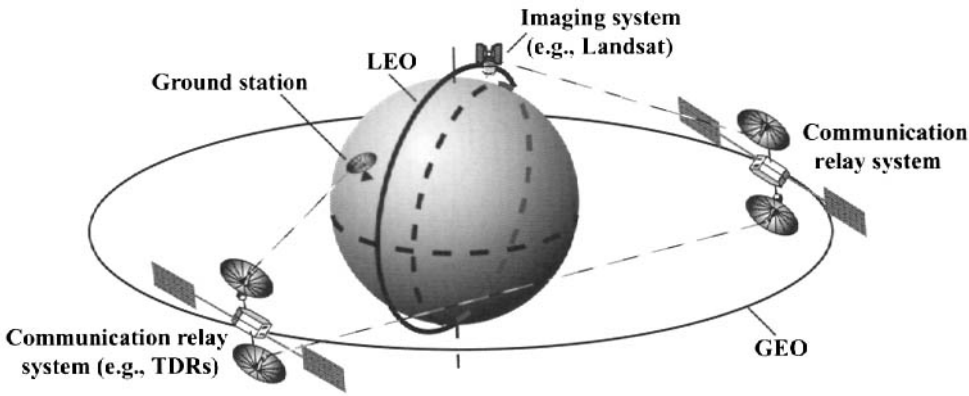


Figure 2.22 Satellites in geostationary orbit can be used to relay image data from LEO to a fixed ground station.

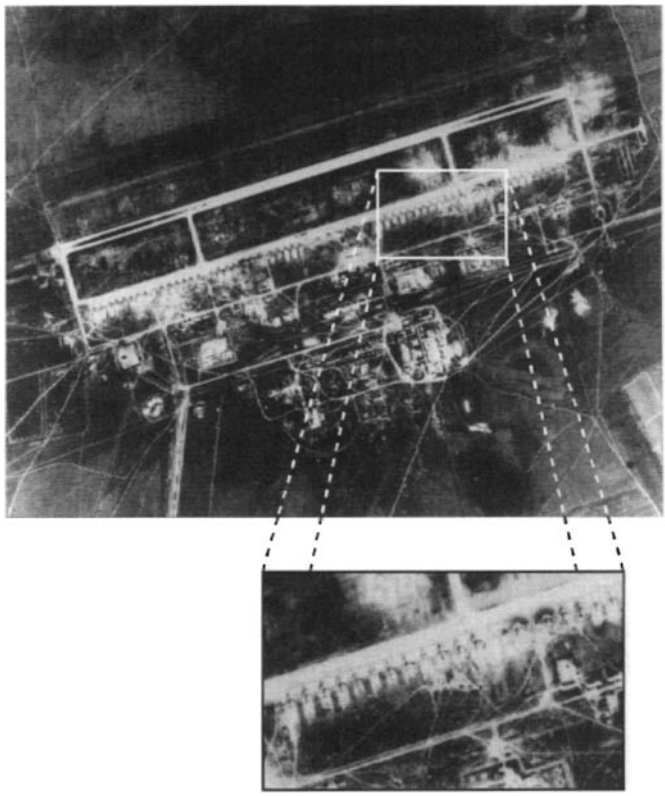


Figure 2.23 This August 20, 1966, photograph of a Soviet Union airfield represents a military target imaged by the CORONA program.

was launched into LEO. This first television and infrared observation satellite (TIROS) carried vidicon cameras that generated low-resolution images suitable for meteorological purposes. With time, a series of TIROS launches led to the improved TIROS operational satellites (ITOS-1 was launched in January 1970). These were called the NOAA (National Oceanic and Atmospheric Administration) series as they became operational and continued to evolve. Eventually, line scanner systems were added to the NOAA series payload and improved to include the current advanced very high-resolution radiometer (AVHRR), which has a 1 km ground spot at nadir and a wide field of view. NASA recently (1999) added the TERRA spacecraft to the list of satellites in LEO. In fact, TERRA, which carries the *moderate resolution image spectrometer* (*MODIS*) along with several other instruments, flies in formation with Landsat following Landsat 7 around the globe about one-half hour later. TERRA's sister spacecraft, AQUA, also carrying several sensors, including another MODIS, was launched in 2002 into a similar orbit but with an afternoon ascending equatorial crossing time. Each of the MODIS instruments images the earth every one or two days, providing moderate resolution 0.25 km to 1 km global coverage across many spectral bands (cf. Chap. 6). This program is continuing with the National Polar-Orbiting Operational Environmental Satellite System (NPOESS), which will continue to obtain global daily coverage of the earth with subkilometer spot sizes and many spectral bands. The Visible Infrared Imager Radiometer Suite (VIIRS) flown on NPOEES will make operational measurements of the globe in a fashion similar to, and motivated by, the MODIS research measurement programs.

2.5 DIGITAL CONCEPTS

With the evolution of digital computing in the 1960s and 1970s, we began to think of images in a new way. In its digital representation, an image is divided into a two-dimensional array of *picture elements* (*pixels*), with each pixel having an integer value associated with it that represents the image brightness at that location (cf. Fig. 2.24). Color or multiple band images can be thought of as being made up of layers of two-dimensional arrays (or three-dimensional arrays with the layers comprising the third dimension), with each layer representing brightness in a different spectral band.

Digital images would eventually offer great advantages because of their ease of manipulation in readily available digital computers. Even in the late 1960s and early 1970s when digital image processing was still in its infancy, recording and transmitting of images in digital form had major benefits. One of the foremost attractions was the value of maintaining the quantitative integrity of the signal during transmission over long distances (e.g., from space). Another attraction was that the digital images could be stored, retrieved, and reproduced many times, or after long periods of time, with no loss in content or quality. For these reasons, starting in the early 1970s, most satellite systems intended for quantitative analysis or for archival purposes began to convert the image to digital form for transmission and storage.

As the user community became more comfortable with digital image concepts, image data were increasingly distributed in digital form.

The Landsat series of satellites typify this evolution to digital image analysis. The first Landsat system originally called the Earth Resources Technology Satellite 1 (ERTS-1) was designed in the late 1960s and launched in 1972. It was the first of a series of operational sensors designed specifically to study the earth's land masses. To accomplish this, it had much better resolution than the operational meteorological systems that preceded it. The multispectral scanner (MSS) on Landsat 1 had an 80-meter spot size on the ground and acquired data in four spectral channels (cf. Chap. 6). The Landsat data were digitized on-board, and, after being transmitted and processed on the ground, they were archived in digital form. Much of the earliest use of the Landsat data was classical photo interpretation of conven-

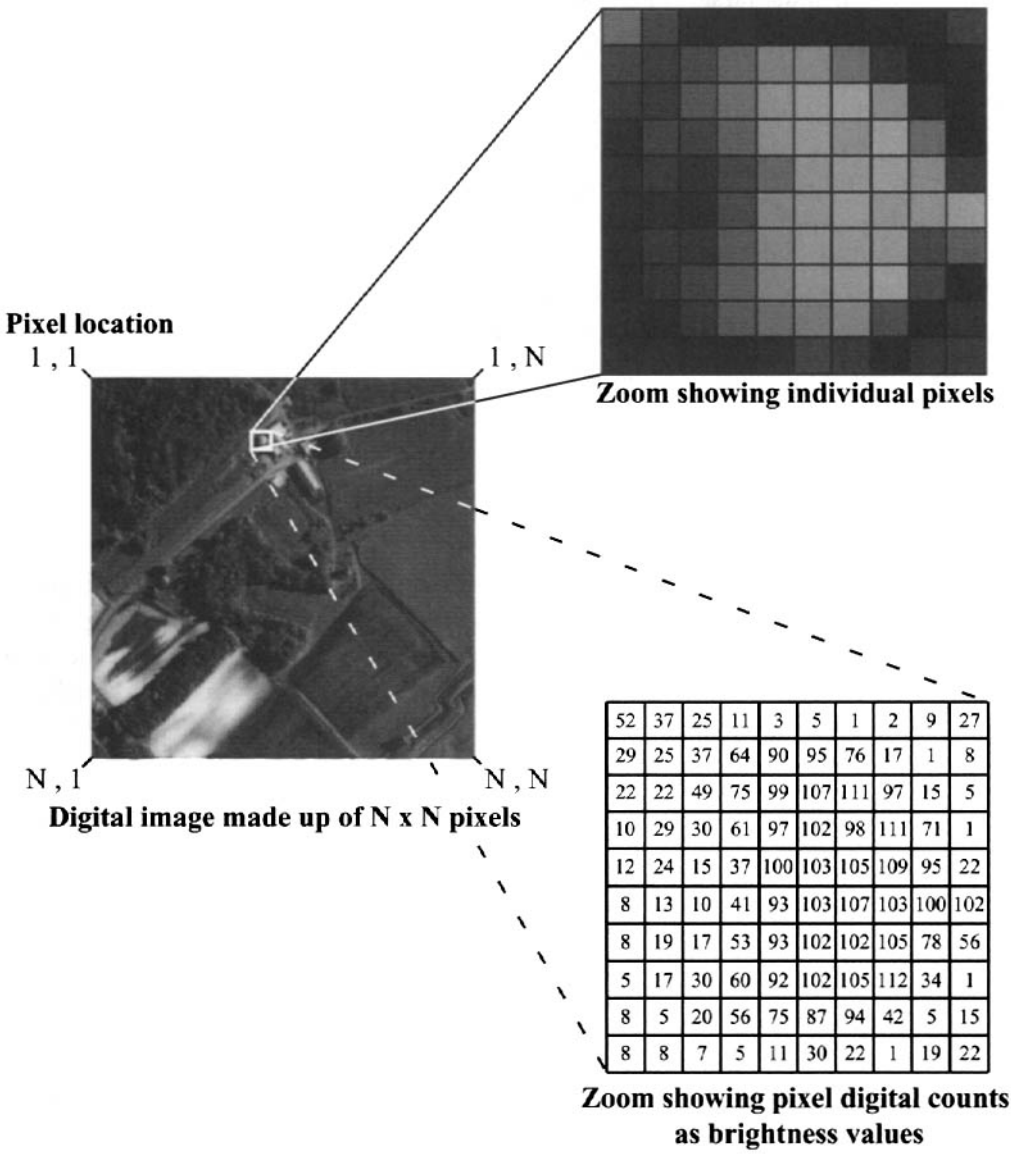


Figure 2.24 Digital image concepts.

tional photographic prints or transparencies made from the digital data. With the rapid evolution of digital computing, interpretation methods quickly changed, with more and more users performing their analysis on digital images produced from the archived records.

This archive of Landsat images contains records of the earth surface condition and how it has changed since 1972. The global perspective, repetitive history, and potentially quantifiable nature of these records are characteristic of the opportunity and challenge facing remote sensing today. How can we unlock and make available to scientists, decision makers, and the common man the wealth of information available from today's remotely sensed systems? Perhaps more importantly, can we decipher the host of archival data that can help us understand the changes taking place on our planet? The remainder of this volume focuses on the tools needed to acquire and analyze remotely sensed data in a manner designed to unlock this information storehouse.

2.6 REFERENCES

Estes, J.E., & Thorley, G.A., eds. (1983). *Vol. 2: Interpretation and Applications*. In Colwell, R.N., ed., *Manual of Remote Sensing*. 2nd ed., American Society of Photogrammetry, Falls Church, VA.

Estes, J.E., Hajic, E.J., & Tinney, L.R. (1983). *Analysis of Visible and Thermal Infrared Data.* In Colwell, R.N., ed., *Manual of Remote Sensing*, Vol. 2. 2nd ed., American Society of Photogrammetry, Falls Church, VA.

James, T.H. (1977). *The Theory of the Photographic Process*. 4th ed., Macmillan, NY.

Lillesand, T.M., Scarpace, F.L., & Clapp, J.L. (1973). "Photographic quantification of water quality in mixing zones." NASA-CR-137268.

McDonald, R.A. (1995). "Opening the cold war sky to the public: declassifying satellite reconnaissance imagery." *Photogrammetric Engineering and Remote Sensing*, Vol. 61, No. 4, pp. 385-390.

McGlone, J.C. ed. (2004). *Manual of Photogrammetry*. 5th ed., ASPRS, Bethesda, MD.

Piech, K.R., & Walker, J.E. (1971). "Aerial color analysis of water quality." *Journal of Surveying and Mapping Division, American Society of Civil Engineers*, Vol. 97, No. SUZ, pp. 185-197.

Piech, K.R., & Walker, J.E. (1974). "Interpretation of soils." *Photogrammetric Engineering*, Vol. 40, pp. 87-94.

Simonett, D.S., ed. (1983). *The Development and Principles of Remote Sensing*. In Colwell, R.N., ed., *Manual of Remote Sensing*. 2nd ed., American Society of Photogrammetry, Falls Church, VA.

Slama, C.G., ed. (1980). *Manual of Photogrammetry*. 4th ed., American Society of Photogrammetry, Falls Church, VA.

Sprechter, M.R., Needler, D., & Fritz, N.L. (1973). "New color film for water-penetration photography." *Photogrammetric Engineering*, Vol. 39, No. 4, pp. 359-369.

Stroebel, L., & Zakia, R., eds. (1993). *The Focal Encyclopedia of Photography*. 3rd ed., Focal Press, Boston, MA.

Wolf, P.R., & DeWitt, B.A. (2000). *Elements of Photogrammetry with Applications in GIS*. 3rd ed., McGraw-Hill, NY.

CHAPTER 3
RADIOMETRY AND RADIATION PROPAGATION

Radiometric image analysis involves a quantitative assessment of the energy or flux recorded by an imaging system and some form of reverse engineering to determine the parameters that controlled the energy levels observed. For example, we might want to know the reflectance of elements in a scene, their temperature, or the distribution of water vapor in the atmosphere above the scene. This chapter contains the basic radiometric principles needed to describe, and eventually analyze, the brightness levels in an image. Before beginning this quantitative review of the fundamental physics, we will conceptually describe the energy paths we may need to analyze in deciphering an image. This "big picture" look will help to clarify what radiometric principles we need to review and will set the stage for the detailed development of the governing equation for radiometric analysis. This treatment is simplified by neglecting the dependency of radiometric signals on spatial or temporal frequency. These dependencies are introduced in Chapters 5 and 13.

3.1 ENERGY PATHS

If we restrict ourselves to passive remote sensing in the 0.4 – 15μm spectral region, the energy paths can be conveniently divided into two groups. The first energy paths are those followed by radiation originating from the sun, while the second set of paths are those associated with radiation resulting from the temperature of objects other than the sun.

3.1.1 Solar Energy Paths

In trying to develop an understanding of radiation propagation and eventually an equation to describe the energy reaching an imaging system, it is useful to try to imagine all the ways electromagnetic energy, "light," could reach an airborne or satellite camera. To keep this mental game somewhat bounded, we should try to place an intuitive estimate on the relative amounts of energy. Terms that are two or three orders of magnitude smaller than the total will not be significant in most applications.

The most significant solar energy paths are illustrated in Figure 3.1. The path we think of first is for packets of energy or photons that originate from the sun, pass through the atmosphere, reflect off the Earth's surface, and propagate back through the atmosphere to the sensor (type A photons in Fig. 3.1). This is the path that lets us "see" sunlit objects with variations in reflectance providing the character or information content of the image. Another important path is followed by photons that originate at the sun, are scattered by the atmosphere, and are then reflected by the Earth to the sensor (type B photons). These photons make up what we commonly refer to as *skylight or sky shine.* If you imagine walking into the shadow of a building on a sunny day (effectively eliminating type A photons), you realize that there is less light, but still plenty by which to see and discriminate objects. Thus, it is clear that at least in the visible region there are typically fewer type B than type A photons, but that the number of type B photons is still significant. Our intuition also correctly tells us that the relative number of type B to A photons will increase as we move from clear to hazy to overcast skies. Another group of photons (type C) that our experience correctly tells us are important are those that originate at the sun and are scattered into the camera's line of sight without ever reaching the Earth. This haze, air light, or flare light, which we will come to call *upwelled radiance*, is what washes out the contrast when we look at air photos or out the window of an aircraft. Clearly, it is also a function of how "hazy" the atmosphere is and can be a very large contributor to the overall flux, in extreme cases (fog and thick cirrus clouds) completely overwhelming the flux reflected from the Earth. Piech and Schott (1974) indicate that in the visible region under clear-sky conditions, this upwelled radiance term can range from an equivalent reflectance of a few percent at low altitudes to 10% or more at higher altitudes and depends strongly on wavelength and atmospheric conditions. If the upwelled radiance were equivalent to a 10% reflector, then if we actually imaged a 10% reflector, half the energy would come from the upwelled radiance term. More to the point, the contrast between a 10% and a 5% reflector rather than being 2 to 1 would be 1.33 to 1.

Another possible photon path is illustrated with type G photons in Figure 3.1. These originate at the sun, propagate through the atmosphere, reflect from background objects, and are then reflected from the object of interest back through the atmosphere to the sensor. This is clearly a more convoluted path than we have considered thus far, involving multiple reflections or bounces of the photons. To attempt to assess the relative magnitude of this phenomenon, imagine a wall painted half black and half white, with grass growing up to it, as shown in Figure 3.2(a).

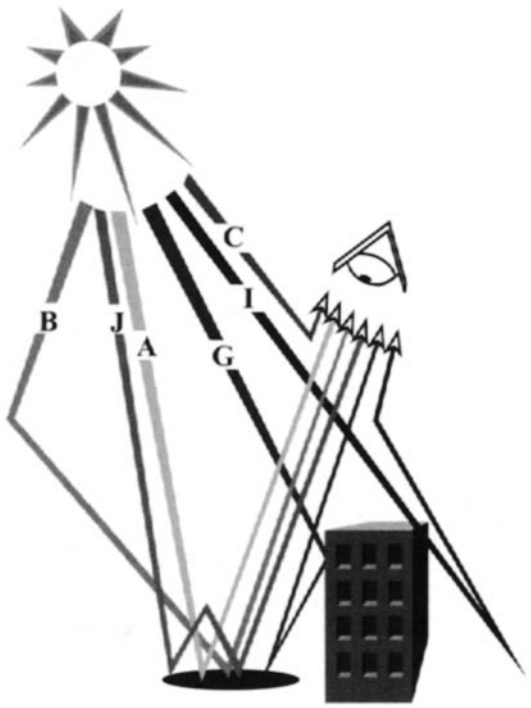
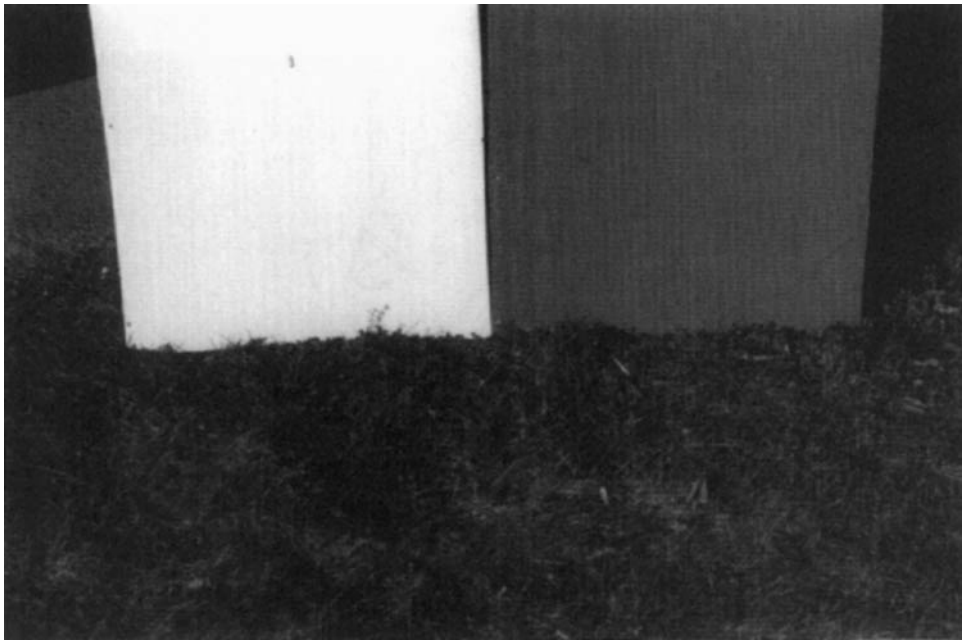


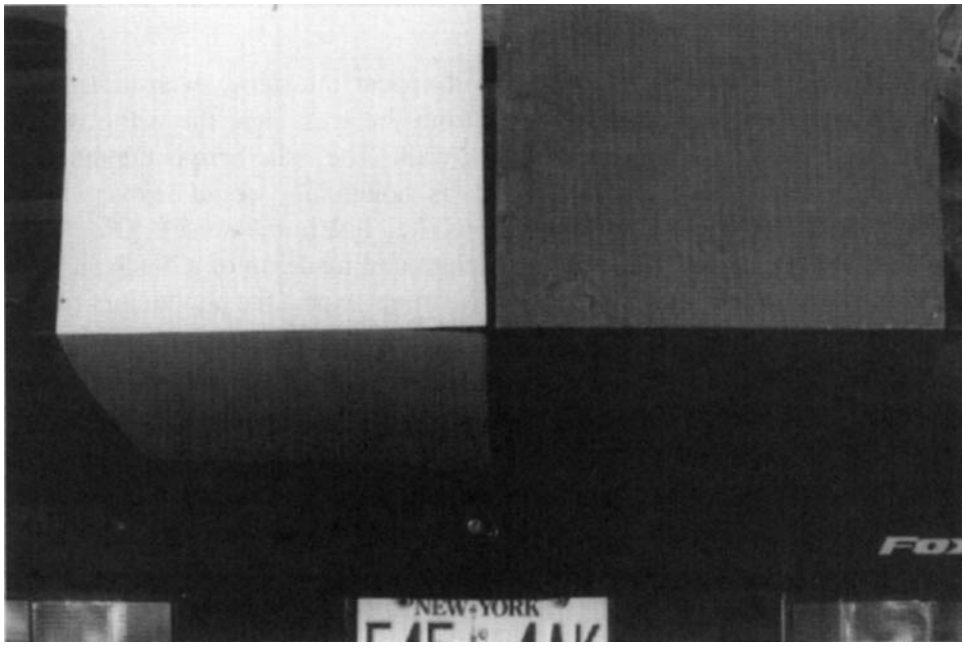
Figure 3.1 Solar energy paths.

Your intuition tells you that the grass will appear the same, even though there must be slightly more photons reflected from the grass near the white wall and fewer photons from the grass by the black wall. The issue here is the number of incremental photons. In this case, it seems as though they would be overwhelmed by the number of type A, B, and C photons (i.e., lost in the noise). On the other hand, consider the same high-contrast background made up of a black and white wall with a dark car parked in front of it. We might expect the photon flux from the background to change the appearance of the car. This is illustrated in Figure 3.2(b), where the combination of the dark car, the high surface reflectance due to the shiny (*specular*) finish, and the high contrast of the background combine to cause the difference in the number of multiple bounce photons to be large enough to affect the appearance (i.e., they are no longer lost in the noise). Thus, we see that the importance of multiple bounce photons depends on the sensitivity of our measurements (i.e., what is the noise level), the roughness of the surface, and the brightness or contrast of the background. The contrast between the background object and the sky is particularly important because the sky would be the source of photons from the background if the object were not present. For most natural (rough) surfaces in reasonably level terrain, the multiple bounce photons do not make a significant contribution and can and will be neglected. However, if specular targets or targets with a high background contrast are of interest, multiple bounce photons must be considered.

Another source of multiple bounce, or bounce and scatter photons (type I in Fig. 3.1), is caused by what has been termed the *adjacency effect* [cf. Kaufman



(a) Grass with black and white backgrounds.



(b) Car hood with black and white backgrounds.

Figure 3.2 Illustration of multiple bounce effects.

(1982)]. These are photons that are reflected from surrounding objects and then scattered into the line of sight of the sensor. In most cases [e.g., where the average background reflectance (*albedo*) is slowly varying], these photons can be treated as a constant and lumped in with the *path radiance* (type C photons). However, if a medium-gray object is imaged with a very dark or very bright background, there will be more photons in the bright background case due to the adjacency effect. (Note: Visually it will appear darker due to psychophysical effects.) Finally, when we are considering multiple scattering, we need to include the *trapping* effect illustrated by path J in Figure 3.1. These are photons reflected from the target into the air column above the target and then scattered (reflected) by the air column back onto the target and finally reflected toward the sensor. Conceivably, a photon could be "trapped" for many bounces between the target and the atmosphere, and all such possible combinations should be summed to account for the total number of photons reaching a sensor. In general, the relative number of these multiply bounced photons becomes important only as the atmosphere becomes relatively thick and multiple scattering becomes important. Since this condition is not uncommon at shorter wavelengths (i.e., the blue-green region or the VNIR under hazy conditions), we will introduce methods to describe this case.

In summary, for energy originating at the sun, we need to always be concerned with type A, B, and C photons. We will not be specifically concerned with type G, I, and J photons, although we should keep in mind that in a number of circumstances they may be important. In practice, in most empirical approaches, type J and G photons are lumped with type A and B photons, respectively, and type I photons are included with type C. Thus, although they are not explicitly addressed, they are at least to first order included in many of the analytical methods we will introduce for atmospheric compensation.

3.1.2 Thermal Energy Paths

Recall from elementary physics that all objects with temperatures above absolute zero radiate energy; therefore, we must also consider electromagnetic energy or photons from these sources. Our intuition is not as useful in understanding these photons, since in the visible region so few photons are emitted by anything other than the sun that they are completely negligible. However, in other portions of the spectrum (e.g., near 10 μm), self-emitted photons become very important, so we will want to consider energy paths that would lead these self-emitted photons to our sensor (cf. Fig. 3.3).

The photons most often of interest are caused by radiation due to the temperature of the target itself. These photons propagate through the atmosphere to the sensor, where their numbers provide information about the temperature of the target (type D photons in Fig. 3.3). Because the atmosphere above the target has some nonzero temperature, it radiates and scatters self-emitted energy down onto the target, which can be reflected and propagated up to the sensor (type E photons in Fig. 3.3). These photons are the thermal equivalent of skylight photons from the sun, and type B and E photons produce what is referred to collectively as *downwelled radiance*. The atmosphere along the line-of-sight path will also emit and

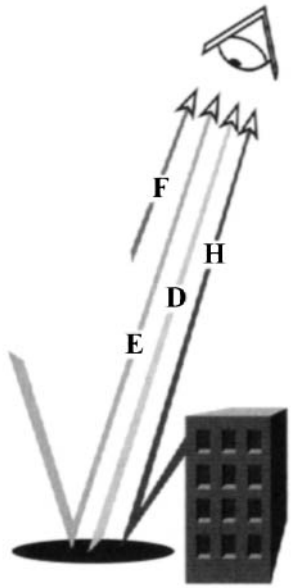


Figure 3.3 Self-emitted thermal energy paths.

scatter self-emitted photons directly to the sensor independent of the surface (type F photons). They are the thermal equivalent of the type C solar scattered photons, and type C and F photons produce the upwelled radiance or path radiance. Finally, background objects obscuring the sky above a target radiate energy due to their temperature and produce photons that can be reflected by the target and propagate to the sensor (type H photons in Fig. 3.3). These are similar to the multiple-bounce type G solar photons in that their relative importance will depend on the diffuseness (roughness) of the target and the brightness temperature of the background. For most natural surfaces without large background objects, type H photons are negligible.

A more quantitative analysis (cf. Sec. 4.6) will show that in the MWIR and LWIR spectral regions there are significant numbers of type D and F photons (more D than F) and smaller numbers of type E and H photons, which, depending on circumstances, may or may not be negligible.

Assessment of the radiation propagation paths shown in Figures 3.1 and 3.3 suggests that in order to describe the radiometry quantitatively, we need to characterize self-radiating sources, propagation of energy from the sun to the Earth, propagation through the atmosphere (e.g., transmission losses), scattering in the atmosphere, reflection by the earth, and the radiometric characteristics of radiation propagation within the sensor. The rest of this chapter deals with the tools needed to characterize the radiometric strand in the image chain. These tools are then applied in Chapter 4 to describe quantitatively the first links in the image chain.

3.2 RADIOMETRIC TERMS

Radiometry is formally defined as the science of characterizing or measuring how much EM energy is present at, or associated with, some location or direction in space. It has evolved separately in the fields of physics, illumination or vision, and engineering, and as a result a host of terms are used to describe various radiometric concepts. Often one concept has several different names, and it is also common for the same term (e.g., intensity) to mean different things to different authors. To provide a common framework, we will briefly review the definitions of the relevant physical parameters and radiometric terms. An emphasis will be placed on the units of measure in this section and throughout the book to ensure a clearer understanding—units are usually designated with square brackets ([ ]) for clarity, and, where relevant, a unit's cancellation analysis may be performed within square brackets. In reading other authors, particularly older works or work drawn from other disciplines, the reader should carefully evaluate the author's definitions and units of measure to determine what term is being applied to each radiometric concept. The definitions used throughout this volume are consistent with those established by the *Commission Internationale de l'Eclairage* (*CIE*) and adopted by most international societies [see CIE (1970)]. In addition, to the extent practical, the parameters, nomenclature, and symbology are consistent with the relevant reference material [see Grum and Becherer (1979) and Nicodemus (1976)].

3.2.1 Definition of Terms

For most radiometric considerations, we can use the ray/particle simplification of optics. This approach is based on geometric optics and assumes that light travels in straight lines and transfers energy in discrete packets or quanta. The physical optics effects of diffraction and interference associated with the wave nature of EM energy can be largely ignored in simple radiometric calculations in the visible and thermal infrared. The wave nature of EM energy is important in image formation but need not be considered in most simple radiometric calculations.

Recall that wavelength (λ [μm]), frequency (ν [sec^{-1}]), and the speed of light (c [m/sec]) are related as

$$c = \lambda\nu \tag{3.1}$$

where *wavelength* is the distance between two similar consecutive elements (same phase) in a wave representation (e.g., peak-to-peak or trough-to-trough). It is commonly referred to in units of microns [μm = 10^{-6} m] or nanometers [nm = 10^{-9} m]. *Frequency* is the number of waves (cycles) that would travel past a fixed point in 1 second and has units of hertz, cycles per unit time [Hz], or [sec^{-1}] The *speed of light* clearly has units of distance per second and in vacuum has a constant value of $2.9979 \cdot 10^8$ m/sec. Spectral references will generally be given with respect to wavelength. However, some computations are more readily represented by the *wave number* ν' [cm^{-1}], which is simply the number of waves that would fit in a 1-centimeter length, i.e.,

$$v' = \frac{1}{\lambda}[\text{cm}^{-1}] = \frac{v}{c} \tag{3.2}$$

The common spectral regions and the nomenclature used in this volume are delineated in Figure 3.4. In radiometric calculations, it is generally easier to think of energy as being transferred in terms of energy packets or quanta in accordance with quantum theory. The particle or energy carrier is called a *photon*, and each photon carries *energy*

$$q = hv = hc/\lambda[\text{joules}] \tag{3.3}$$

where $h = 6.6256 \cdot 10^{-34}$ [joules·sec] is Planck's constant, and energy is expressed as joules [j]. Thus, we see that shorter wavelength photons carry more energy than longer wavelength photons. This becomes very important when we begin to look at the spectral response of detectors in Chapter 5. The total energy (Q) in a beam or ray is a function of the number and spectral makeup of the photons according to

$$Q = \sum q_i = \sum_{i=1} n_i h v_i \tag{3.4}$$

where the sum is over all frequencies present and n_i is the number of photons at each frequency.

It is usually more convenient to think of a beam or bundle of rays not in terms of the total energy but rather in terms of the rate at which the energy is passing or propagating [cf. Fig. 3.5(a)]. This rate of flow of energy is called the *radiant flux, or power* (Φ), and is defined as the first derivative of the radiant energy with respect to time (t), i.e.,

$$\Phi = \frac{dQ}{dt}[\text{watts, W}] \tag{3.5}$$

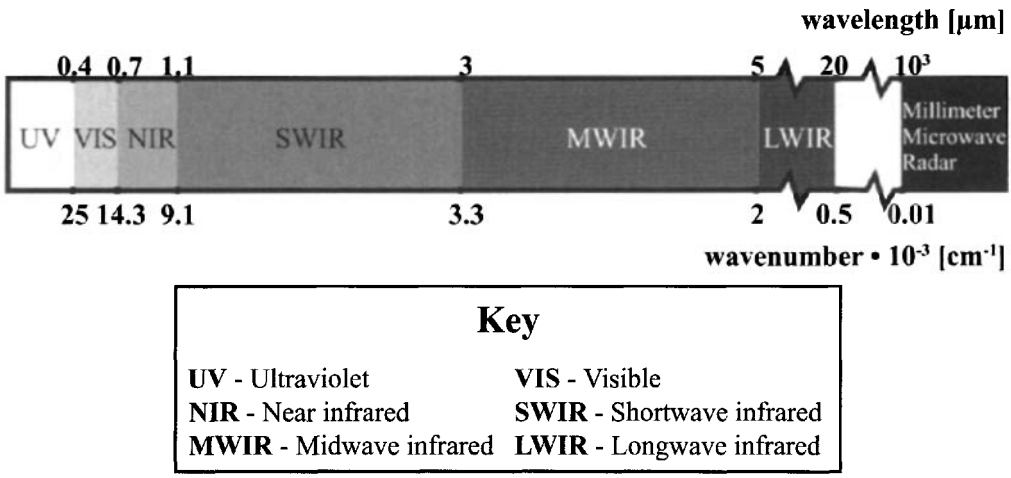


Figure 3.4 Nomenclature for various regions of the electromagnetic spectrum.

Often we are interested in the rate at which the radiant flux is delivered to a surface (e.g., the responsive surface of a detector). This concept is given the term *irradiance* (E) and is defined as

$$E = E(x,y) = \frac{d\Phi}{dA} [\mathrm{Wm}^{-2}] \tag{3.6}$$

where dA [m^2] is an area element on the surface of interest, and (x,y) are generic spatial location parameters that, for convenience, will generally not be explicitly expressed. [Eq. (3.6) is characteristic of a shorthand we will use to indicate a simplification of notation where E and $E(x,y)$ are identical, but the dependence on x and y will only be explicitly stated where it is required for clarity.] Irradiance, as illustrated in Figure 3.5(b), is the flux per unit area *onto* a surface. It is very similar to radiant exitance, which is defined as

$$M = M(x,y) = \frac{d\Phi}{dA} [\mathrm{Wm}^{-2}] \tag{3.7}$$

and describes the flux per unit area *away from* a surface [cf. Fig. 3.5(c)]. This term describes the power per unit area radiated by a source or reflected from a surface.

Both the irradiance and the exitance provide spatial information about the flux, but no angular or directional information. The simplest term used to describe directional or dispersive information about the flux is the *radiant intensity* (I), defined as:

$$I = I(\theta,\phi) = \frac{d\Phi}{d\Omega} [\mathrm{Wsr}^{-1}] \tag{3.8}$$

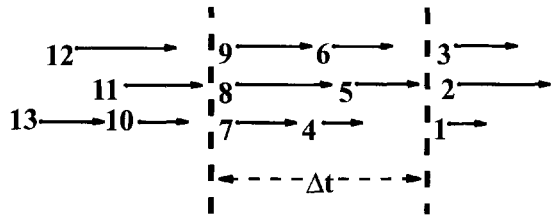
where $d\Omega = dA/r^2$ [*steradian*, sr] is the element of *solid angle*. The element of solid angle is defined as the conic angle encompassing the area element dA on the surface of a sphere of radius r [cf. Fig. 3.5(d)], and θ and ϕ are generic orientation angles, as illustrated in Figure 3.5(e), which will not be explicitly expressed unless required.

The radiant intensity describes the flux per unit solid angle from a point source into a particular direction. While the intensity provides directional information, it does not provide any spatial information. The use of the *radiance* term (L [$\mathrm{wm}^{-2}\mathrm{sr}^{-1}$]) to characterize the flux overcomes this limitation. It is the most complex of the radiometric terms we will consider, but also the most useful and ubiquitous. It is defined as

$$L = L(x,y,\theta,\phi) = \frac{d^2\Phi}{dA\cos\theta\, d\Omega} = \frac{dE}{d\Omega\cos\theta} = \frac{dI}{dA\cos\theta} = \frac{dM}{d\Omega\cos\theta} \tag{3.9}$$

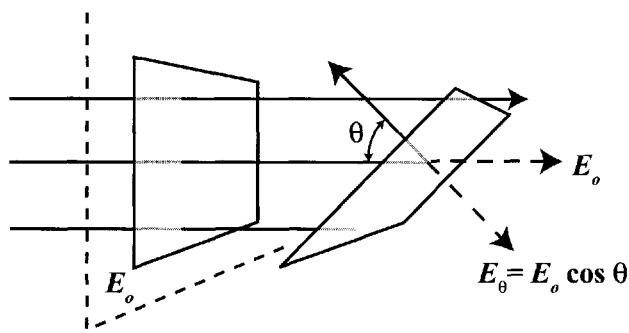
where x and y define a location in the plane of interest, and θ and ϕ are angles that define the direction of interest relative to the normal to the plane. The radiance is the flux per unit projected area (at the specified location in the plane of interest) per unit solid angle (in the direction specified relative to the reference plane). Note that while radiant exitance and intensity are generally source terms and irradiance is generally associated with receivers or detectors, radiance can be used to characterize the flux from or onto a surface, as well as the flux through any arbitrary

$$\Phi = \frac{dQ}{dt} = \frac{\sum_{i=4}^{9} q_i}{\Delta t}$$



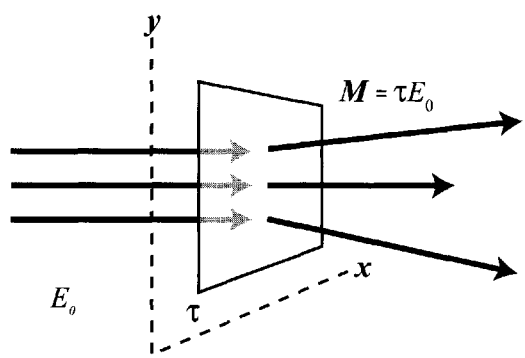
(a) Radiant flux: time rate of energy delivery, production, or propagation.

$$E = \frac{d\Phi}{dA}$$



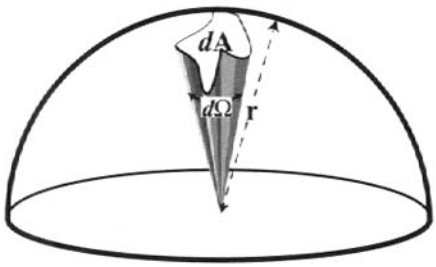
(b) Irradiance: flux per unit area onto a surface. (The first surface is perpendicular to the incident flux and has irradiance E_0. The second is rotated through the angle θ and has a flux E_θ.)

$$M = \frac{d\Phi}{dA}$$



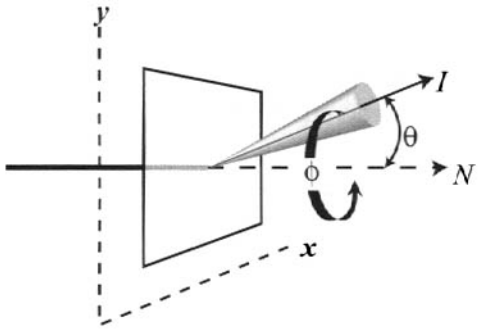
(c) Radiant emitance or radiant exitance: flux per unit area away from the surface. (A surface with E_0 irradiance from the left and transmission τ would have exitance M away from the righthand side.)

Figure 3.5 Illustration of radiometry definitions.



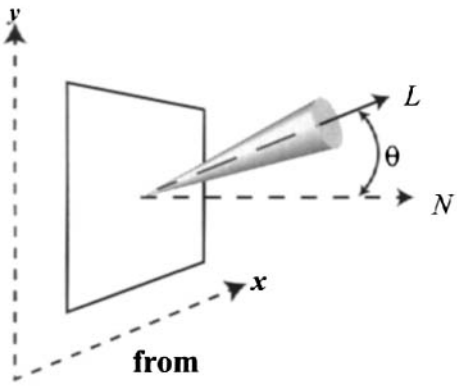
$$d\Omega = \frac{dA}{r^2}$$

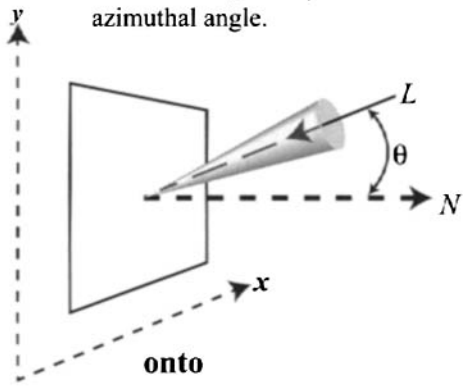
(d) Element of solid angle [steradian].

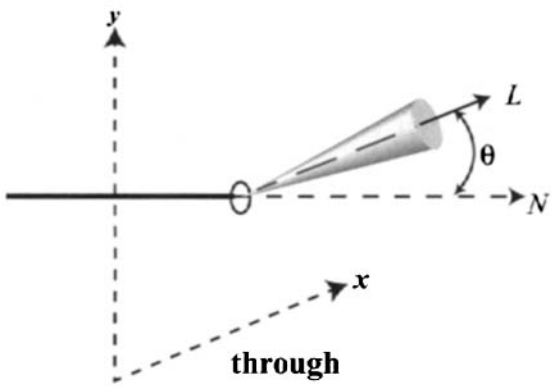


$$I = \frac{d\Phi}{d\Omega}$$

(e) Radiant intensity: flux per unit solid angle into the direction defined by θ and ϕ, where θ is the angle from the normal to a reference surface, and ϕ is an azimuthal angle.







$$L = \frac{d^2\Phi}{dA\, d\Omega\, \cos\theta}$$

(f) Radiance: flux per unit projected area per unit solid angle from, onto, or through the plane. N is the normal to the plane.

Figure 3.5 continued. Illustration of radiometry definitions.

surface in space [cf. Fig. 3.5(f)]. In addition, it has some very useful properties of constancy of propagation that make it an attractive parameter to use in most treatments of radiation propagation.

Nicodemus (1976) demonstrates the important concept of the constancy of radiance through an isotropic lossless media (i.e., no transmission losses and unit index of refraction). Referring to Figure 3.6, we assume a beam of energy with constant radiance across the profile of the beam. We select two arbitrary points along the beam and two surfaces with arbitrary orientation containing those points. It may be convenient to conceptualize the first surface as a source (i.e., the Earth's surface) and the second as a sensor. If we consider the flux associated with a bundle of rays at surface 1 contained in a surface element dA_1, which are also contained in dA_2 on surface 2, we see that in a lossless media, the flux $d\Phi_1$ through dA_1 must equal the flux $d\Phi_2$ through dA_2. We want to evaluate how the radiance at surface 1 (L_1) is related to the radiance at surface 2 (L_2). We see that the radiance L_1 along the primary ray at p_1 toward the surface element dA_2 is, by definition,

$$L_1 = \frac{d^2\Phi_1}{dA_1 \cos\theta_1 d\Omega_{12}} \tag{3.10}$$

and the radiance at p_2 is

$$L_2 = \frac{d^2\Phi_2}{dA_2 \cos\theta_2 d\Omega_{21}} \tag{3.11}$$

where θ_1 and θ_2 are the angles from the ray normal to the surface to the primary ray, and $d\Omega_{12}$ is the element of solid angle encompassed by the area element dA_2 at p_1, and similarly $d\Omega_{21}$ is the element of solid angle encompassed by the area element dA_1 at p_2. If we let r represent the arbitrary distance between p_1 and p_2, we see that the throughput $\mathcal{T}$ can be expressed as

$$\mathcal{T}_1 = dA_1 \cos\theta_1 d\Omega_{12} = dA_1 \cos\theta_1 \frac{dA_2 \cos\theta_2}{r^2}$$

and

$$\mathcal{T}_2 = dA_2 \cos\theta_2 d\Omega_{21} = dA_2 \cos\theta_2 \frac{dA_1 \cos\theta_1}{r^2} \tag{3.12}$$

and that $\mathcal{T}_1 = \mathcal{T}_2 = \mathcal{T}$. Expressing the radiance as

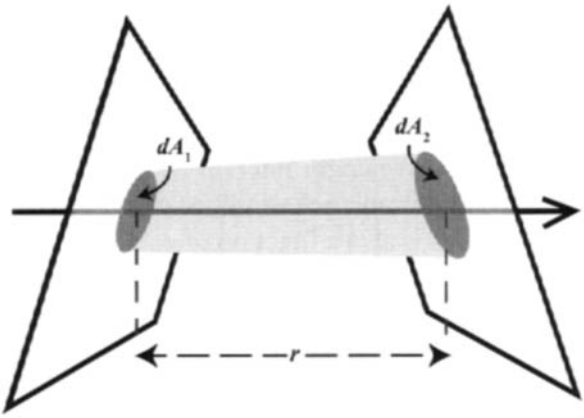
$$L_1 = \frac{d^2\Phi_1}{\mathcal{T}_1} \tag{3.13}$$

and

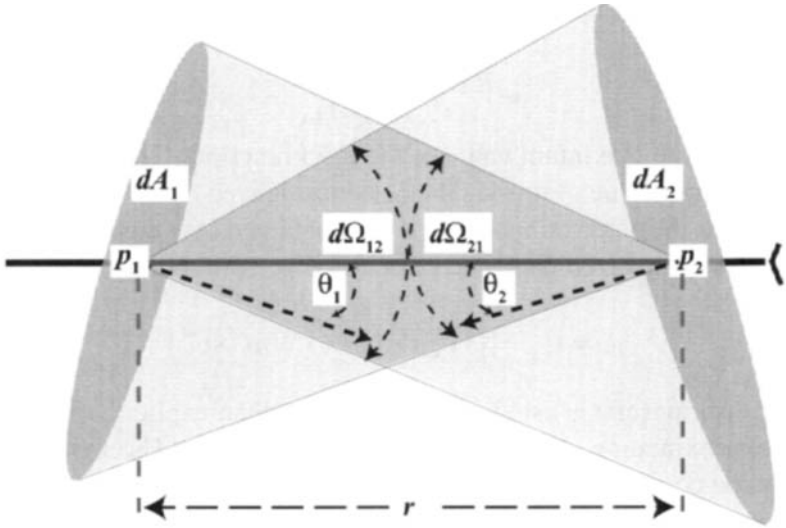
$$L_2 = \frac{d^2\Phi_2}{\mathcal{T}_2} \tag{3.14}$$

and recalling that $d\Phi_1 = d\Phi_2 = d\Phi$, we have

$$L_1 = L_2 = \frac{d^2\Phi r^2}{dA_1 \cos\theta_1 dA_2 \cos\theta_2} = \frac{d^2\Phi}{\mathcal{T}} \tag{3.15}$$



(a) Radiance along a beam.



(b) Illustration of the terms used in the definition of radiance at the points p_1 and p_2 along a beam.

Figure 3.6 Constancy of radiance.

Since all the terms in this analysis were completely arbitrary, we see that the radiance along a ray is constant over distance in a lossless media. Thus, it is the term most readily used for radiation propagation, since it is independent of geometric considerations and only losses due to the medium need to be considered (e.g., absorption and scattering).

To this point, we have ignored the spectral character of the radiometric terms. In fact, the flux is spectrally variable, and therefore, each of the radiometric terms will vary with wavelength. In general, we will be interested in *spectral density* expressed as flux per unit wavelength interval and designated with a wavelength subscript. Thus, the spectral irradiance would be expressed as E_λ [$Wm^{-2}\mu m^{-1}$]. The responsivity of the detectors is also a function of wavelength and must be cascaded with the spectral flux to generate effective bandpass values for the radiometric terms (i.e., what is the effective magnitude of the radiometric term relative to the spectral response of the detector?). The *responsivity* at each wavelength is defined as the signal out (S) per unit flux incident on the detector at the wavelength of interest. Therefore, the *spectral response* function is defined as

$$R(\lambda) = \frac{dS}{d\Phi(\lambda)} \tag{3.16}$$

with units of [amps W^{-1}] or [volts W^{-1}], depending on the signal out of the detector. The unitless peak normalized spectral response function is

$$R'(\lambda) = \frac{R(\lambda)}{R(\lambda)_{max}} \tag{3.17}$$

where $R(\lambda)_{max}$ is the maximum value of the $R(\lambda)$ function. Thus, $R'(\lambda)$ is normalized to a maximum value of unity. The peak normalized *effective value* of a radiometric term over the detector bandpass is then obtained by weighting the radiometric term by this normalized response value (cf. Fig. 3.7). For example,

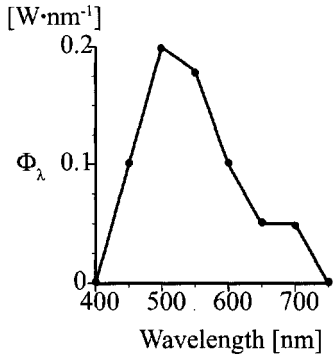
$$L = L_{eff} = \int_0^\infty L_\lambda R'(\lambda) d\lambda \ [Wm^{-2}sr^{-1}] \tag{3.18}$$

where the subscript (eff) is usually implied, rather than explicitly indicated, and a numerical approximation to the integral is used in practice. The signal output from a sensor can be computed by integrating the spectral flux weighted by the spectral response function according to

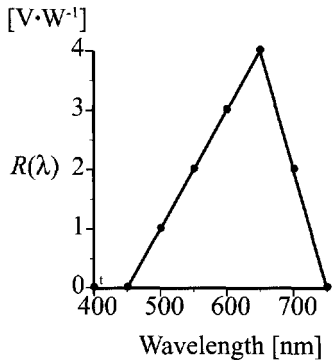
$$S = \int_0^\infty \Phi_\lambda R(\lambda) d\lambda \tag{3.19}$$

where the output signal (S) has units of amps or volts depending on the type of detector, and the integral or its numerical approximation need only be performed over the nonzero spectral response range. It is convenient to define the effective bandpass responsivity (R) such that the product of the effective responsivity and the total flux yields the observed signal, i.e.,

$$R\Phi_{tot} = R\Phi = S = \int_0^\infty \Phi_\lambda R(\lambda) d\lambda \tag{3.20}$$



	Given		Calculated	
λ [nm]	Φ_λ [Wnm^{-1}]	$R(\lambda)$[V/W]	$R'(\lambda)$	$\Phi_\lambda/\int\Phi_\lambda d\lambda$
400	0	0	0	0
450	0.1	0	0	2.94×10^{-3}
500	0.2	1	0.25	5.88×10^{-3}
550	0.18	2	0.5	5.29×10^{-3}
600	0.1	3	0.75	2.94×10^{-3}
650	0.05	4	1	1.47×10^{-3}
700	0.05	2	0.5	1.47×10^{-3}
750	0	0	0	0

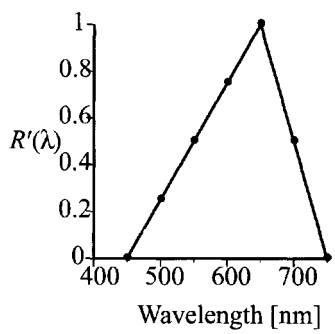


$$\Phi = \int\Phi_\lambda d\lambda = \Sigma\Phi_\lambda\Delta\lambda = 34 \text{ [W]}$$

$$\Phi_{eff} = \int\Phi_\lambda R'(\lambda)d\lambda = \Sigma\Phi_\lambda R'(\lambda)\Delta\lambda = 14.5 \text{ [W]}$$

$$R = \int R(\lambda)\frac{\Phi_\lambda}{\int\Phi_\lambda d\lambda}d\lambda = \Sigma R(\lambda)\frac{\Phi_\lambda}{\Sigma\Phi_\lambda d\lambda}\Delta\lambda = 1.7 \text{ [V/W]}$$

$$S = \int \Phi_\lambda R(\lambda)d\lambda = 58 \text{ [V]}$$



$$S = R\Phi \cong 58 \text{ [V]}$$

$$\Phi_{\lambda eff} = \int\Phi_\lambda R(\lambda)d\lambda / \int R(\lambda)d\lambda = 58[\text{v}]/600 \text{ [VW}^{-1}\text{nm]}$$
$$= 0.097 \text{ [Wnm}^{-1}\text{]}$$

Figure 3.7 Effective flux and responsivity sample calculations.

Substituting the integral form of Φ_{tot} and rearranging yields an expression for the *effective responsivity* of

$$R = \frac{\int_0^\infty \Phi_\lambda R(\lambda) d\lambda}{\int_0^\infty \Phi_\lambda d\lambda} \tag{3.21}$$

It is important to recognize that when the bandpass value for responsivity is used, it is calculated for a specific source spectral distribution. The same detector will exhibit different bandpass responsivity values when irradiated by sources with differing spectral distributions.

In dealing with imaging spectroscopy, it is often useful to use effective spectral radiometric terms. This is the constant value the radiometric term would have to have over the sensor bandpass to generate the observed signal. For *effective spectral radiance*, this would require the following expression to be true:

$$S = \int L_\lambda R(\lambda) A\Omega \cos\theta d\lambda = L_{\lambda\text{eff}} \int R(\lambda) A\Omega \cos\theta d\lambda \tag{3.22}$$

yielding an expression for the effective spectral radiance of

$$L_{\lambda\text{eff}} = \frac{\int L_\lambda R(\lambda) d\lambda}{\int R(\lambda) d\lambda} = \frac{\int L_\lambda R'(\lambda) d\lambda}{\int R'(\lambda) d\lambda} = \frac{L_{eff}}{\Delta\lambda_{eff}} [\text{Wm}^{-2} sr^{-1} \mu\text{m}^{-1}] \tag{3.23}$$

where $\Delta\lambda_{\text{eff}}$ is sometimes referred to as the effective spectral bandwidth. Similar expressions can be derived for all the radiometric terms. In general, for narrow spectral bands and spectrally smooth equations, the effective spectral values are good estimates of the actual spectral values.

3.2.2 Blackbody Radiators

One of the cornerstones of modern physics and a critical element of quantitative radiometric remote sensing is the formula for spectral exitance from a blackbody radiator. A *blackbody* is an idealized surface or cavity that has the property that all incident electromagnetic flux is perfectly absorbed and then reradiated (i.e., the reflectivity is zero and absorptivity is one). Planck (1901) derived an expression for the spectral radiant exitance from a blackbody based on statistical calculation of the vibrational energy states between the atoms and the assumption that the vibrational resonation between the atoms could only emit or absorb energy in discrete levels proportional to the frequency of the oscillation state. Thus, all the energy states are defined by $Q = mh\nu$, where m can only take on integer values, h was an empirically derived value that we now refer to as *Planck's constant*, and ν is the frequency of oscillation. Einstein's later work on the quantum theory of light and the concept of the photon provided the theoretical foundation for Planck's results [cf. Einstein (1905)]. The Planck or blackbody radiation equation for the spectral radiant exitance from a surface is

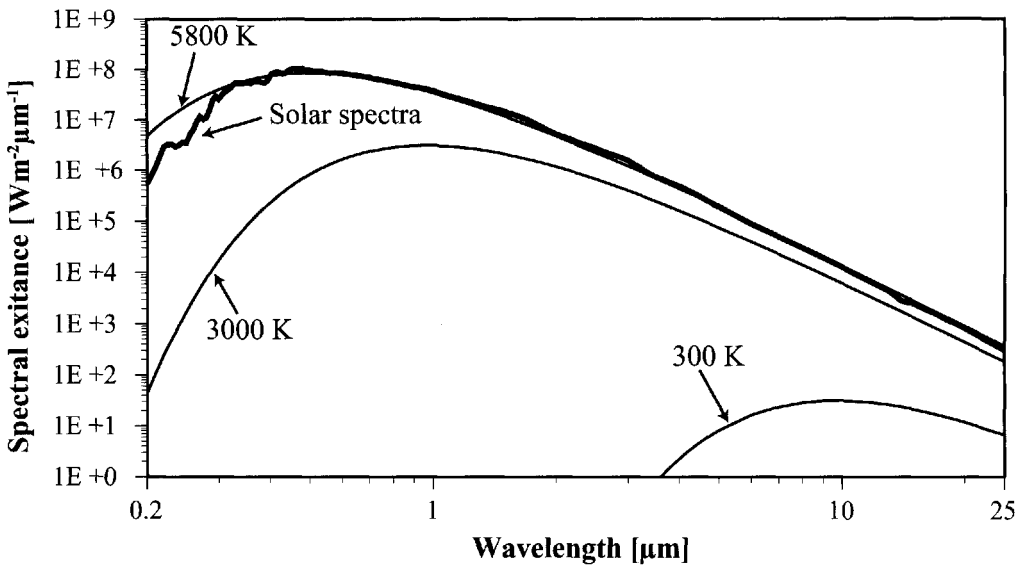


Figure 3.8 Blackbody curves and solar exitance spectra.

$$M_\lambda = 2\pi hc^2\lambda^{-5}\left(e^{\frac{hc}{\lambda kT}} - 1\right)^{-1} [Wm^{-2}\mu m^{-1}] \tag{3.24}$$

where T is the temperature in degrees Kelvin, k is the Boltzmann gas *constant* ($1.38 \cdot 10^{-23}$ jK^{-1}), and h and c are the Planck constant and the speed of light as previously defined. Examination of the Planck equation shows that radiant exitance is a function of both temperature and wavelength. By holding temperature fixed at selected values, a family of blackbody curves can be generated as shown in Figure 3.8 relating spectral exitance to wavelength. These curves show how the exitance increases with temperature and that it is a well-behaved function whose peak value shifts to shorter wavelengths as the temperature increases. In practice, the ideal blackbody can only be approximated by imperfect absorbers. To describe this phenomenon, we introduce the concept of *emissivity,* $\varepsilon(\lambda)$, defined as the ratio of the spectral exitance $M_\lambda(T)$ from an object at temperature T to the exitance from a blackbody at that same temperature $M_{\lambda BB}(T)$, i.e.,

$$\varepsilon(\lambda) = \frac{M_\lambda(T)}{M_{\lambda BB}(T)} \tag{3.25}$$

The emissivity describes how well an object radiates energy compared to the perfect blackbody radiator and is a unitless value with a range from 0 to 1. Objects whose emissivity is approximately constant with wavelength are referred to as *gray bodies*, while objects with spectrally varying emissivities are called *selective radiators*. Objects that approximate *gray-body* radiators over all or part of the spectrum are often described or approximated by a blackbody that would produce the equivalent exitance. The exitance from the sun can be approximated by a blackbody at approximately 5800 K (cf. Fig. 3.8).

It is important to recognize that emissivity is a fundamental property of matter just like absorptivity, reflectivity, and transmissivity. In the formalism we have introduced, the *transmissivity* is the ability of the material to allow the flux to propagate through it. The *transmittance* or *transmission* τ can be expressed as the unitless ratio of the exitance from the back of a sample (M_τ) to the irradiance on the front of the sample (E_i), i.e.,

$$\tau = \frac{M_\tau}{E_i} \tag{3.26}$$

Clearly, the spectral transmittance $\tau(\lambda)$ is simply the ratio of the spectral exitance and spectral irradiance. Similarly, the *reflectivity* is the ability of the material to turn incident flux back into the hemisphere above the material, and the *reflectance* r can be expressed as the ratio of the exitance from the front of a sample (M_r) to the irradiance onto the front of the sample, i.e.,

$$r = \frac{M_r}{E_i} \tag{3.27}$$

Finally, the *absorptivity* is the ability of the material to remove electromagnetic flux from the system by converting incident flux to another form of energy (e.g., thermal energy). The *absorptance* α can be expressed as the ratio of the flux per unit area incident on the surface that is converted to another form of energy (M_α) to the irradiance onto the surface, i.e.,

$$\alpha = \frac{M_\alpha}{E_i} \tag{3.28}$$

Since conservation of energy requires all the incident flux to be absorbed, transmitted, or reflected, we have

$$\alpha + \tau + r = 1 \tag{3.29}$$

or in the case of an opaque material, where τ is zero, we have

$$\alpha + r = 1 \tag{3.30}$$

Furthermore, according to Grum and Becherer (1979), *Kirchoff's law* states that the emissivity must be numerically equal to the absorptance for surfaces in thermodynamic equilibrium (i.e., good absorbers are good emitters). Therefore, we can also express the conservation of energy relationship as

$$\varepsilon + \tau + r = 1 \tag{3.31}$$

or for opaque objects

$$\varepsilon + r = 1 \tag{3.32}$$

We can compute the total exitance from a blackbody by integrating the Planck equation over all wavelengths. This yields the familiar Stefan-Boltzmann equation, i.e.,

$$M = \int_0^\infty M_\lambda d\lambda = \int 2\pi hc^2 \lambda^{-5} (e^{\frac{hc}{\lambda kT}} - 1)^{-1} d\lambda$$

$$= \frac{2\pi^5 k^4 T^4}{15c^2 h^3} = \sigma T^4 \tag{3.33}$$

where σ ($5.67 \cdot 10^{-8} \mathrm{Wm^{-2}K^{-4}}$) is the Stefan-Boltzmann constant. It is important to recognize that this fourth-power relationship holds only for the integral over all wavelengths and is mostly of use for energy exchange calculations in thermodynamics. The exitance within a bandpass can be expressed as

$$M = \int_{\lambda_1}^{\lambda_2} M_\lambda d\lambda \tag{3.34}$$

and must be solved in numerical form since no closed form solution exists.

Another fundamental natural law can be derived from the Planck equation by taking the first derivative with respect to wavelength and setting it equal to zero, i.e.,

$$\frac{dM_\lambda}{d\lambda} = 0 \tag{3.35}$$

Since we have already seen that the Planck equation is well behaved with a single maximum, the zero point in the first derivative will yield the wavelength of maximum exitance. Solving Eq. (3.35) and rearranging produces the Wien displacement law, i.e.,

$$\lambda_{\max} = \frac{A}{T} \tag{3.36}$$

where A (2898 μm•K) is the Wien displacement constant. This expression predicts that the peak radiance from the sun at approximately 6000 K will occur in the visible portion of the spectrum at approximately 0.5 μm and that the peak flux for an object near the earth's ambient temperature of 300 K will occur at approximately 10 μm (cf. Fig. 3.8). This is conveniently in the center of an atmospheric transmission window (cf. Fig. 1.5), which is extensively used for studying the thermal characteristics of the earth.

3.2.3 Polarization Concepts

Recall that electromagnetic energy can be represented as two mutually perpendicular electric ($\mathcal{E}$) and magnetic ($\mathcal{H}$) waves traveling in a direction z perpendicular to both (cf. Fig. 3.9). Randomly polarized EM energy is composed of many superimposed waves whose electric fields vary so rapidly in orientation that no preferential orientation can be discerned [cf. Hecht (1990) for a more complete treatment of *polarization* principles]. In contrast, for a plane polarized wave, the electric (and therefore also the magnetic) field only varies in a single plane. More generally, due to phase differences in the constituent waves, the plane of polarization may rotate over time and the amplitude may vary. This results in the general case of ellipti-

cally polarized radiation. The amplitude of the electric field vector appears to trace out an ellipse when viewed backward along the direction of propagation (cf. Fig. 3.10). Circular and plane polarization can be thought of as special cases of elliptical polarization. For circular polarization, the amplitude remains constant while the orientation changes, and for linear polarization, the magnitude varies while the orientation remains fixed.

In most cases, EM radiation can be thought of as a combination of randomly polarized radiation and polarized radiation. If our sensor is insensitive to polarization (i.e., responds to all polarizations equally), then we must simply deal with the aggregate average flux. On the other hand, if our sensor, either by design or by happenstance, is polarization sensitive, then we will need to be concerned about the relative response of the sensor to different states of polarization. To deal with these cases, we will introduce the *Stokes vector* terminology as it characterizes the time-averaged polarization state most directly applicable to operational sensing. Sir George Gabriel Stokes (1819–1903) developed a method to characterize the polarization state of EM energy that can be directly related to a simple set of measurements of the irradiance transmitted through a conceptual set of four filters, as illustrated in Figure 3.11 [cf. Stokes (1852)]. The filters are a neutral density (polarization insensitive), two linear polarizers oriented at 45° to each other, and a right-hand circular polarizer. All have nominal transmittance values of 50% to randomly polarized radiation (in practice, corrections must be made for actual transmittance values). The Stokes parameters can then be defined as

$$\hat{S}_0 = 2E_0 \tag{3.37}$$

$$\hat{S}_1 = 2(E_1 - E_0) \tag{3.38}$$

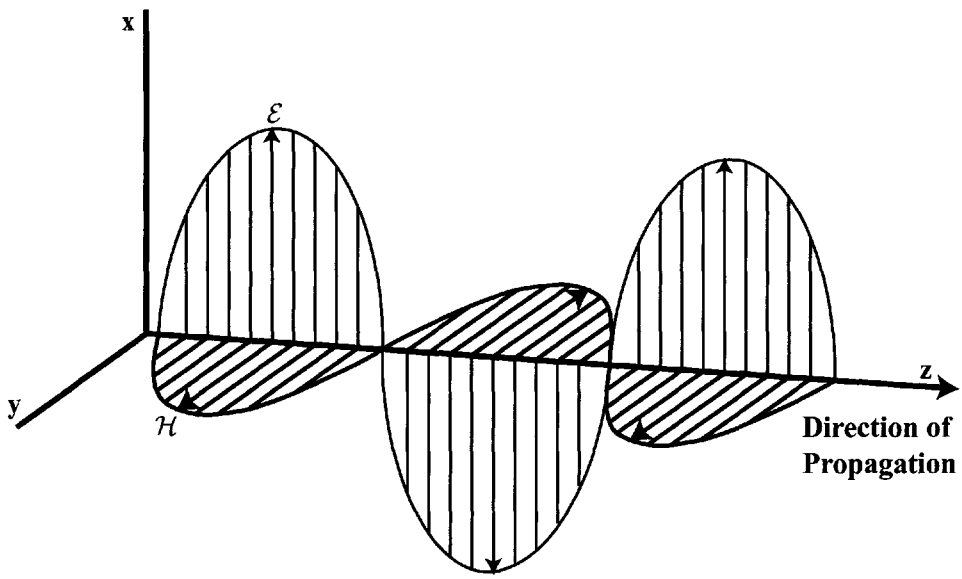
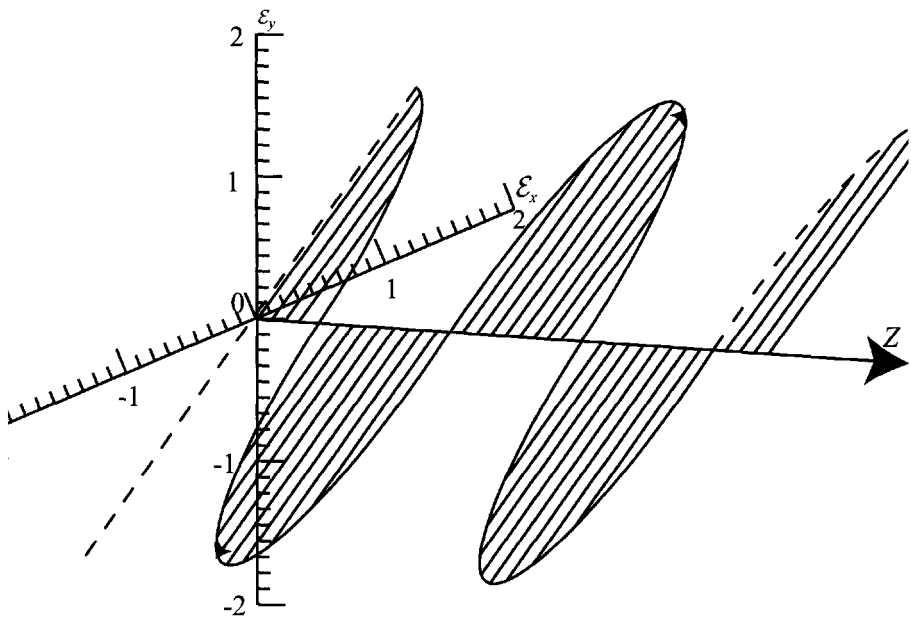
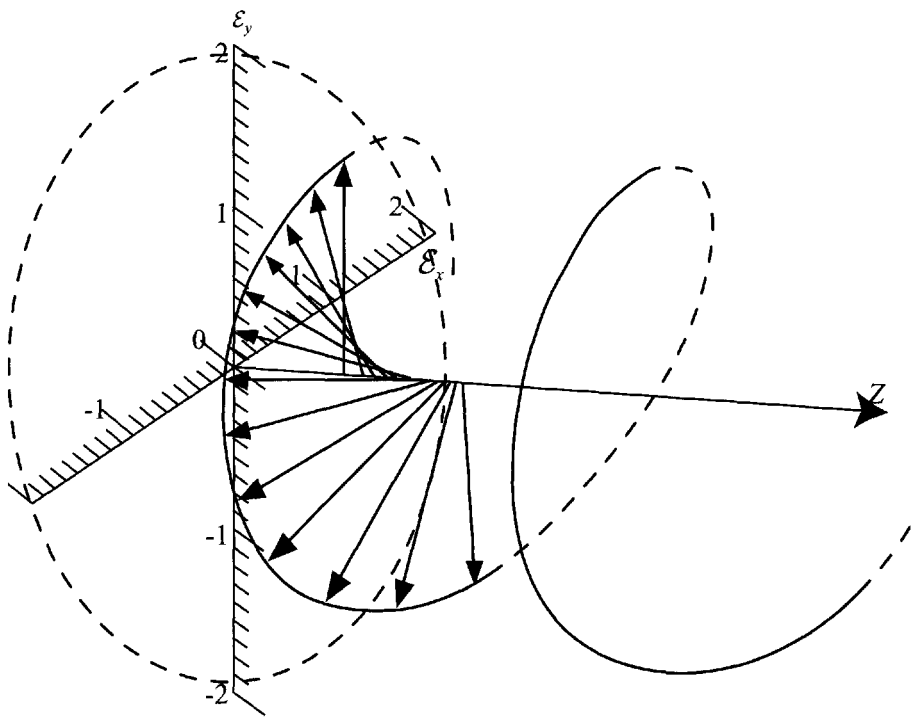


Figure 3.9 Fluctuations in the electric ($\mathcal{E}$) and magnetic ($\mathcal{H}$) fields giving rise to propagation of electromagnetic radiation in the z direction.

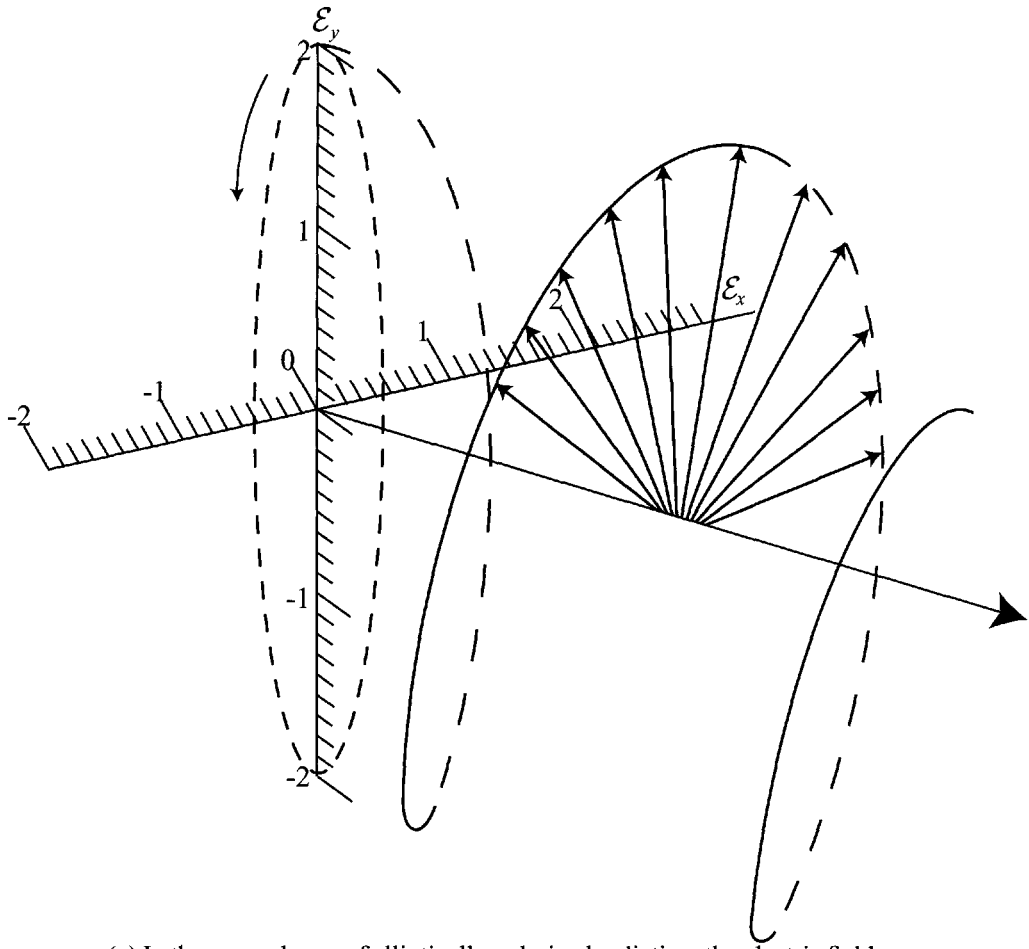


(a) In linear polarization, the electric field vector varies in a fixed plane that would form a line if projected onto the x,y plane.



(b) For circular polarization, the electric field magnitude vector rotates such that it would trace a circle (clockwise in this case) when projected onto the x,y plane.

Figure 3.10 Polarization concepts.



(c) In the general case of elliptically polarized radiation, the electric field vector rotates and changes in magnitude such that its tip would trace an ellipse (counterclockwise in this case) when projected onto the x,y plane.

Figure 3.10 continued. Polarization concepts.

$$\hat{S}_2 = 2(E_2 - E_0) \tag{3.39}$$

$$\hat{S}_3 = 2(E_3 - E_0) \tag{3.40}$$

where $\hat{S}_0$ is the incident irradiance, and $\hat{S}_1$ is related to horizontal polarization (assuming the E_1 polarizer is oriented to transmit horizontally polarized radiation) and will be positive when horizontal radiation dominates over vertical and vice versa. $\hat{S}_2$ is related to polarization at 45° to the horizontal and will be positive when +45° polarization dominates over -45° and vice versa. $\hat{S}_3$ is related to circular polarization and will be positive when right-hand polarization dominates over left and vice versa. E_0 through E_3 are the measured irradiance values as illustrated in Figure 3.11.

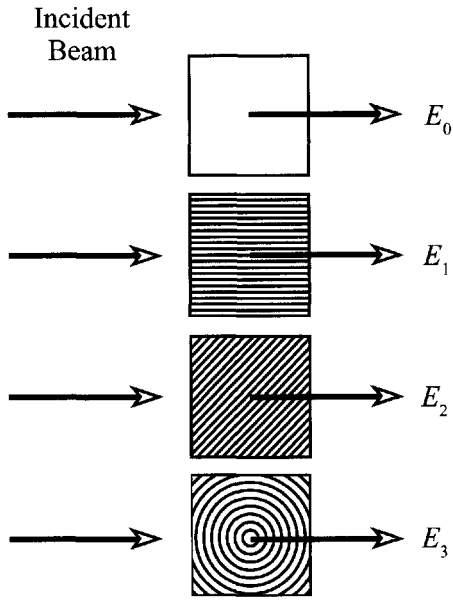


Figure 3.11 Polarization filters used to characterize the Stokes vector.

The Stokes parameters are often normalized through division by the $\hat{S}_0$ value and collected into a vector that carries the polarization characteristics of the ray, i.e.,

$$S = \begin{bmatrix} \hat{S}_0 / \hat{S}_0 \\ \hat{S}_1 / \hat{S}_0 \\ \hat{S}_2 / \hat{S}_0 \\ \hat{S}_3 / \hat{S}_0 \end{bmatrix} = \begin{bmatrix} S_0 \\ S_1 \\ S_2 \\ S_3 \end{bmatrix} \tag{3.41}$$

From the Stokes vector parameters, we can easily characterize the *degree of polarization* (*DOP*) according to

$$\text{DOP} = \frac{\left(S_1^2 + S_2^2 + S_3^2\right)^{\frac{1}{2}}}{S_0} \tag{3.42}$$

and the degree of linear polarization (DOLP) according to

$$\text{DOLP} = \frac{\left(S_1^2 + S_2^2\right)^{\frac{1}{2}}}{S_0} \tag{3.43}$$

We will see in Chapter 4 how each energy–matter interaction (e.g., reflection and transmission) will change the Stokes vector and how in certain cases the Stokes vector can aid in the characterization of materials when polarization sensitive sensors (e.g., sensors with filters such as illustrated in Fig. 3.11) are available.

3.3 RADIOMETRIC CONCEPTS

This section introduces some basic radiometry concepts that draw on the definitions and terms we have introduced thus far, which are needed to define the equations governing radiometric remote sensing of the earth.

3.3.1 Inverse-Square Law for Irradiance from a Point Source

It is often useful to know how irradiance varies with distance from a small source (i.e., in the ideal, a point source). Consider two spherical surfaces at distances r_1 and r_2 from a point source located at p as shown in Figure 3.12(a). We construct two area elements dA_1 and dA_2 defined by the extent of two elements of angle $d\theta$ and $d\phi$ as illustrated. If we assume that the irradiance (E_1) at surface element dA_1 is known, then we wish to know how the irradiance (E_2) at surface dA_2 is related to E_1. If we assume a lossless isotropic medium, then by construction the element of flux ($d\Phi_1$) through the first surface must be identical to the element of flux ($d\Phi_2$) through the second surface. This is because every ray (or photon) contained within the angle elements $d\theta$ and $d\phi$ must pass through dA_1 and dA_2, i.e.,

$$d\Phi_1 \equiv d\Phi_2 [W] \tag{3.44}$$

By definition, the irradiance on each surface can be expressed as

$$E_1 = \frac{d\Phi_1}{dA_1}[Wm^{-2}]$$

and

$$E_2 = \frac{d\Phi_2}{dA_2}[Wm^{-2}] \tag{3.45}$$

but the area elements dA can be represented as the product of the arc lengths $rd\theta$ and $rd\phi$, i.e.,

$$dA_1 = r_1 d\theta r_1 d\phi [m^2]$$

and

$$dA_2 = r_2 d\theta r_2 d\phi [m^2] \tag{3.46}$$

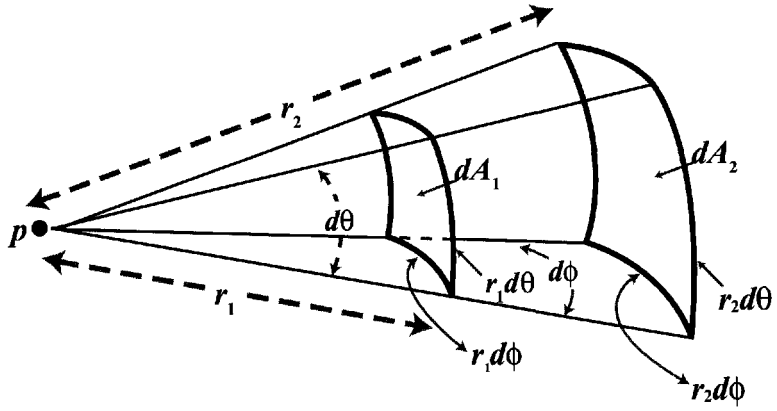
Substituting Eqs. (3.46) and (3.45) into Eq. (3.44) and rearranging, we have

$$E_1 r_1^2 d\theta d\phi = d\Phi_1 = d\Phi_2 = E_2 r_2^2 d\theta d\phi [W] \tag{3.47}$$

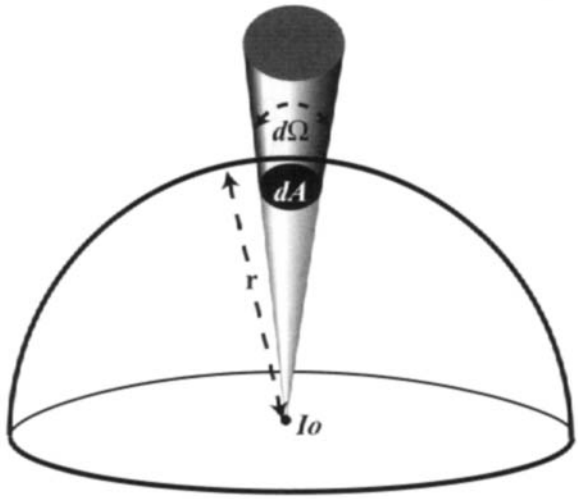
or

$$E_2 = \frac{E_1 r_1^2}{r_2^2} \tag{3.48}$$

The irradiance varies inversely with the square of the distance from a point source. A similar relationship can be derived for the irradiance at a distance r [m]



(a) The irradiance E_2 at r_2 compared to the irradiance at r_1 is given by $E_2 = E_1 r_1^2 r_2^{-2}$.



(b) A source of radiant intensity I_o produces an irradiance at a distance r of I_o / r^2.

Figure 3.12 Inverse-square law for irradiance variation from a point source.

from a source of known radiant intensity I_0 [Wsr^{-1}], as shown in Figure 3.12(b). If we consider an element of solid angle $d\Omega$ encompassing an area element dA on the surface of interest, the irradiance onto the surface must by definition be

$$E = \frac{d\Phi_E}{dA} \tag{3.49}$$

where $d\Phi_E$ is the element of flux onto the surface dA. Also by definition, the radiant intensity is

$$I_0 = \frac{d\Phi_I}{d\Omega} \tag{3.50}$$

where $d\Phi_i$ is the element of flux encompassed by the element of solid angle $d\Omega$. Clearly, in a lossless isotopic media $d\Phi_E = d\Phi_I$, or substituting

$$EdA[Wm^{-2}m^2] = d\Phi_E[W] = d\Phi_I[W] = Id\Omega[Wsr^{-1}sr] \tag{3.51}$$

Recall that by definition $d\Omega = dA/r^2$. Substituting and rearranging yields

$$E = \frac{I}{r^2}[\mathrm{Wm}^{-2}] \tag{3.52}$$

which again reflects a variation in irradiance from a point source that is inversely proportional to the distance squared.

To see where this concept comes into play, let us consider how much energy from the sun reaches the outside of the Earth's atmosphere in terms of irradiance, i.e., the *exoatmospheric irradiance*. We then wish to know how much of the exoatmospheric irradiance would reach the Earth in the absence of an atmosphere. To begin, we will make the simplifying assumption that if we compute the total flux from the sun and then treat the sun as a point source radiating equally in all directions, the irradiance we compute at the Earth–sun distance will be a reasonable approximation of the true irradiance. Grum and Becherer (1979) suggest that if the ratio of the source sensor distance to the source radius is more than a factor of 10, then point source approximations should introduce relatively small error.

To begin, we assume that if the sun is a 5800 K blackbody radiator, then the exitance from the solar surface is

$$M = \sigma T^4 = 5.6697 \cdot 10^{-8}[\mathrm{Wm}^{-2}K^{-4}](5800[K])^4 = 6.42 \cdot 10^7[\mathrm{Wm}^{-2}] \tag{3.53}$$

and the total flux from the solar surface is

$$\begin{aligned}\Phi = M \cdot 4\pi r_s^2 &= 6.42 \cdot 10^7[\mathrm{Wm}^{-2}] \cdot 4\pi \cdot (695.5 \cdot 10^6[\mathrm{m}])^2 \\ &= 3.9 \cdot 10^{26}[\mathrm{W}]\end{aligned} \tag{3.54}$$

where $r_s = 695.5 \cdot 10^6$ [m] is the mean radius of the sun.

If we assume this flux comes from a point source at the center of the sun such that it would produce the observed exitance at the sun's surface, then it should produce a flux at the mean Earth–sun distance $r'_{es} = 149.5 \cdot 10^9$ [m] of

$$\begin{aligned}E_{ex} &= \frac{\Phi}{4\pi r_{es}'^2} \\ &= \frac{3.9 \cdot 10^{26}[\mathrm{W}]}{4\pi(149.5 \cdot 10^9[\mathrm{m}])^2} = 1390[\mathrm{Wm}^{-2}]\end{aligned} \tag{3.55}$$

These crude calculations are in good agreement with an early estimate for the "solar constant" (exoatmospheric irradiance at the mean Earth–sun distance) of 1373 ± 20 [Wm^{-2}] as put forward by Frohlich (1977) based on a number of measurement programs. If we assume the Earth's atmosphere to be approximately 200-km thick, then, according to Eq. (3.48), the variation in irradiance from the exoatmospheric value (E_{ex}) to the value at the Earth's surface due strictly to geometric effects will be

$$E_s = \frac{E_{ex} r_{es}^{2}}{(r_{es}^{'} + 200 \cdot 10^{3}[m])^{2}} = E_{ex}(0.999997) \cong E_{ex} \tag{3.56}$$

The only losses that must be considered in propagating the solar beam are due to atmospheric absorption and scattering, since no geometric effects need to be considered. On the other hand, Chen (1985) points out that the variation in the Earth–sun distance due to the ellipticity of the Earth's orbit can produce as much as a 3.4% variation in the exoatmospheric irradiance from the value at the mean Earth–sun distance. We will often be interested in the exact exoatmospheric spectral irradiance. This value can be computed from tabulated spectral values as found in Thekaekara (1972), or in the MODTRAN radiative transfer code [cf. Berk et al. (1989)], for the exoatmospheric irradiance at the mean Earth–sun distance $E_{g\lambda}$ corrected for the Earth–sun distance on the day of interest according to

$$E_{ex\lambda} = E_{g\lambda} \frac{r_{es}^{'2}}{r_{es}^{2}} \tag{3.57}$$

where $E_{ex\lambda}$ is the exoatmospheric spectral irradiance on a specific day when the Earth–sun distance is r_{es}.

3.3.2 Projected Area Effects (cos θ)

Most radiance calculations involve dealing with radiation propagating through or onto a surface that is not perpendicular to the propagating beam. In these cases, it is important to take into account the effects of projected area. Consider the problem illustrated in Figure 3.13(a). In case 1, the element of flux $d\Phi$, associated with the ray **p**, is incident on the element of unit area dA_u such that the normal to the surface **N** is coincident with **p** (i.e., the beam is perpendicular to the surface). In this case, the irradiance onto the surface is by definition

$$E_0 = \frac{d\Phi_0}{dA_u} \tag{3.58}$$

where the subscript 0 refers to the angle from the normal to the surface to the incident ray. In case 2, illustrated in Figure 3.13(b), the same amount of incident flux is spread over a larger area, since the surface is rotated through an angle θ. The irradiance in this case is

$$E_\theta = \frac{d\Phi_\theta}{dA_\theta} = \frac{d\Phi_0}{dA_u/\cos\theta} = E_0 \cos\theta \tag{3.59}$$

where θ is the angle from the normal to the surface to the incident ray, and we have used the geometric construction that, in order for the area element to capture the same amount of flux (i.e., for $d\Phi_\theta = d\Phi_0$), the area must be increased by a factor of cos θ. That is, the area element A_θ is the projection of the area element perpendicular to the ray (A_u) onto the plane of the surface of interest ($A_0/\cos\theta$). Many people find it easier to think of the irradiance onto the surface as the component of the irradiance vector along the ray, which is normal to the surface as illustrated in Figure 3.13(c).

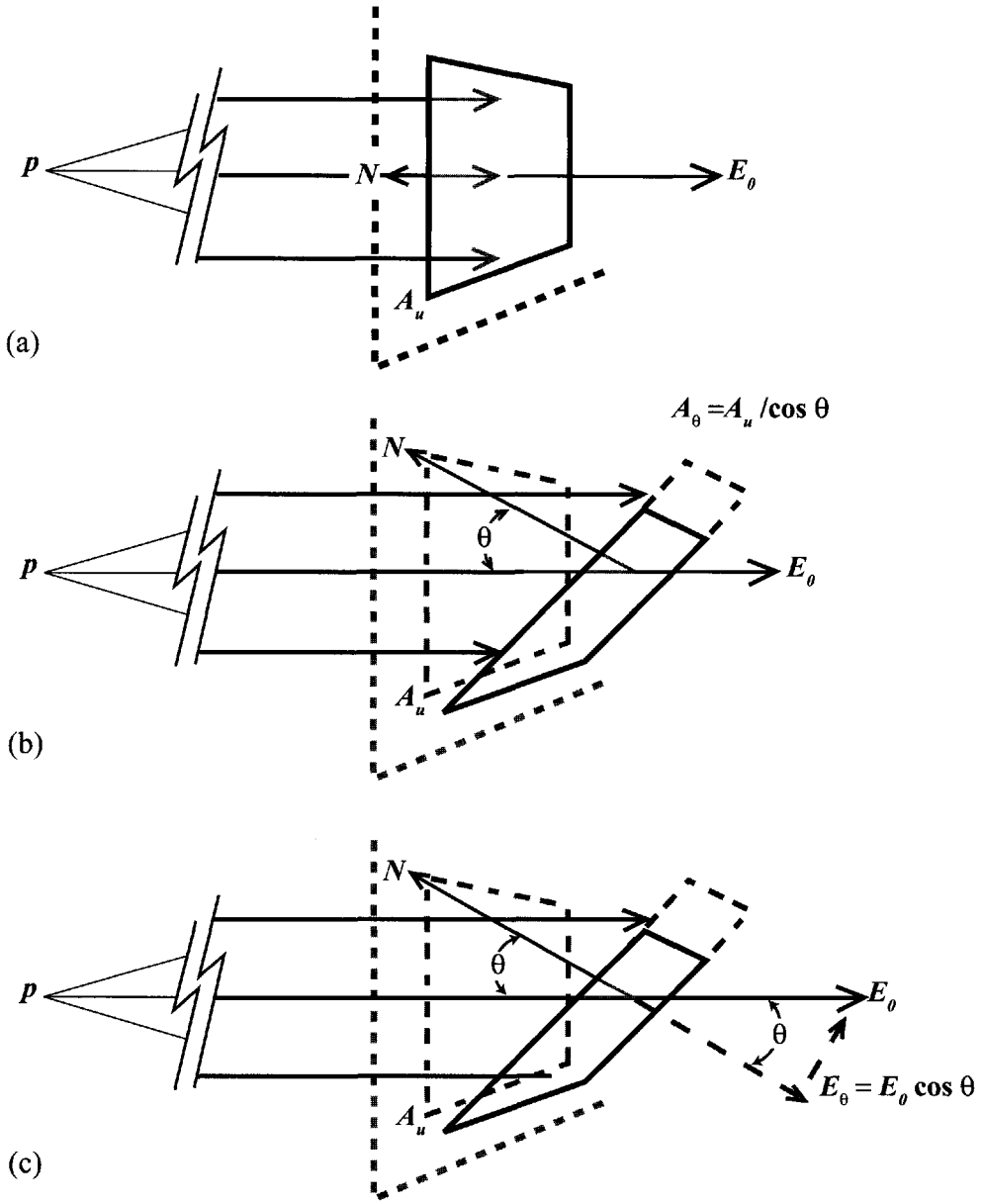


Figure 3.13 Projected area effect (cos θ)].

3.3.3 Lambertian Surfaces

One of the most difficult questions we will have to deal with in remote sensing is the question of how the energy leaving a surface is angularly distributed into the hemisphere above the surface. Is all of it perpendicular to the surface? Is the radiant intensity the same in all directions? Can the intensity distribution be characterized in any functional form? We will often make use of an idealized surface known as a *Lambertian surface* to help us deal with these questions. A Lambertian radiator, or reflector, is characterized by a well-behaved variation in radiant intensity according to

$$I_\theta = I_0 \cos\theta \tag{3.60}$$

where I_0 is the intensity normal to the surface and θ is the angle from the normal to the surface to the direction of interest. Thus, a Lambertian surface will have a steady decrease in the intensity, approaching zero at grazing angles. It is often more intuitive to consider the angular distribution of radiance from a Lambertian surface. Visual response is proportional to radiance, and this will tell us how a Lambertian surface will "look" when viewed at different angles.

By definition, the radiance along the normal from a Lambertian surface will be

$$L_0 = \frac{dI_0}{dA\cos(0)} = \frac{dI_0}{dA} \tag{3.61}$$

Similarly, the definition of radiance into any direction θ from the normal is

$$L_\theta = \frac{dI_\theta}{dA\cos\theta} \tag{3.62}$$

Substituting Eqs. (3.60) and (3.61) into Eq. (3.62) yields

$$L_\theta = \frac{dI_\theta}{dA\cos\theta} = \frac{dI_0\cos\theta}{dA\cos\theta} = \frac{dI_0}{dA} = L_0 \tag{3.63}$$

which states that the radiance from a Lambertian surface is the same in all directions (i.e., the decreasing intensity with angle is exactly compensated by a decrease in projected area). Since perceived brightness is proportional to the radiance in the visible region, this means that a Lambertian surface would look the same from all directions. There are many examples in nature of surfaces that are approximately Lambertian, particularly if grazing angles are avoided. Thus, while we cannot assume that all surfaces are Lambertian, we will often use Lambertian assumptions as approximations or as a point of reference for discussion of less well-behaved surfaces.

3.3.4 Magic π

Because of the importance of Lambertian surfaces as reference materials, we will consider the relationship between the exitance from a Lambertian surface and the radiance from that surface. For example, consider the case illustrated in Figure 3.14(a) of a Lambertian surface having reflectance r exposed to a beam having irradiance E_0 onto an imaginary surface perpendicular to the beam at the surface. We know the radiance from the surface will be the same in all directions, but we seek its magnitude. The irradiance onto the surface from vector analysis of Figure 3.14(a) is

$$E_\theta = E_0 \cos\theta \ [\mathrm{Wm}^{-2}] \tag{3.64}$$

where θ is the angle from the normal to the surface to the incident ray. From the definition of reflectance Eq. (3.27), we have the exitance expressed as

$$M = E_\theta r = E_0 r \cos\theta \ [\mathrm{Wm}^{-2}] \tag{3.65}$$

The radiance by definition can then be expressed as:

$$L = \frac{dM}{d\Omega \cos\theta} [Wm^{-2} sr^{-1}] \tag{3.66}$$

or rearranging

$$M = \int dM = \int L \cos\theta d\Omega [Wm^{-2}] \tag{3.67}$$

If we take the integral of both sides to find the total exitance into the hemisphere above the reflecting surface, we have

$$dM = Ld\Omega \cos\theta [Wm^{-2}] \tag{3.68}$$

where the left-hand side is simply the exitance we already know from Eq. (3.65), and the right-hand side is the integral with respect to solid angle over the hemisphere above the surface. Cos θ changes with respect to solid angle, so we need to convert $d\Omega$ into a more tractable form to perform the integration. If we construct the element of solid angle associated with an arbitrary (or unit) sphere of radius ρ, as illustrated in Figure 3.14(b), then we have, by definition,

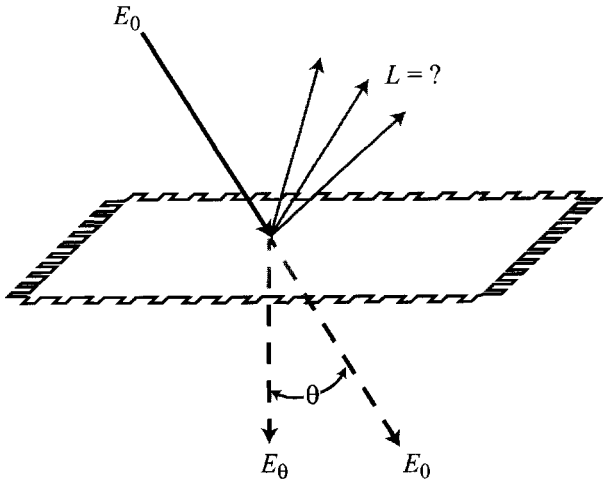
$$d\Omega = dA/\rho^2 \tag{3.69}$$

and by construction,

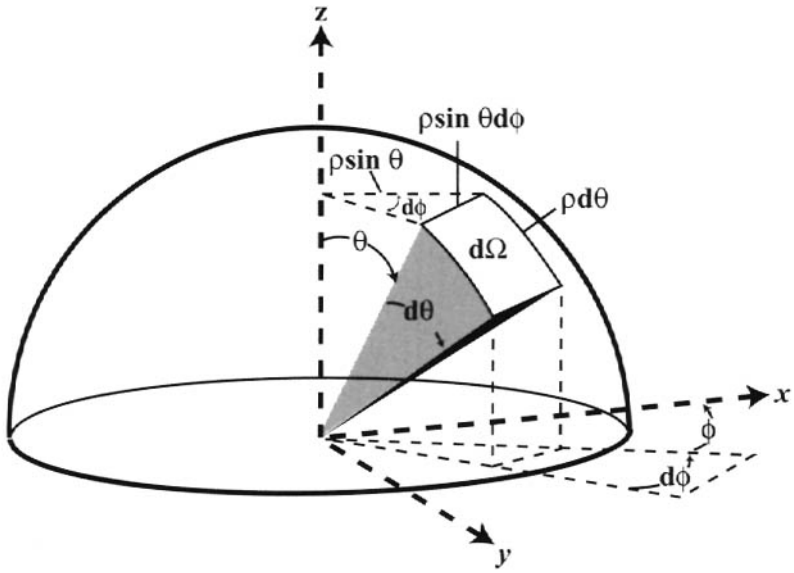
$$dA = \rho d\theta \rho \sin\theta d\phi [\mathrm{m}^2] \tag{3.70}$$

where the surface is defined by the x,y plane, θ is the angle from the normal to the surface (i.e., from the z axis), and φ is the azimuthal angle swept about the z axis in any plane parallel to the x,y plane. Substituting Eq. (3.70) into Eq. (3.69), we see that the dependence on the size of the hemisphere vanishes, leaving

$$d\Omega = \frac{\rho^2 d\theta \sin\theta d\phi}{\rho^2} \frac{[m^2]}{[m^2]} = \sin\theta d\theta d\phi [sr] \tag{3.71}$$



(a) Statement of a problem in terms of reflected radiance.



(b) The element of solid angle $d\Omega$ can be expressed as $d\Omega = \sin\theta d\theta d\phi$.

Figure 3.14 Magic π.

Substituting this expression for the element of solid angle in Eq. (3.68), we are left with a double integral on θ and ϕ of the form

$$M = \int\int L\cos\theta\sin\theta d\theta d\phi \tag{3.72}$$

where inspection of Figure 3.14(b) for the limits of integration over the hemisphere yields

$$M = \int_{\phi=0}^{2\pi}\int_{\theta=0}^{\frac{\pi}{2}} L\cos\theta\sin\theta d\theta d\phi \tag{3.73}$$

If *and only if* we have a Lambertian surface, the radiance is a constant and can be taken outside the integral. The integral on θ can then be seen to be

$$M = L\int_{\phi=0}^{2\pi}\left[\int_{\theta=0}^{\frac{\pi}{2}}\cos\theta\sin\theta d\theta\right]d\phi$$

$$= L\int_{\phi=0}^{2\pi}\left[\frac{\sin^2\theta}{2}\right]\Bigg|_0^{\frac{\pi}{2}} d\phi = \frac{L}{2}\int_{\phi=0}^{2\pi} d\phi \tag{3.74}$$

and the integral on ϕ yields

$$M = L[Wm^{-2}sr^{-1}]\pi[sr] = L\pi[Wm^{-2}] \tag{3.75}$$

or on rearranging, we see that for a Lambertian surface the radiance and radiant exitance are related by a factor of π, i.e.,

$$L = \frac{M}{\pi} \tag{3.76}$$

or in our case

$$L = \frac{M}{\pi} = \frac{E_0 r\cos\theta}{\pi} \tag{3.77}$$

This factor of π [sr] shows up repeatedly (magically) in radiometric calculations (usually without its derivation and often inappropriately), but it is generally valid only for treatment of Lambertian surfaces or when a comparison to a Lambertian surface is used as is the case with the reflectance factors that are introduced in the next chapter (cf. Sec. 4.2.1). For example, if we assume a blackbody is a Lambertian radiator, the Planck equation can be expressed in terms of spectral radiance as

$$L_\lambda = \frac{M_\lambda}{\pi} = 2hc^2\lambda^{-5}(e^{\frac{hc}{\lambda kT}} - 1)^{-1}[\mathrm{Wm^{-2}sr^{-1}\mu m^{-1}}] \tag{3.78}$$

It is important to realize that in order to find the relationship between radiance and exitance [i.e., to solve Eq. (3.73)], the relative variation in radiance into all possible directions in the hemisphere above the surface would have to be known. It

is the magnitude of this task that forces us to make the assumption of Lambertian behavior whenever possible.

3.3.5 Lens Falloff

In Chapter 2, we introduced the concept of using a film-based camera as a radiometer, and in the sections that follow we will expand on this theme by addressing the use of a variety of imaging systems as quantitative radiometers. The imaging nature of these radiometers will cause many of them to have a substantial angular field of view. To see what effect this off-axis viewing will have on the radiometry, we will consider the simplest case of a pinhole camera and examine the variation in exposure as a function of the view angle (θ). A pinhole camera is simply a light-tight box (e.g., a shoe box) with a sheet of film at the back, as illustrated in Figure 3.15. The image is acquired by opening a shutter (removing your finger from the hole in the shoe box) and exposing the film for a short period of time. The small aperture acts as a lens by only allowing flux from a very limited direction to hit each point on the film. In our case, we want to imagine imaging a scene of uniform radiance (e.g., a uniform gray Lambertian reflector) and determine if the exposure will vary across the image format (i.e., will two identical surfaces appear different due solely to their location in the image?). To begin, we compute the irradiance (E_0) on the focal plane along the primary axis. If the scene has radiance L [$Wm^{-2}sr^{-1}$], then in a lossless medium we will have radiance $L_0 = L$ reaching the aperture along a $\theta = 0$ view angle (i.e., along the primary axis). The radiant intensity through the aperture in the $\theta = 0$ direction is

$$dI_0 = L_0 dA \cos\theta \tag{3.79}$$

from Eq. (3.9). For small apertures, we can assume that θ and dI_0 do not vary across the aperture, so

$$I_0 = L_0 A \cos(0)[\mathrm{Wm^{-2}sr^{-1}}][\mathrm{m^2}] = L_0 A[\mathrm{Wsr^{-1}}] \tag{3.80}$$

where A is the area of the aperture [m^2]. From Eq. (3.52), the irradiance onto the center of the focal plane a distance r_0 away is

$$E_0 = \frac{I_0}{r_0^2}[\mathrm{Wm^{-2}}] = \frac{L_0 A}{r_0^2}[\mathrm{Wm^{-2}}] \tag{3.81}$$

and recalling that the exposure is simply the irradiance times the time of exposure (t) yields the on-axis exposure of

$$H_0 = E_0 t[\mathrm{Wm^{-2}}][\mathrm{sec}] = \frac{L_0 A t}{r_0^2}[\mathrm{joules\ m^{-2}}] \tag{3.82}$$

Similarly, in a lossless medium the radiance reaching the pinhole headed in a direction defined by the view angle (θ) from the primary axis is

$$L_\theta = L = L_0 \tag{3.83}$$

By analogy to Eq. (3.80), the radiant intensity in the direction defined by the view angle θ radiating from the aperture is

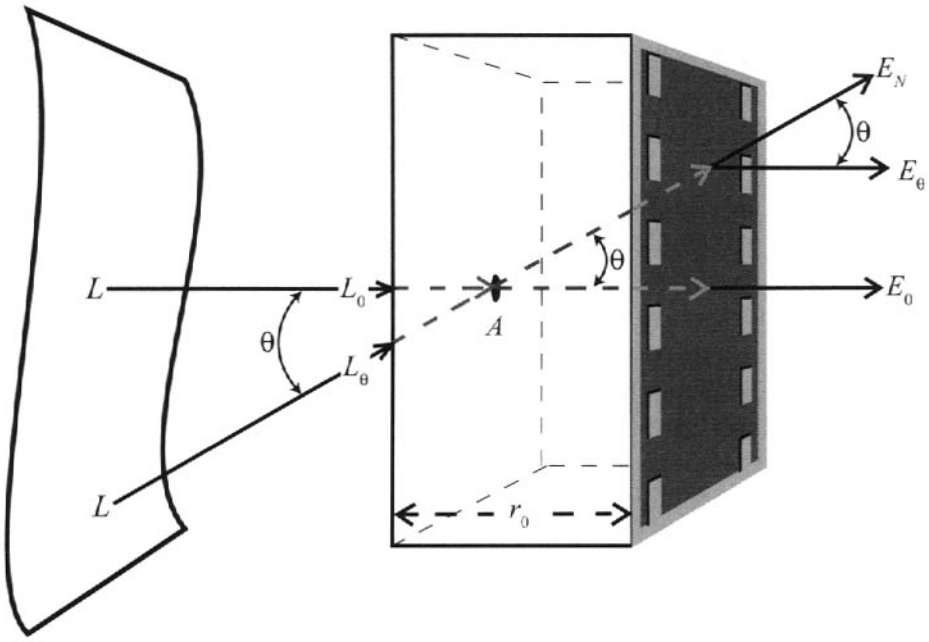


Figure 3.15 $\cos^4(\theta)$ lens falloff for a pinhole camera.

$$I_\theta = L_0 A \cos\theta \tag{3.84}$$

i.e., the radiant intensity is reduced due to the aperture appearing smaller for off-axis viewing. The irradiance E_N that would be observed on a surface normal to the ray at the focal plane (cf. Fig. 3.15) is

$$E_N = \frac{I_\theta}{r_\theta^2} = \frac{I_\theta}{(r_0/\cos\theta)^2} = \frac{L_0 A \cos^3\theta}{r_0^2} \tag{3.85}$$

i.e., the irradiance is reduced by the square of the increased distance from the aperture to the focal plane at the view angle θ. To find the irradiance onto the focal plane, we take the vector component of E_N perpendicular to the focal plane, yielding

$$E_\theta = E_N \cos\theta = \frac{L_0 A \cos^4\theta}{r_0^2} \tag{3.86}$$

Note that this derivation can also be done by rearranging the definition of radiance relative to irradiance to yield

$$dE = Ld\Omega\cos\theta \tag{3.87}$$

where dE is the element of irradiance onto the focal plane from the radiance at the aperture. Recognizing that L has the constant value L_0 over the aperture and that θ is approximately constant over Ω since the pin hole is so small and substituting the definition for solid angle, we have

$$dE = L_0 \frac{dA_\theta}{r_\theta^2} \cos\theta \tag{3.88}$$

Integration by the area element over the aperture as viewed by the receiver at an angle θ and substituting for $r_\theta = r_0/\cos\theta$ yields

$$E = \frac{L_0 (A\cos\theta)\cos\theta}{(r_0/\cos\theta)^2} = \frac{L_0 A \cos^4\theta}{r_0^2} \tag{3.89}$$

In either case, the resulting exposure would be

$$H_\theta = E_\theta t = \frac{L_0 A t}{r_0^2} \cos^4\theta = H_0 \cos^4\theta \tag{3.90}$$

indicating that the exposure would fall off from the center of the format as the cosine of the view angle to the fourth power.

Since most of us don't do radiometric image analysis with pinhole cameras, we need to know how this derivation relates to conventional imaging systems. From the standpoint of radiometry, the major function of a lens is to gather flux more effectively from all directions. As a result, the effective size of the aperture is more nearly constant with view angle—i.e., $\cos^4\theta$ becomes approximately $\cos^3\theta$. With a conventional camera, the effects due to increase in path and projected area of the sensor in the off-axis direction remain approximately the same as for the pinhole camera. The fact that the aperture may not be a good approximation of a point source results in a more complicated version of Eq. (3.52) that must account for variation in the distance and irradiation angles from the various elements of area across a finite aperture. For a real camera system, the relation between irradiance across the format can usually be approximated as

$$E_\theta = E_0 \cos^n(\theta) \tag{3.91}$$

where for simple cameras *n* will usually take on a value of approximately 3, with the exact value normally obtained by curve-fitting through experimental data.

To evaluate the magnitude of this effect, consider first a 70-mm format camera flown with an 80-mm focal length lens. The actual image format is approximately 55-mm or 28-mm from center to edge for a field of view of

$$\theta = \tan^{-1}\left(\frac{28mm}{80mm}\right) \cong 19° \tag{3.92}$$

If an optical bench analysis of the lens yields a value for *n* in Eq. (3.91) of 3.2, then the falloff in exposure will be

$$H_{19} = H_0 \cos^{3.2} 19° \cong 0.83 H_0 \tag{3.93}$$

i.e., there will be only 83% as much exposure at the edges of the format as at the center for identical objects. This is clearly a nonnegligible effect that must be compensated for in quantitative image analysis. Because it is a well-behaved repeatable effect, it is easily accounted for by measuring the exposure (H_θ) or irradiance at any point of interest on the focal plane and then correcting for what that exposure

would have been if the object were imaged at the center of the format. The corrected value H_0' would be

$$H_0' = \frac{H_\theta}{\cos^n(\theta)} \tag{3.94}$$

If we consider the exposure falloff from a standard 9-inch format camera (with actual image area 230 mm times 230 mm) with a 6-inch focal length lens, we have a field of view to the edges of the format of

$$\theta = \tan^{-1}\left(\frac{115mm}{152mm}\right) \cong 37° \tag{3.95}$$

and 47° to the corners of the format. This results in exposures of 51% and 32% of the on-axis values for $n = 3$. This effect is very dramatic and causes the edges of the format to be underexposed and difficult to interpret when the center of the format is properly exposed. To compensate for this effect, antivignetting filters are often used with this type of camera to partially account for the lens falloff effects. These filters have lower transmission in the center of the format than at the edges where the transmission is near unity. The combined effects of lens falloff and an antivignetting filter are illustrated in Figure 3.16. The $\cos^3\theta$ curve shows how exposure would vary with angle with no filter present. The filter transmission with angle $\tau(\theta)$ is also shown, with the actual exposure variation resulting from the product of the $\cos^3\theta$ curve and the transmission curve. In general, this will still result in considerable residual variation across the format, which must be measured and corrected. Making an exact antivignetting filter [i.e., $\cos^n(\theta)\tau(\theta)$ = a constant] is seldom attempted. Besides the technical difficulties, the central transmission would be so low that excessively long exposure times would be required to properly expose the film.

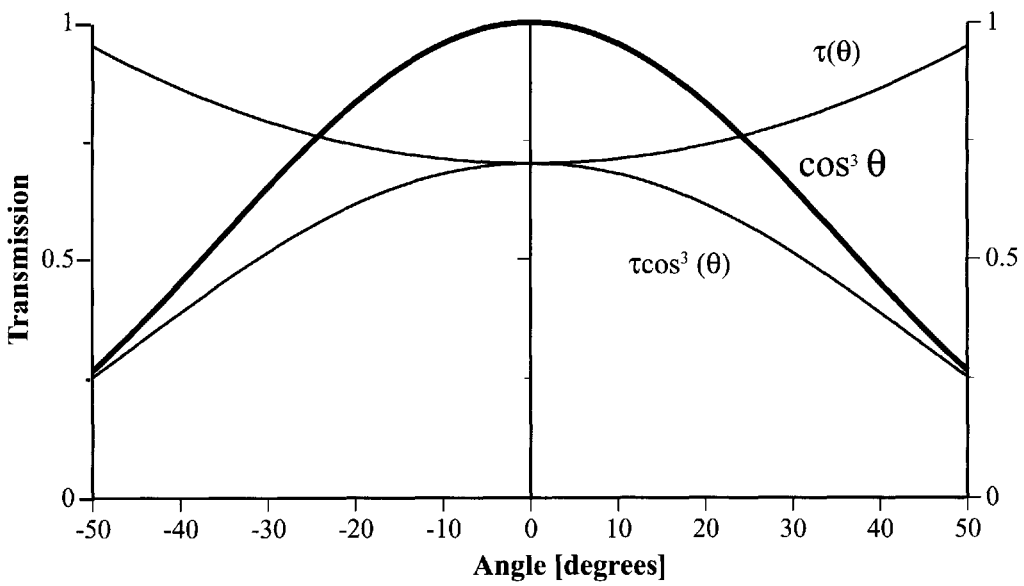


Figure 3.16 Lens falloff and antivignetting filter effects.

3.4 ATMOSPHERIC PROPAGATION

One of the most critical factors affecting radiometric remote sensing is the effect of the atmosphere on the propagating energy. In most cases, the atmosphere is perceived as a hostile entity whose adverse impacts must be neutralized or eliminated before remotely sensed data can be properly analyzed. In this section, we will examine the underlying principles that describe radiation propagation through the atmosphere. These principles will be used in the next chapter to describe the governing equation for the radiance reaching a remote sensing system.

3.4.1 Atmospheric Absorption

We want to examine the effect of the atmosphere on a beam of energy propagating through the atmosphere (e.g., a beam associated with the exoatmospheric solar irradiance as it propagates to the Earth's surface). To begin, we will consider an idealized atmosphere that only exhibits absorption effects (i.e., nonscattering). We define *absorption* as the process of removal of energy (photons) from the beam by conversion of the electromagnetic energy to another form (usually thermal). For atmospheric propagation, this results from absorption of the photons by the constituent molecules in the atmosphere when a photon induces a molecular vibration, rotation, or electron orbital transition to an alternate energy state. Recall from modern physics that these are discrete (i.e., quantized) energy transitions, so only photons with selected energy levels can be absorbed. In the simplest case, we would have absorption occurring in narrow spectral lines associated with those photons having the exact energy ($h\nu$) needed to induce an allowable energy transition. Because of the low densities in the atmospheric media of interest, we can assume that each absorption event is discrete and that the total absorption is simply the sum of the individual events (i.e., the absorption of each photon only impacts the molecule that absorbs it). This would lead to an atmospheric absorption spectrum made up of discrete lines, as illustrated in Figure 3.17(a). In fact, several factors contribute to the broadening of these discrete lines [cf. USAFGL (1985), Chap. 18, for a general treatment]. The Heisenberg uncertainty principle indicates that the time interval of energy transitions can only be known within set limits, and, therefore, the associated emission, or absorption frequency, will have some uncertainty or width. However, this effect is small compared to Doppler and pressure broadening of the absorption lines. Doppler broadening is due to the fact that a particle moving toward or away from an observer will emit or absorb frequency-shifted photons in accord with conventional Doppler theory. This results in a random distribution of the absorption frequencies about the nominal absorption line caused by the random distribution of particle velocities relative to the incident beam. Since particle motion is increased at higher temperature and dampened by inertia, the extent of Doppler broadening increases with temperature and decreases with molecular mass according to

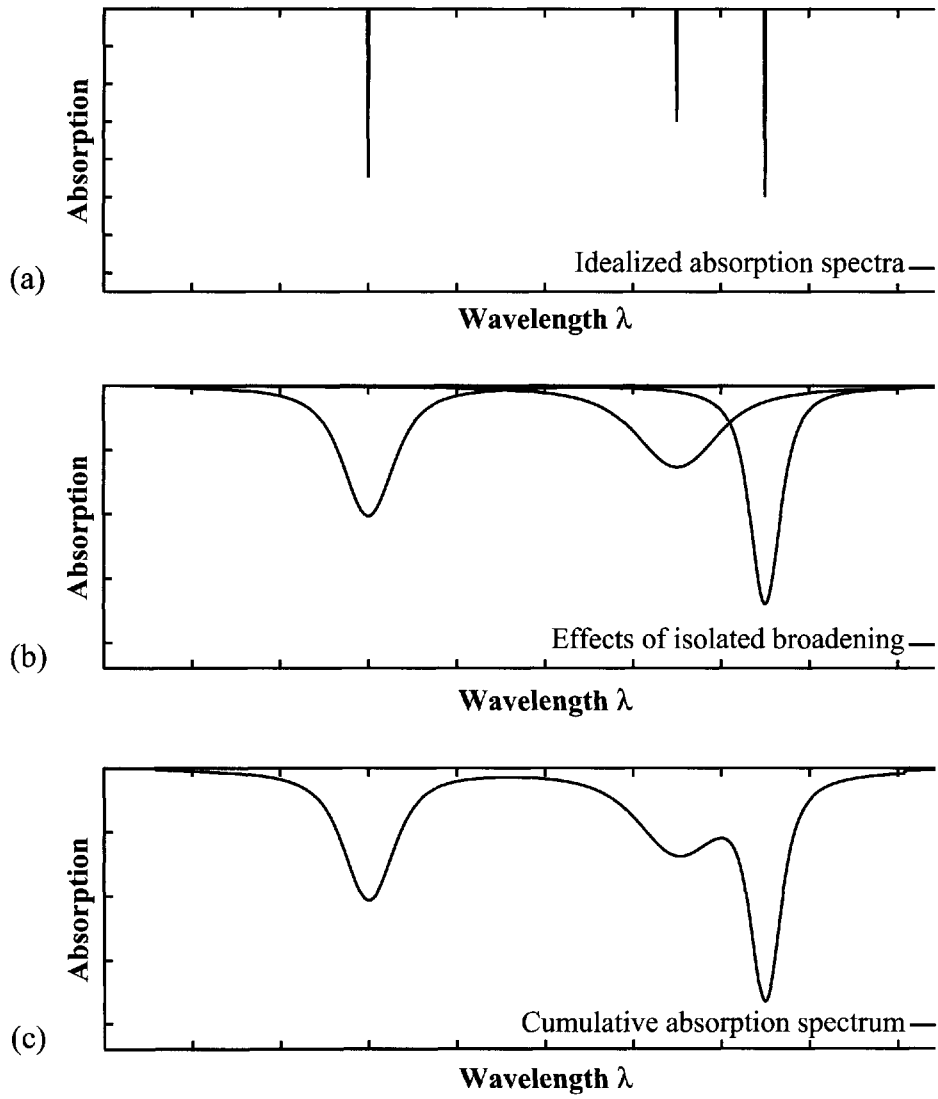


Figure 3.17 Characteristics of absorption spectra.

$$\Delta v \propto \sqrt{\frac{T}{M}} \tag{3.96}$$

where Δv is the width of the absorption frequency due to broadening, T is the temperature in Kelvin, and M is the gram molecular weight.

Pressure or collision broadening of the absorption lines results from the interaction or collision of one molecule with another perturbing the energy states. The degree of broadening (Δv) is proportional to the relative pressure (p). The overall theory of line broadening is beyond the scope of this treatment, and we must be satisfied at this point to illustrate the net result in Figure 3.17. What we see is that the absorption spectrum is made up of many discrete lines that are broadened to various extents. Furthermore, when many broadened lines are close together, we

tend toward a continuum of absorption where the discrete line nature is largely lost. Depending on the spectral region of interest and the spectral resolution of the sensor, both the individual broadened lines and the continuum concept will be important for remote sensing.

In simplified form, we can derive the following relationships between the number and efficiency of absorbers and their effect on the propagating beam. At each wavelength, we define the *absorption cross section* C_α to be the effective size of a molecule relative to the photon flux at that wavelength. Conceptually, this can be expressed as

$$C_\alpha = C_g \xi = \pi r^2 \xi [\mathrm{m}^2] \tag{3.97}$$

where C_g [m^2] is the geometric cross section for a molecule of radius r [m] and ξ is a unitless wavelength-dependent efficiency factor that is proportional to the molecule's ability to absorb flux. Values of C_α are available as a function of wavelength for the atmospheric constituents. These values can be derived for particular temperatures and pressures from experimental data or through molecular energy theory, and then adjusted for the effects of the temperature and pressure of interest, as alluded to above. The molecule is then assumed to be a perfect absorber over that cross-sectional area. To compute the fractional amount of energy lost per unit length of transit in a propagating beam, we need to know the number density of the molecules. Referring to Figure 3.18, we let m' be the number of molecules in a unit volume of side dimension l [m]. Then, the area blocked (A_b) by the molecules is

$$A_b = m' C_\alpha [\mathrm{m}^2] \tag{3.98}$$

The area on the face of the volume (A_f) onto which the molecules were projected is

$$A_f = l^2 [\mathrm{m}^2] \tag{3.99}$$

The fraction of the face blocked by the absorbing molecules (F) is

$$F = \frac{m' C_\alpha}{l^2} \left[\frac{\mathrm{m}^2}{\mathrm{m}^2} \right] \tag{3.100}$$

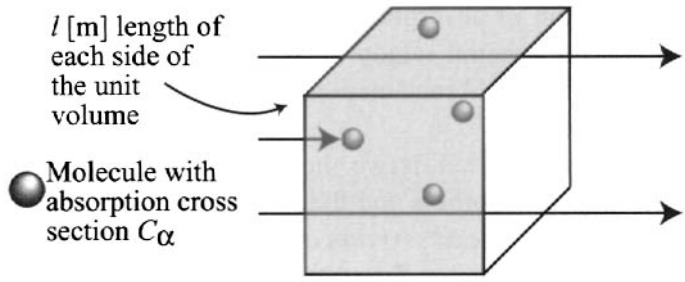
Therefore, the fractional amount of flux absorbed β_α per unit length of transit (l) is

$$\beta_\alpha = \frac{F}{l} = \frac{m'}{l^3} C_\alpha [\mathrm{m}^{-1}] = \frac{m'}{V} C_\alpha = m C_\alpha [\mathrm{m}^{-1}] \tag{3.101}$$

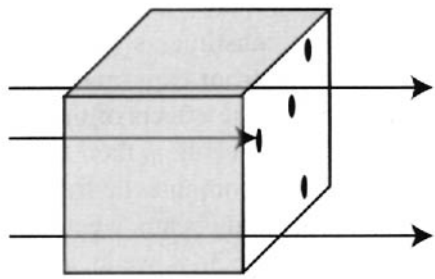
where V is the unit volume, m is the *number density* of molecules, defined as the number of molecules per unit volume, and β_α is the *absorption coefficient*, defined as the fractional amount of flux lost to absorption per unit length of transit in a propagating beam.

According to Grum and Becherer (1979), for an element of path length dz[m] in the medium, the element of fractional flux lost can be expressed as:

$$\frac{d\Phi}{\Phi} = -\beta_\alpha(z) dz \tag{3.102}$$



(a) A unit volume containing m' absorption centers. We assume that the medium has a large mean free path such that in a small volume, if we project the molecules onto one face, there will be no overlap.



(b) Projection of absorbers onto the face of the volume.

Figure 3.18 Computation of the absorption coefficient.

i.e.,

$$\frac{d\Phi}{dz} = -\beta_\alpha(z)\Phi \tag{3.103}$$

where we have made the dependence of β_α on location in the media explicit and the negative sign indicates a loss in energy with distance. For propagation along a finite path starting at distance zero where we have initial flux Φ_0 to distance z where we have flux Φ_z, we have

$$\int_{\Phi_0}^{\Phi_z} \frac{d\Phi}{\Phi} = \int_0^z -\beta_\alpha(z)dz = ln\Phi \Big|_{\Phi_0}^{\Phi_z} = \int_0^z -\beta_\alpha(z)dz = ln\Phi_z - ln\Phi_0$$

$$= ln\left(\frac{\Phi_z}{\Phi_0}\right) = \int_0^z -\beta_\alpha(z)dz \tag{3.104}$$

Making both sides powers of e to simplify the left-hand side yields

$$\frac{\Phi_z}{\Phi_0} = e^{-\int_0^z \beta_\alpha(z)dz} \tag{3.105}$$

Recognizing the left-hand side as a definition of transmission and solving for the simplified case of a homogeneous medium, we have

$$\tau = \frac{\Phi_z}{\Phi_0} = e^{-\beta_\alpha \int_0^z dz} = e^{-\beta_\alpha z} \tag{3.106}$$

which is variously known as Lambert's law or Bouguer's law.

The product $\beta_\alpha z$ is generally referred to as the *optical depth* (δ), i.e.,

$$\delta_\alpha = \beta_\alpha z \tag{3.107}$$

To this point, we have implicitly assumed a medium containing a single constituent. For a homogeneous atmosphere containing many types of molecules, we introduce the subscript i to denote the particular constituent. If we assume that the molecules interact independently with the propagating flux, we can express the transmission as

$$\tau = \prod \tau_i = e^{-\sum \delta_i} = e^{-\sum \beta_{\alpha i} z} = e^{-\sum m_i C_{\alpha i} z} = e^{-\beta_\alpha z} = e^{-\delta_\alpha} \tag{3.108}$$

where Π designates the product of the transmission values for each constituent if computed separately, the summation (Σ) is over all constituents, and we redefine $\beta_\alpha = \Sigma\beta_{\alpha i}$ to be the composite absorption coefficient and δ_α to be the composite optical depth due to absorption.

From Eq. (3.108), it is clear that to find the transmission along a beam, we need to know the number density of each constituent molecule along the path, the absorption cross section as a function of wavelength for each constituent, and how that cross section varies with environmental parameters (e.g., T, p) along the path. Often the atmosphere can be treated as a series of homogeneous layers, as shown in Figure 3.19. The overall optical depth can then be found from a numerical integration of Eq. (3.104) to be a simple sum of the optical depths for each of the homogeneous layers. The transmission is then given by

$$\tau = e^{-\sum_j \left(\sum_i m_{ij} C_{\alpha ij} z_j \right)} = e^{-\sum_j \delta_{\alpha j}} = e^{-\delta_\alpha} \tag{3.109}$$

where the subscript i differentiates the constituents in the atmosphere, and j the layers, and the terms are as described in Figure 3.19.

In many cases, the absorption is expressed in terms of the fractional amount of energy lost for a path length through a gas with a concentration expressed in parts per million. This is a logical way to express absorption given how it is measured in the lab and how it might be used to estimate the concentration of a layer of gas in the atmosphere of a known thickness. The transmission can be expressed in terms of this concentration-depth absorption in the following fashion:

$$\tau = e^{-\beta_{cz} c_{ppm} z} = e^{\beta z} = e^{-C_\alpha m z} \tag{3.110}$$

where β_{cz} is the absorption per part per million meter [$\text{ppm}^{-1}\text{m}^{-1}$] and c_{ppm} is the gas concentration in parts per million. Recall from the universal gas law that the total number of molecules per unit volume of a gas can be expressed as

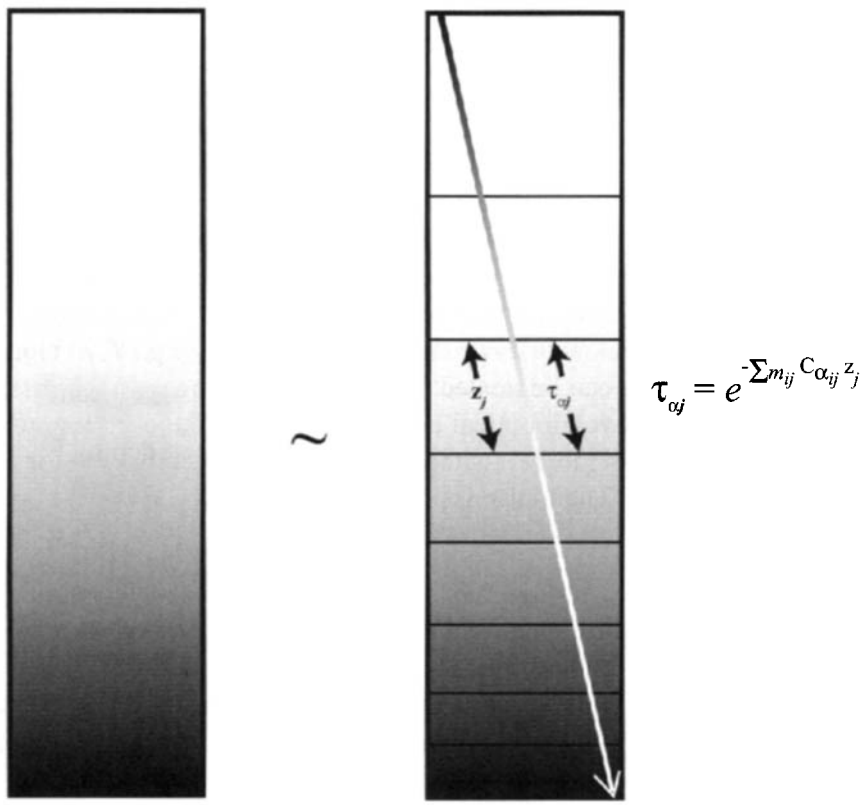
$$n/V = p/kT\left[m^{-3}\right] \tag{3.111}$$

where p is pressure, T is temperature, V is volume, n is the total number of molecules, and k is the universal gas constant (8.31 j/mole K or 1.38 10^{-23} j/molecule K). Then, a simple proportion can be used to find the number of molecules of the constituent gas per unit volume (m) according to

$$\frac{c_{ppm}}{10^6} = \frac{m}{p/kT} \tag{3.112}$$

or

$$m = \frac{c_{ppm}\, p/kT}{10^6} \tag{3.113}$$



(a) Continuously varying atmosphere.

(b) Approximation to atmosphere (a) with nine layers, each of which is homogeneous.

Figure 3.19 Transmission calculations using homogeneous layers to approximate a stratified atmosphere. Each constituent designated by the subscript i has an absorption cross section $C_{\alpha ij}$ and a number density m_{ij} where the subscript j designates the layer.

equating the exponents in Eq. (3.110) and substituting yields

$$\beta_{cz} c_{ppm} z = \frac{C_\alpha c_{ppm} pz}{kT\,10^6} \tag{3.114}$$

which yields a simple relationship between the traditional absorption cross section (C_α) and the absorption per concentration depth (β_{cz}) of

$$C_\alpha = \frac{\beta_{cz} \bullet 10^6}{p/kT} [m^2] \tag{3.115}$$

which allows these terms to be used interchangeably.

3.4.2 Atmospheric Scattering

In addition to absorption losses, energy may leave the beam due to *scattering*. This process is most easily thought of as a disturbance of the electromagnetic field by the constituents in the atmosphere resulting in a change in the direction and spectral distribution of the energy in the beam. The details of scattering theory and the related radiation propagation are quite complex, and only a limited treatment will be given here. The simplified treatment needed for our purposes has been reduced from the rigorous coverage by Chandrasekhar (1960), Van de Hulst (1981), and Bohren and Huffman (1998). Scattering can be described in terms of one of three theoretical approximations. Rayleigh scatter is the result of the EM wave interacting with the very small particles or molecules that we think of as making up the atmosphere itself. It occurs when the particles are much smaller than the wavelength of the incident flux. Because it can be described in a closed form, we will treat it in some detail and use the formalism developed as a model for our more general treatment of Mie scatter and nonselective scatter. Mie scattering theory applies when the wavelength of the energy is approximately equal to the size of the particles. Aerosols, small dust particles, fossil fuel combustion products, and suspended sea salts are some of the major atmospheric constituents in this size range. Nonselective scattering theory applies when the particles are very large compared to the wavelength of the incident energy. Large dust particles, water droplets, and ice crystals are sources of nonselective scattering.

3.4.2.1 Rayleigh Scatter

The fractional amount of energy scattered into a solid angle at an angle θ from the propagation direction per unit length of transit in the medium was first characterized by Lord Rayleigh (1871) in an attempt to explain the blue color of the sky. His expression for randomly polarized flux is still valid and can be written as

$$\beta_r(\theta) = \frac{2\pi^2}{m\lambda^4}(n(\lambda)-1)^2(1+\cos^2\theta)[m^{-1}sr^{-1}] \tag{3.116}$$

where $\beta_r(\theta)$ is the Rayleigh angular scattering coefficient defined as the fractional amount of energy scattered into the direction θ per unit solid angle about θ per

unit length of transit, θ is the deflection angle from the beam direction, $n(\lambda)$ is the wavelength dependent index of refraction of the medium, and m is the number density as previously introduced. We explicitly write the functional dependence of the *angular scattering coefficient* $\beta_r(\theta)$ on θ to differentiate it from the total scattering coefficient β_r to be introduced shortly. Inspection of Eq. (3.116) shows an inverse dependence with wavelength to the fourth power and an inverse dependence on number density. The wavelength dependence can explain why the sky is blue, since more short-wavelength energy will be scattered out of the beam. It also explains why the solar disk is red at sunrise and sunset, since over long paths the fractional amount of energy in the beam will shift to longer wavelengths as the shorter wavelengths are disproportionately scattered out of the beam.

The inverse dependence of the Rayleigh law on number density appears to contradict our intuition. Since the number density near the earth's surface is higher, we should expect less scattering because of the inverse relationship. However, recognize that the index of refraction will also become larger as we approach the Earth's surface and that the term $[n(\lambda) - 1]^2$ will increase more rapidly than m as we approach the Earth. So our intuition is satisfied with scattering increasing in the lower atmosphere.

The Rayleigh expression for scattering expressed in Eq. (3.116) does not directly tell us the energy loss from the beam per unit length (cf. Fig. 3.20). In order to find this, we must integrate the angular scattering coefficient over all possible solid angles according to

$$\beta_r[m^{-1}] = \int \beta_r(\theta) d\Omega = \int_{\phi=0}^{2\pi} \int_{\theta=0}^{\pi} \frac{2\pi^2}{m\lambda^4} (n(\lambda)-1)^2 (1+\cos^2\theta) \sin\theta d\theta d\phi$$

$$= \frac{32\pi^3 (n(\lambda)-1)^2}{3\lambda^4 m} [m^{-1}] \tag{3.117}$$

where we have substituted $\sin\theta d\theta d\phi$ for $d\Omega$ (cf. Fig. 3.14) and β_r is the Rayleigh *scattering coefficient* and can be interpreted in a fashion analogous to the absorption coefficient β_α.

Also by analogy to the absorption cross section, we introduce the wavelength-dependent Rayleigh *scattering cross section* to be

$$C_r[\mathrm{m}^2] = \frac{\beta_r}{m} \tag{3.118}$$

to define the effective size of a scattering center in terms of how efficiently it scatters flux from the beam. The Rayleigh optical depth can then be expressed as

$$\delta_r = \beta_r z = m C_r z \tag{3.119}$$

which results in a transmission value of

$$\tau = e^{-\delta_r} \tag{3.120}$$

The angular scattering coefficient is often broken down into a product of two terms. The first term defines the *magnitude* or amount of scattering, and the

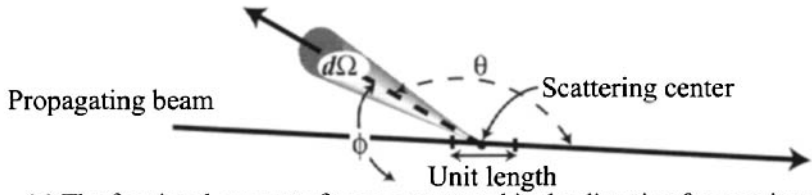
second, referred to as the *scattering phase function*, describes how the scattered energy is angularly distributed. Van de Hulst (1981) describes the scattering phase function $p(\theta)$ as the ratio of the energy scattered per unit solid angle into a particular direction to the average energy scattered per unit solid angle in all directions. This can mathematically be represented as

$$p(\theta) = \frac{\beta_r(\theta)}{\frac{\int \beta_r(\theta) d\Omega}{4\pi}} = \frac{4\pi\beta_r(\theta)}{\beta_r} \tag{3.121}$$

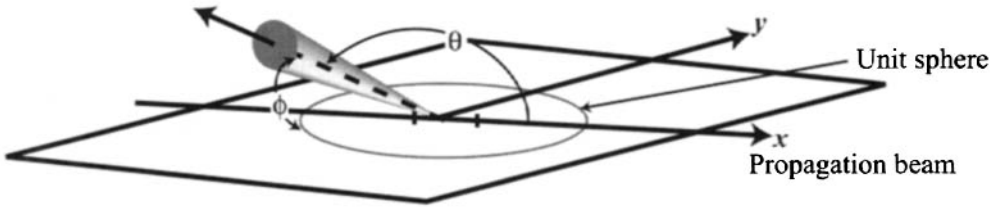
or rearranging we have

$$\beta_r(\theta) = \frac{\beta_r}{4\pi} p(\theta) \tag{3.122}$$

where we see that the magnitude of the angular scattering term is simply the scattering coefficient divided by 4π steradians, and the scattering phase function is normalized such that its integral over all solid angles (i.e., over the sphere about the scattering center) must equal 4π. We could therefore express the Rayleigh scattering law of Eq. (3.116) as



(a) The fractional amount of energy scattered in the direction θ per unit solid angle per unit length through the medium.



ϕ is an azimuthal angle about the primary ray relative to an arbitrary x,y plane

(b) Geometric terms.

Figure 3.20 Angular scattering coefficient.

$$\beta_r(\theta) = \frac{8\pi^2(n(\lambda)-1)^2}{3m\lambda^4} p(\theta) = \frac{8\pi^2(n(\lambda)-1)^2}{3m\lambda^4}\left[\frac{3}{4}(1+\cos^2\theta)\right] \quad (3.123)$$

where $p(\theta) = 3/4\,(1 + \cos^2\theta)$ is the Rayleigh scattering phase function and is plotted in Figure 3.21 along with other scattering phase functions.

It is important to recognize that Eq. (3.116) for Rayleigh scatter is for randomly polarized incident flux and that the scattered flux will be polarized. Bohren and Huffman (1998) derive the following expression for the *degree of polarization* induced by scattering from a small particle exposed to randomly polarized flux:

$$\text{DOP} = \frac{1\text{-}\cos^2\theta}{1+\cos^2\theta} \quad (3.124)$$

from which we see that the forward and back scattered flux will remain randomly polarized, but the scattered flux will become highly polarized at right angles to the incident beam. As a result, we can expect skylight to be polarized for angles well away from the sun (cf. Fig. 3.22). Recognize that the overall degree of polarization and the orientation of the polarization are complex functions of the relative orientation of the illumination and view angles, as well as the number of scattering interactions that may occur [cf. Lee (1998)].

3.4.2.2 Mie and Nonselective Scatter

It is not as straightforward to write an expression for the angular scattering coefficient associated with Mie scattering as it is for Rayleigh scattering. In general, the magnitude and scattering phase function are dependent on the number density, particle size distribution, and type (which controls the shape and complex index of refraction) of the scatterers. All the input parameters are generally not available for a rigorous solution to Mie scattering theory. As a result, empirical approximations are often used based on general classes of aerosols. The magnitude and scattering phase functions are available for a variety of atmospheric types based on data derived from empirical measurements [cf. Kneizys et al. (1988)]. In general, Mie-type scattering is loosely referred to as *aerosol scattering* and represented with a subscript *a*. Similar to Rayleigh scatter, we can define an angular scattering coefficient, $\beta_a(\theta)$, made up of a magnitude $\beta_a/4\pi$ and phase $p_a(\theta)$. Also by analogy, we can define an aerosol scattering cross section C_a and optical depth δ_a.

Mie scattering differs from Rayleigh in that it is highly forward-scattered as seen in Figure 3.21. The Rayleigh scatter is symmetric with significant and equal amounts of forward and backscatter. Both types of scattering show relatively little energy scattered at right angles to the propagating beam. Mie scattering is also less a function of wavelength. Curcio (1961) indicates that the magnitude of the scattering term dependency ranges from λ^{-2} to $\lambda^{0.6}$ for various atmospheres. Like Rayleigh scatter, Mie scatter will induce polarization characteristics to a randomly polarized beam.

The term *nonselective scattering* is used to describe the scattering from particles that are large compared to the wavelength of the EM energy. Nonselective scattering (denoted with the subscript *ns*) shows little dependency on wavelength

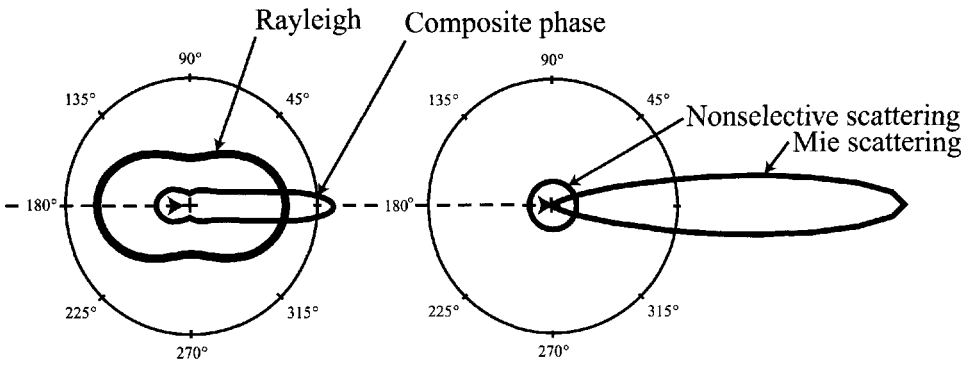


Figure 3.21 Shapes of scattering phase functions.

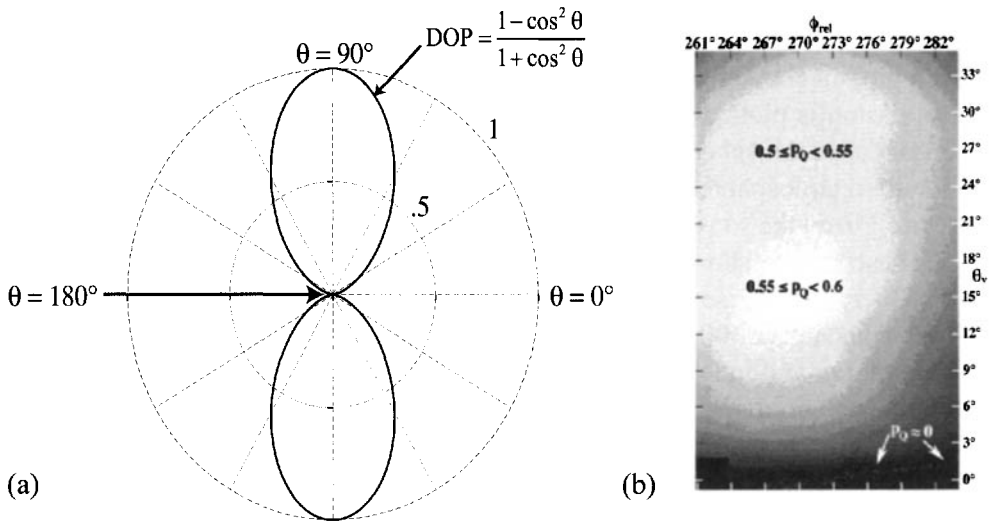


Figure 3.22 Polarization characteristics of scattered flux. (a) Polar plot of the degree of polarization induced by Rayleigh scatter of randomly polarized flux incident on a scattering center. (b) Illustration of the relative polarization of a clear sky. The image is a gray-scale representation produced from two images taken with linear polarizers oriented at right angles to each other such that the gray scale is proportional to $P_Q = \frac{L(0^0) - L(90^0)}{L(0^0) + L(90^0)}$. The sun is at an elevation of -0.46° and 0° azimuth. The gray scale quantization is set at ΔPQ =0.05.

and is approximately the same in all scattering directions (cf. Fig. 3.21). This type of scattering is most often associated with large dust particles, water droplets, ice, and hail (e.g., smoke and fog).

Scattering by all sources can be treated as independent and combined with absorption effects to yield the overall transmission along a beam according to

$$\begin{aligned} \tau &= e^{-(\delta_\alpha+\delta_r+\delta_a+\delta_{ns})} = e^{-\delta} \\ &= e^{-(\beta_\alpha+\beta_r+\beta_a+\beta_{ns})z} = e^{-\beta_{ext}z} \end{aligned} \tag{3.125}$$

where the *extinction coefficient* (β_{ext}) has been introduced as the sum of the absorption and scattering coefficients, and the total optical depth is simply designated δ. It is important to realize that each extinction term or optical depth term in Eq. (3.125) may be made up of contributions from several constituents in the atmosphere. Furthermore, for a nonhomogeneous atmosphere, the atmosphere would be approximated as a series of layers, and the optical depths summed in a piecewise fashion over each layer, as indicated in Figure 3.19.

3.5 CHARACTERISTICS OF THE EM SPECTRUM

In order to evaluate the relative importance of many of the concepts introduced in this chapter, we need to combine them as a function of wavelength. Figure 3.23 is a graphical attempt to combine several of these concepts. The atmospheric transmission is plotted as a function of wavelength for transmission through a single standard atmosphere (i.e., vertical transmission from Earth to space for the U.S. standard atmosphere). Also plotted is the exoatmospheric spectral irradiance [$Wm^{-2}\mu m^{-1}$] from the sun and the radiant exitance [$Wm^{-2}\mu m^{-1}$] from a 300 K blackbody. These curves show that from the visible through the short-wave infrared (SWIR), there will be several orders of magnitude more flux from the sun than from self-emission at 300 K. In general, we will refer to this spectral range as the *reflective region* and ignore thermal or self-emitted flux when we are considering this region (i.e., referring to Fig. 3.1, only type A, B, and C photons are relevant). In the 8 to 14 μm window, just the opposite occurs, and there will be several orders of magnitude more flux from self-emission than from reflected solar flux. Only approximately one photon in 4000 to 5000 (for a 10% reflector at 300 K) will be from the sun. Since we will rarely be able to make measurements where the noise limits are better than 1 part in 1000, we can generally ignore solar photons in the 8 to 14 μm region (i.e., referring to Fig. 3.3, only type D, E, and F photons are relevant). In the midwave infrared (MWIR) window between 3 and 5 μm, the situation is considerably more complicated. In this region, the number of solar and self-emitted photons are of the same order of magnitude. Therefore, under daylight conditions we must consider both solar and self-emitted photon paths. The photon flux in this spectral region simplifies somewhat at night and becomes similar to the 8 to 14 μm region. It is important to realize that Figure 3.23 represents a simplified case with direct sunlight and a 300 K target. Under low-sun conditions, or for hotter objects, the relative amounts of self-emitted flux can be more important, particularly at the longer wavelength end of the SWIR region.

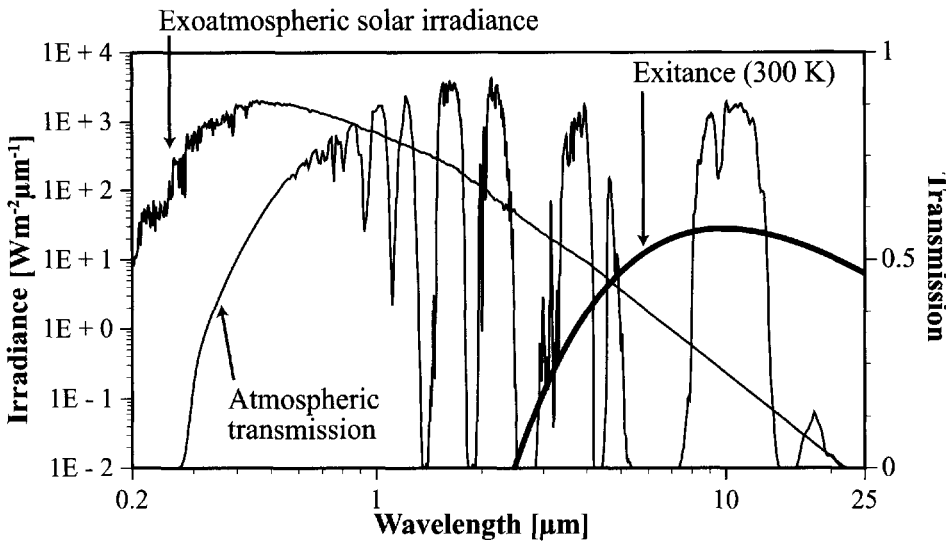


Figure 3.23 Comparison of atmospheric transmission, solar irradiance, and self-emission spectra.

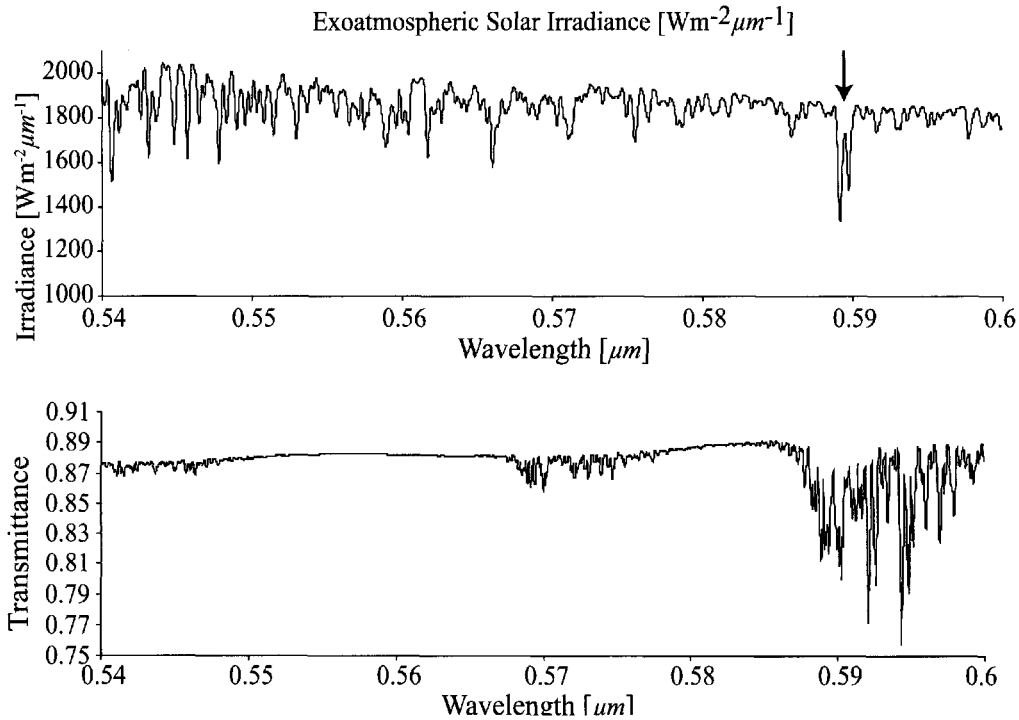


Figure 3.24 Expanded solar irradiance and transmission plots showing absorption line effects.

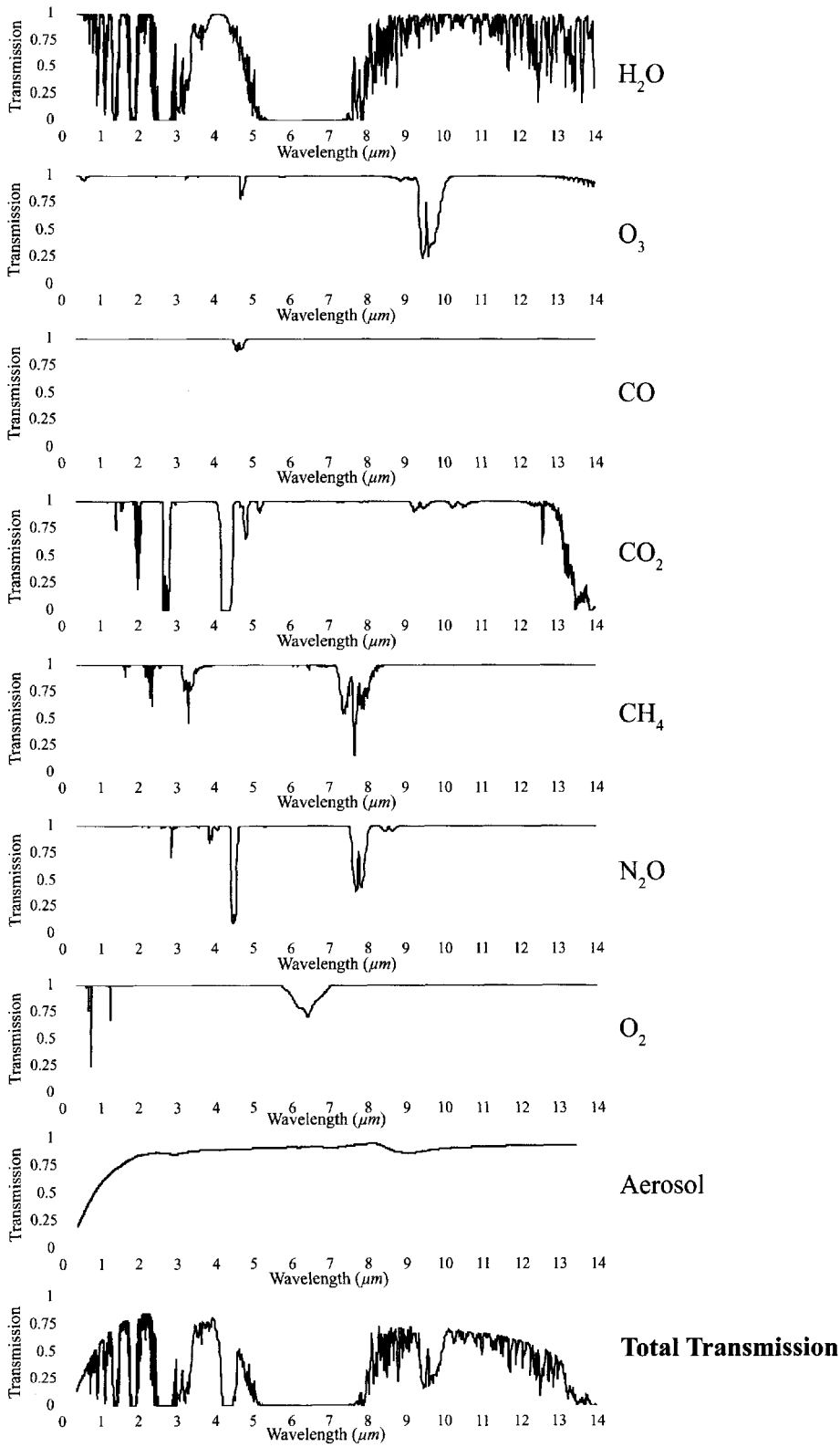


Figure 3.25 Absorption spectra of various atmospheric constituents and overall atmospheric transmission as derived from MODTRAN [cf. Berk et al. (1989)].

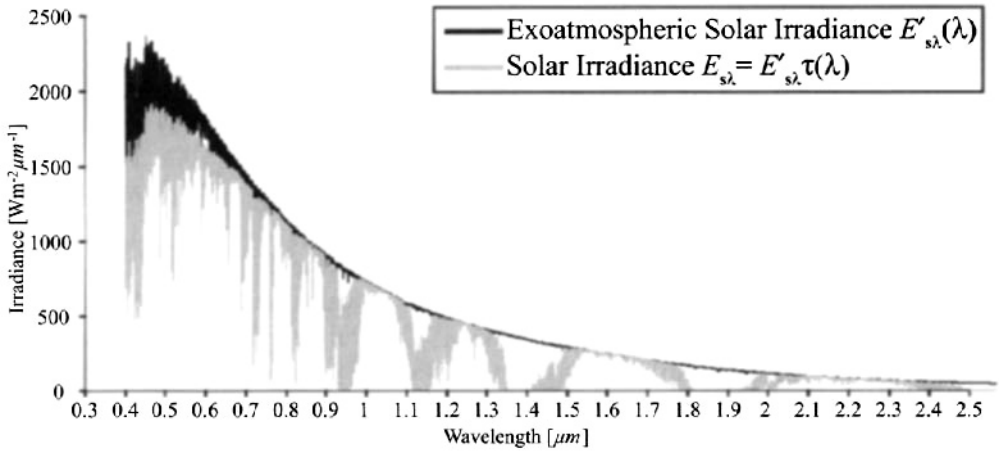


Figure 3.26 Effect of atmospheric transmission on the solar spectral irradiance reaching the earth.

The wavelength scale of Figure 3.23 makes the solar irradiance and transmission curves deceptively smooth. In Figure 3.24 [derived from the MODTRAN code described in Berk et al. (1989)], we show an expanded wavelength scale that more clearly shows some of the spectral structure. The absorption line highlighted in the irradiance spectra is one of the *Fraunhofer lines*. These are very narrow deep lines caused by absorption in the solar atmosphere of the near-blackbody radiation from the sun. If we were to image the earth in this spectral region, even on a sunlit day, we would see little or no reflected solar flux. In Section 6.3, we will see how we can take advantage of this to study solar-induced luminescence. The transmission plot shows how the transmission spectra is actually made up of the overlapping absorption characteristics of the various constituents in the atmosphere. Figure 3.25 shows the absorption spectra of various constituents in the atmosphere for a single pass through the U.S. standard atmosphere along a 45° solar illumination path. The spectra are cascaded together in the lowest curve representing the actual atmospheric transmission. In Figure 3.26, we see the exoatmospheric spectral irradiance ($E'_{s\lambda}$) cascaded together with the atmospheric transmission spectra [$\tau(\lambda)$] to yield the spectral irradiance reaching the Earth, $E_{s\lambda}$. These spectra show how different absorbers will be important in different wavelength regions. It also shows where we might select spectral bands to avoid or to study certain atmospheric effects.

In this chapter we have defined the basic radiometric terms and shown how radiometric concepts can be used to describe radiation propagation in the atmosphere. In the next chapter we will forge the first few links in the image chain by using these principles to derive the governing equation for the radiance reaching the sensor.

3.6 REFERENCES

Berk, A., Bernstein, L.S.' & Robertson, D.C. (1989). "MODTRAN: a moderate resolution model for LOWTRAN 7," GL-TR-89-0122, Spectral Sciences, Burlington, MA.

Bohren, F., & Huffman, D.R. (1998). *Absorption and Scattering of Light by Small Particles*. John Wiley & Sons, NY.

Chandrasekhar, S. (1960). *Radiative Transfer*. Dover, Minneola, NY.

Chen, H.S. (1985). *SPACE Remote Sensing Systems*. Academic Press, Orlando, FL.

CIE. (1970). "International lighting vocabulary," CIE Publication No. 17 (E-1.1), International Commission on Illumination (CIE), Paris.

Curcio, J.A. (1961). Evaluation of atmospheric aerosol particle size distribution from scattering measurements in the visible and infrared. *Journal of the Optical Society of America*, Vol. 51, No. 5, pp. 548-551.

Einstein, A. (1905). "On a Heuristic Viewpoint Concerning the Production and Transformation of Light." *Annalen der Physik*, Vol. 17 [cf. English translation by A.B. Arons and M.B. Peppard (1968), *American Journal of Physics*, Vol. 33, No. 5, pp. 367-374].

Frohlich, C. (1977). Contemporary measures of the solar constant. In *The Solar Output and Its Variation*, edited by O.R. White, Colorado Associated University Press, Boulder, CO.

Grum, F., & Becherer, R.J. (1979). Optical radiation measurements. In *Vol. 1, Radiometry*. Academic, NY.

Hecht, E. (1990). *Optics*. 2nd ed., Addison-Wesley, Reading, MA.

Kaufman, Y.J. (1982). Solution of the equation of radiative transfer for remote sensing over nonuniform surface reflectivity. *Journal of Geophysical Research*, Vol. 87, No. C6, pp. 4137-4147.

Kneizys, F.X., Shettle, E.P., Abreu, L.W., Chetwynd, J.H., Anderson, G.P., Gallery, W.O., Selby, J.E.A., & Clough, S.A. (1988). "Users Guide to LOWTRAN7," AFGL-TR-88-0177, Environmental Research Papers, No. 1010, Air Force Geophysics Laboratory, Optical/Infrared Technology Division, Hanscom AFB, MD.

Lee, R.L. (1998). Digital imaging of clear-sky polarization. *Applied Optics*, Vol. 37, No. 9, pp. 1467-1476.

Nicodemus, F. E., ed. (1976). "Self-study manual on optical radiation measurements: Part 1-Concepts, Chaps. 1 to 3," NBS Technical Note 910-1, U.S. Dept. of Commerce, National Bureau of Standards, U.S. Printing Office, Washington, DC.

Piech, K.R., & Schott, J.R. (1974). Atmospheric corrections for satellite water quality studies. Proceedings of the SPIE, Vol. 57, pp. 84-89.

Planck, M. (1901). *Annalen der Physik*, Vol. 4, No. 3, p.553.

Lord Rayleigh (J.W. Stratt) (1871). *Philosophical Magazine*, Vol. 41, pp. 107-120, 274-279.

Stokes, G.G. (1852). On the composition and resolution of streams of polarized light from different sources. *Trans. Camb. Philos. Soc.*, Vol. 9, pp. 399-416 (reprinted in *Mathematical and Physical Papers*, Vol. III, Cambridge University Press, Cambridge, 1901).

Thekaekara, M.P. (1972). Evaluating the light from the sun. *Optical Spectra*, Vol. 6, No. 3, pp. 32-35.

U.S. Air Force Geophysics Laboratory. (1985). *Handbook of Geophysics and the Space Environments.*, 4th ed., Air Force Geophysics Laboratory, Bedford, MA.

Van de Hulst, H.C. (1981). *Light Scattering by Small Particles*. Dover, Minneola, NY.

This page intentionally left blank

CHAPTER 4
THE GOVERNING EQUATION FOR RADIANCE REACHING THE SENSOR

In the last chapter we encountered the basic concepts of atmospheric propagation and the fundamentals of radiometry. In this chapter we will use these principles to derive an expression for the radiance reaching a remote sensor. This governing equation must describe all the significant paths illustrated in Figures 3.1 and 3.3. To determine the relative importance of each of the terms in the governing equation, we will perform a simple sensitivity analysis (Sec. 4.6) on the final equation. This governing equation represents several major links in the imaging chain that are critical to quantitative radiometric image analysis. As a result, we will derive a governing radiometric propagation equation that is quite complex. In many cases, this level of detail may not be required. We will find that simplifying assumptions can often be made without introducing significant error. At other times, unexpected errors can occur because we violate some of the simplifying assumptions. For these cases, it is important to have available the more complete treatment. The approach we will use deviates somewhat from more traditional radiative transfer discussions. We try to follow the flow of photons in as intuitive a fashion as is feasible while yielding results consistent with a more rigorous traditional radiative transfer approach, [cf. Chandrasekhar (1960)].

4.1 IRRADIANCE ONTO THE EARTH'S SURFACE

In developing the governing equation, we follow the flow of photons to the sensor. Initially, we will consider only solar photons, with the self-emitted photons introduced in Section 4.4. In tracing the photon flux, we will first try to determine an expression for all the flux incident on a reflecting surface.

4.1.1 Solar Irradiance

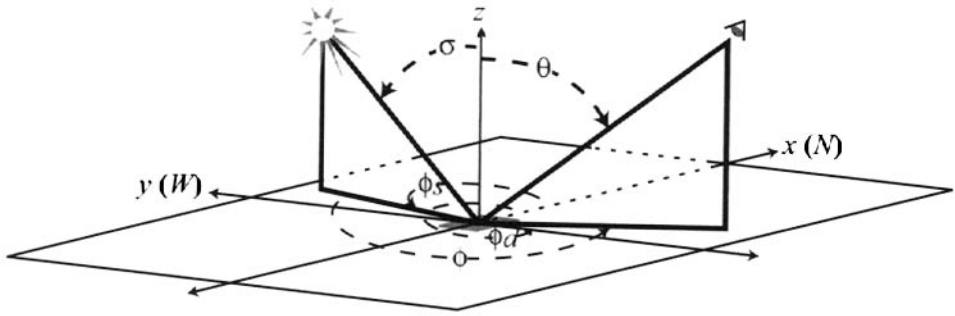
In general, the primary source of flux incident on a reflecting surface is the sun. In order to use the generic expressions developed in the previous section for the specific case of interest to us, we must introduce two related sets of orientation angles. These angles are depicted in Figure 4.1. The first set of angles describes the relative source-target-sensor orientation in the world coordinate system. In this system, the x,y plane is defined to be tangent to the Earth at the target, with the x axis due north, the y axis due west, and the z axis normal to the Earth at the target. The sun location is then defined by its zenith angle σ from the normal to the Earth and an azimuthal angle ϕ_s defined as the angle going west from the x axis (north) to the projection of the sun onto the x,y plane. The sensor, or detector, is located through a view angle θ from the normal to the Earth to the sensor and an azimuthal angle ϕ_d defined using the same measurement convention as ϕ_s. The relative azimuthal angle between the source and the detector we denote simply as ϕ.

In many cases, we need to define the source-target-sensor orientation in a target-centered coordinate system [cf. Fig. 4.1(b)]. Here we define the x,y plane as the plane of the target or the plane tangent to the target at the point of interest with the x axis located along the projection of the Earth normal onto the plane of the target (defined to be north when the target plane and the plane of the Earth are coincident). The sun location relative to the target plane is defined by the solar zenith relative to the target-centered z axis (i.e., the normal to the target) identified as σ' and an azimuthal angle ϕ_s' measured from the target's x axis in the positive y direction to the projection of the sun onto the target plane. Similarly, the sensor location is defined by the angle from the normal to the target to the sensor (θ') and the azimuthal angle ϕ_d' to the projection of the detector onto the target plane. The relative azimuthal angle in the target plane is denoted by ϕ'.

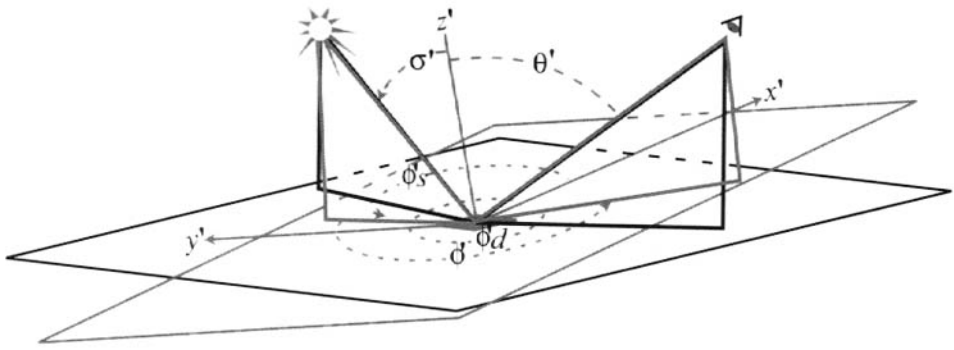
When the target is a horizontal surface, the two coordinate systems coincide and $\theta = \theta'$, $\sigma = \sigma'$, and $\phi = \phi'$. For all other cases, a set of relative orientation angles is required to define the relationship between the coordinate systems. These angles are not required for our treatment here, so a naming convention will not be introduced. The reader should recognize that the geometric convention introduced here is arbitrary, and many naming conventions and reference coordinate systems are used in the literature.

Given the angular definitions of Figure 4.1, we can define the solar spectral irradiance onto a target as

$$E_{s\lambda}\cos\sigma' = E'_{s\lambda}\tau_1(\lambda)\cos\sigma' = E'_{s\lambda}e^{-\delta(\lambda)\sec\sigma}\cos\sigma' [W\mathrm{m}^{-2}\mu\mathrm{m}^{-1}] \qquad (4.1)$$



(a) World coordinate system.



(b) Target coordinate system.

Figure 4.1 Definition of orientation angles.

where $E_{s\lambda}'$ is the exoatmospheric spectral irradiance [$Wm^{-2}\mu m^{-1}$] onto a surface perpendicular to the incident beam [cf. Eq. (3.57)] and $\tau_1(\lambda)$ is the atmospheric transmission along the sun-target path, which can be written as

$$\tau_1(\lambda) = e^{-\delta_1(\lambda)} \approx e^{-\delta(\lambda)\sec\sigma} \tag{4.2}$$

where $\delta_1(\lambda)$ is the optical depth along the sun-target path, which for single scattering in a plane, parallel atmosphere, and ignoring refractive index changes can be approximated as

$$\delta_1 = \delta(\lambda)\sec\sigma \tag{4.3}$$

where $\delta(\lambda)$ is defined to be the optical depth vertically through the Earth's atmosphere and $\sec\sigma = 1/\cos\sigma$. This approximation will be used in the next section and is reasonably valid for clear atmospheres and σ ranging from 0 to 60°.

It is important to realize how much the solar irradiance can change due to the relative orientation of the target to the sun even for surfaces with relatively shallow slopes (cf. Fig. 4.2). The solar irradiance will also change with atmospheric make-

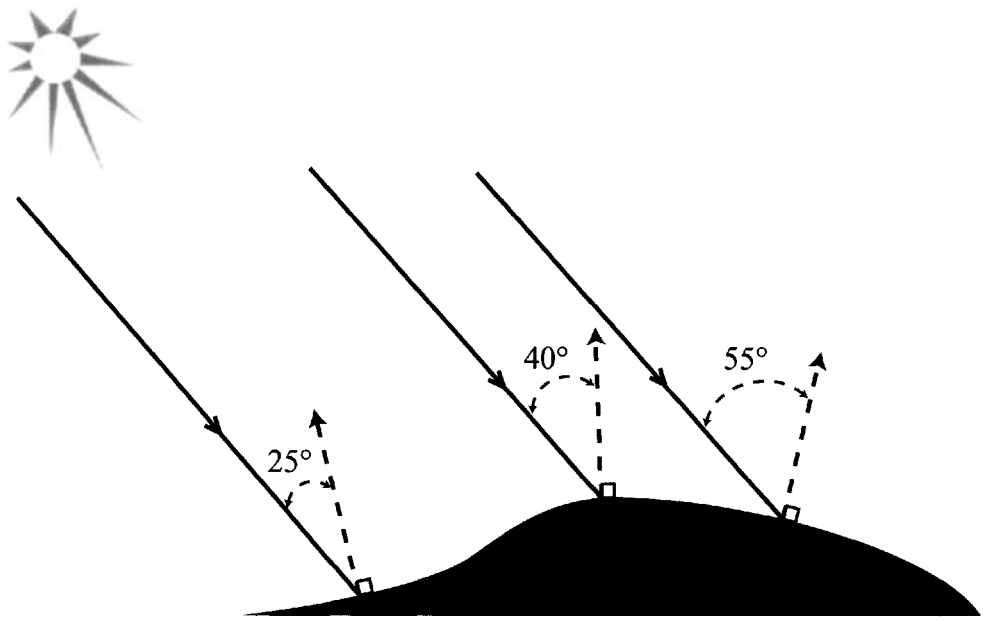


Figure 4.2 Effects of target slope on solar irradiance.

up (e.g., number and type of aerosols) and with elevation, since from inspection of Eq. (3.104) it is clear that the total optical depth will decrease as the elevation increases. This is particularly important because the atmosphere is denser near the earth's surface. The lowest layers make large contributions to the optical depth that are eliminated for targets at higher elevations.

4.1.2 Downwelled Radiance (Skylight)

In addition to the direct solar irradiance term, we also recognized in Figure 3.1 that solar photons would be incident on the target due to scattering from the atmosphere. We refer to this downward scattering as *downwelled radiance* or, in the visible, as *skylight*. In order to characterize the flux onto the target due to downwelled radiance, let's first consider the contribution from a small volume in the atmosphere, as shown in Figure 4.3. The spectral irradiance onto the face of a volume element located perpendicular to the incident flux is

$$E_{v\lambda} = E'_{s\lambda}\tau_{L1}(\lambda)[Wm^{-2}\mu m^{-1}] \tag{4.4}$$

where $\tau_{L1}(\lambda)$ is the transmission along the first leg of the L-shaped path from the sun to the scattering volume element to the target. If we take a unit area on the face of the volume perpendicular to the propagating ray and consider the element of spectral radiant intensity scattered toward the target in traversing a unit length along the beam (i.e., from the unit volume), we have

$$dI_{v\lambda}[\mathrm{Wsr^{-1}\mu m^{-1}}] = E_{v\lambda}\beta_{sca}(\lambda,\theta_v)dV[\mathrm{Wm^{-2}\mu m^{-1}m^{-1}sr^{-1}m^3}] \tag{4.5}$$

where $\beta_{sca}(\lambda,\theta_v)$ is the spectrally dependent angular scattering coefficient for the composite atmosphere, θ_v is the angle between the incident ray and the ray to the target, and dV is the volume element. Adapting Eqs. (3.52) and (3.59) for this case, we have the element of spectral irradiance onto the target from the volume element:

$$dE_{d\lambda}[Wm^{-2}\mu m^{-1}] = \frac{dI_{v\lambda}\tau_{L2}(\lambda)\cos\sigma_v}{r_v^2} = \frac{E_{v\lambda}\beta_{sca}(\lambda,\theta_v)\tau_{L2}(\lambda)\cos\sigma_v dV}{r_v^2} \tag{4.6}$$

or, on substituting Eq. (4.4),

$$dE_{d\lambda} = \frac{E'_{s\lambda}\tau_{L1}(\lambda)\beta_{sca}(\lambda,\theta_v)\tau_{L2}(\lambda)\cos\sigma_v dV}{r_v^2} \tag{4.7}$$

where $\tau_{L2}(\lambda)$ is the transmission along the second leg of the L-shaped sun-volume-target path, σ_v is the angle between the normal to the target and the ray from the volume, and r_v [m] is the path length along the volume-target path. The total downwelled spectral irradiance $E_{d\lambda}$ could then be computed by integrating over the hemisphere above the target, yielding

$$E_{d\lambda} = \int \frac{E'_{s\lambda}\tau_{L1}(\lambda)\tau_{L2}(\lambda)\beta_{sca}(\lambda,\theta_v)\cos\sigma_v dV}{r_v^2} \tag{4.8}$$

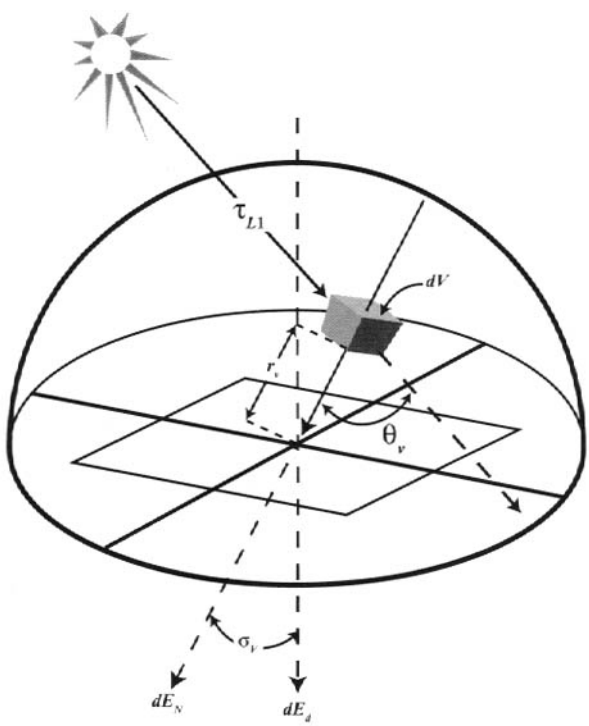


Figure 4.3 Contributions to downwelled irradiance from a unit volume.

The limits of integration would extend to space in all directions above the target, and all the parameters on the right-hand side of Eq. (4.8) except $E_{s\lambda}{}'$ are dependent on location (i.e., their dependence on dV must be accounted for to perform the integration). This approach is the most straightforward for seeing the dependence of the downwelled radiance (or irradiance) on the contribution from the volume scattering element. However, a numerical integration of Eq. (4.8) is somewhat laborious to set up.

A slightly more tractable approach to finding the downwelled radiance from the sky dome is to consider the directional downwelled radiance reaching the target from a particular direction and then integrating over all directions. This approach is particularly attractive because at times we are interested in the downwelled radiance from a particular direction. This approach is also attractive because it roughly parallels the numerical integration used in the MODTRAN/LOWTRAN-style computer codes that we will address in Chapter 7 [cf. Kneizys et al. (1988)]. If we consider a scattering volume element dV [m^3] subtended by the element of solid angle $d\Omega$ as illustrated in Figure 4.4, we can write the relationship

$$dV = dAdr = d\Omega r_v^2 dr \tag{4.9}$$

where dA is the area element subtended by the solid angle element $d\Omega$ at a radial distance r_v from the target and dr is the element of radial distance. By combining Eqs. (4.4) and (4.5), we can write the element of spectral radiant intensity [$Wsr^{-1}\mu m^{-1}$] scattered toward the target from a volume element as

$$dI_{v\lambda} = E'_{s\lambda}\tau_{L1}(\lambda)\beta_{sca}(\lambda,\theta)dV \tag{4.10}$$

Substituting Eq. (4.9) for dV, we have

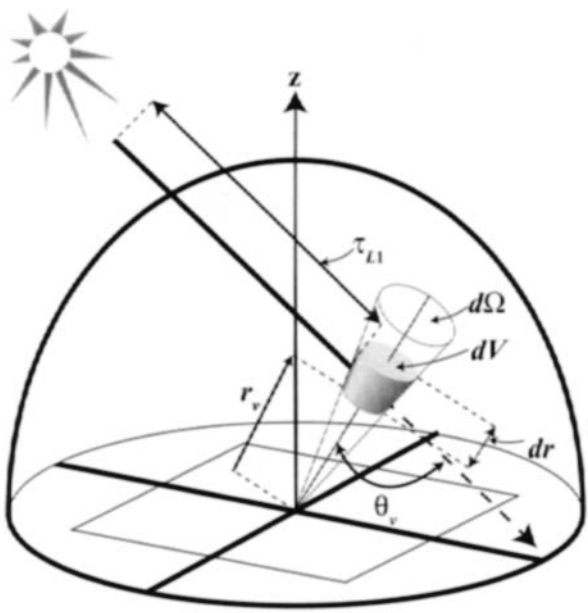


Figure 4.4 Procedure for calculation of angular downwelled radiance.

$$dI_{v\lambda} = E'_{s\lambda}\tau_{L1}(\lambda)\beta_{sca}(\lambda,\theta_v)d\Omega r_v^2 dr \tag{4.11}$$

Using Eq. (3.52), the element of spectral irradiance [$Wm^{-2}\mu m^{-1}$] onto a surface perpendicular to the scattered ray at the target is then given by

$$dE_{n\lambda} = \frac{E'_{s\lambda}\tau_{L1}(\lambda)\tau_{L2}(\lambda)\beta_{sca}(\lambda,\theta_v)d\Omega r_v^2 dr}{r_v^2} \tag{4.12}$$

The element of downwelled spectral radiance [$Wm^{-2}sr^{-1}\mu m^{-1}$] along the scattered ray at the target's surface can then be expressed from the definition of radiance as

$$dL_{d\lambda}(\sigma,\phi) = \frac{E'_{s\lambda}\tau_{L1}(\lambda)\tau_{L2}(\lambda)\beta_{sca}(\lambda,\theta_v)d\Omega dr}{d\Omega} \tag{4.13}$$

where we have introduced the generic direction angles σ and ϕ to describe the radiance reaching the target from the direction described by the zenith angle σ and the azimuthal angle ϕ. The total downwelled spectral radiance reaching the target from the σ,ϕ direction can be expressed by integrating over r from zero to the top of the atmosphere according to

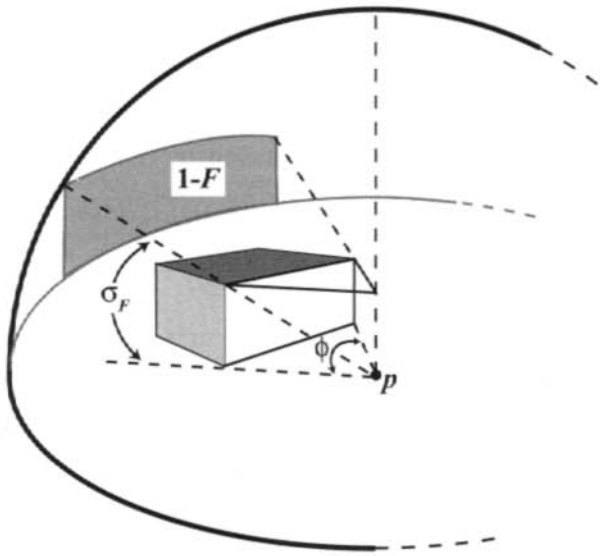
$$L_{d\lambda}(\sigma,\phi) = E'_{s\lambda}\int \tau_{L1}(\lambda)\tau_{L2}(\lambda)\beta_{sca}(\lambda,\theta_v)dr \tag{4.14}$$

The L-path transmission values (τ_{L1} and τ_{L2}) can be computed for each r value by numerical integration of the optical depth if the number densities and extinction cross sections are known along the L-path. Finally, if the angular scattering function is known along r in the σ,ϕ direction, then Eq. (4.14) can be solved numerically.

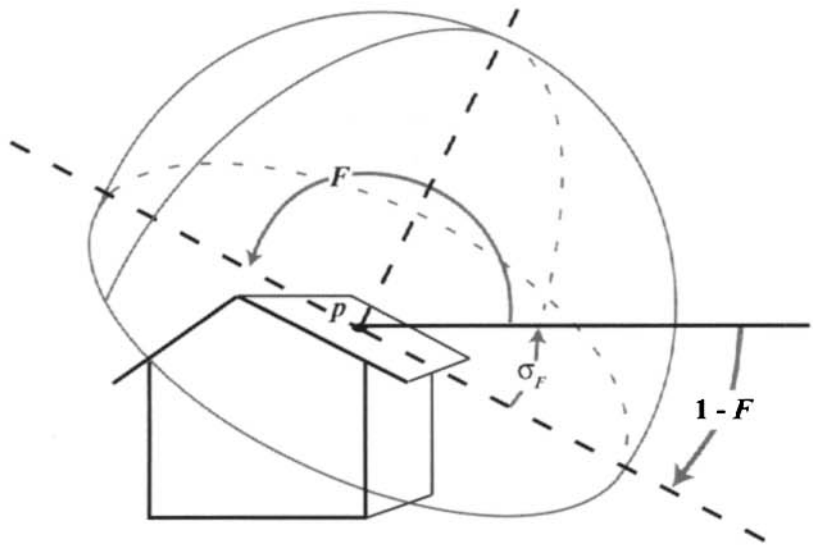
The total downwelled spectral irradiance can be computed from the angular downwelled radiance by integration over the hemisphere above the target according to

$$E_{d\lambda} = \int L_{d\lambda}(\sigma,\phi)\cos\sigma\, d\Omega = \int_{\varphi=0}^{2\pi}\int_{\sigma=0}^{\frac{\pi}{2}} L_{d\lambda}(\sigma,\phi)\cos\sigma\sin\sigma\, d\sigma\, d\phi \tag{4.15}$$

In the analysis of downwelled radiance presented thus far, we have implicitly assumed that the target is horizontal and unobstructed such that the entire hemisphere above the target is sky. This is a situation that generally only occurs in textbooks. In reality, if the target has any slope or if there are adjacent objects obstructing the sky dome, the downwelled radiance onto the target will be reduced (cf. Fig. 4.5). The most rigorous way to compute the effect from adjacent objects is to simply change the limits of integration of Eq. (4.15) so that the integral only includes that portion of the hemisphere that is sky. In most cases, it is more convenient to compute separately, or estimate, the fraction of the hemisphere above the target which is sky (F). This can most easily be accomplished by simply integrat-



(a) The building blocks the shaded portion of the sky from irradiating the point p.



(b) The slope of the roof reduces the amount of sky seen by the point p.

Figure 4.5 Shape factor F of the exposed sky.

ing the solid angle element ($d\Omega$) over the solid angle obstructed by background and dividing by the solid angle of the hemisphere (2π). This yields the fraction of the hemisphere that is background, with the remaining fraction (i.e., 1 minus the background fraction) being sky. The computational form of this can be expressed as

$$F = 1 - \frac{\int d\Omega}{2\pi} = 1 - \frac{\int\int \sin\sigma d\sigma d\phi}{2\pi} \tag{4.16}$$

Two simple examples of the use of this expression are to compute the shape factor (F) for the case of an adjacent building and for a sloped roof, as illustrated in Figure 4.5. Piech and Walker (1971) solved for the case of an adjacent building to yield an approximation, for the point p in Figure 4.5(a), of the form

$$F = 1 - \frac{\int_{\phi=0}^{\phi} \int_{\sigma=\frac{\pi}{2}-\sigma_F}^{\frac{\pi}{2}} \sin\sigma d\sigma d\phi}{2\pi} = 1 - \frac{\phi\cos\left(\frac{\pi}{2} - \sigma_F\right)}{2\pi} \tag{4.17}$$

This would produce the intuitively correct result if p were at the base of the building,

$$F \cong 1 - \frac{\pi\cos\left(\frac{\pi}{2} - \frac{\pi}{2}\right)}{2\pi} = 0.5 \tag{4.18}$$

i.e., half the sky is visible.

For the case of a point on a sloped roof [cf. Fig. 4.5(b)], F can be approximated from

$$F \cong 1 - \frac{\int_{\phi=0}^{\pi} \int_{\sigma=\frac{\pi}{2}-\sigma_F}^{\frac{\pi}{2}} \sin\sigma d\sigma d\phi}{2\pi} = 1 - \frac{1}{2}\cos\left(\frac{\pi}{2} - \sigma_F\right) \tag{4.19}$$

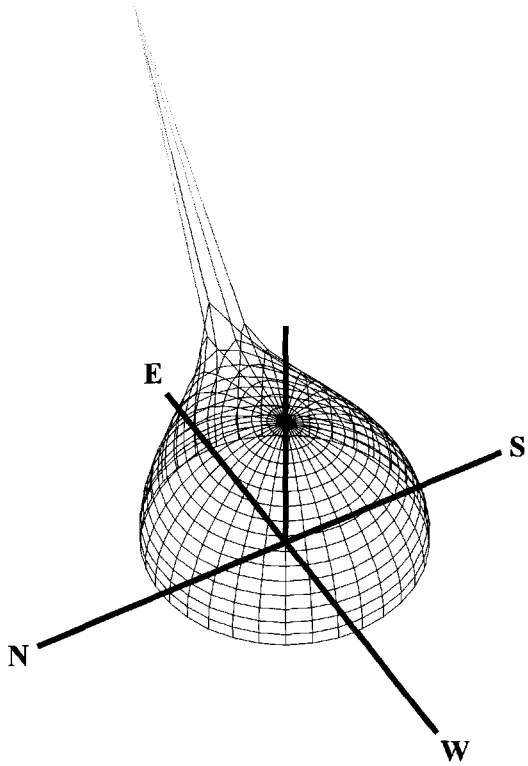
The downwelled irradiance from the sky onto the target can then be expressed as

$$E_{d\lambda sky} = \int\int L_{d\lambda}(\sigma,\phi)\cos\sigma\sin\sigma d\sigma d\phi \tag{4.20}$$

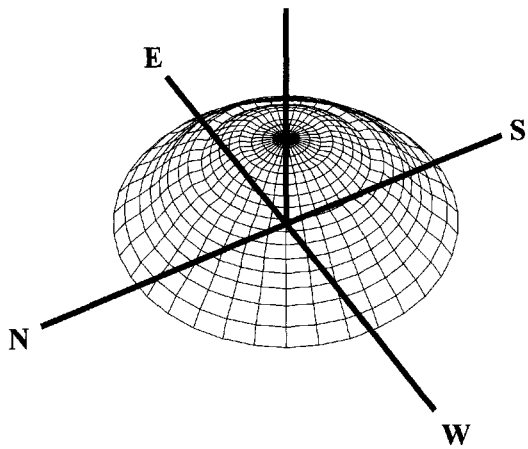
where the limits on the integrals are set to encompass the fraction of the hemisphere above the target that is sky. This can be approximated by

$$E_{d\lambda sky} = FE_{d\lambda} = F\int_{\phi=0}^{2\pi}\int_{\sigma=0}^{\frac{\pi}{2}} L_{d\lambda}(\sigma,\phi)\cos\sigma\sin\sigma d\sigma d\phi \tag{4.21}$$

where $E_{d\lambda}$ is the total downwelled spectral irradiance from the sky in the absence of background effects, and $E_{d\lambda\text{sky}}$ is the spectral downwelled irradiance from the fraction of the hemisphere above the target that is sky. Clearly, Eq. (4.21) will be a better approximation when the radiance from the sky is approximately a constant. Figure 4.6 illustrates that the downwelled radiance from the sky on a clear day is



(a) Polar plot of downwelled radiance from the morning sky in the visible region of the spectrum.



(b) Polar plot of downwelled radiance from the morning sky in the LWIR region of the spectrum.

Figure 4.6 Angular variation in downwelled radiance.

a function of both σ and ϕ in the reflective region and at least of σ in the thermal region. In general, under clear conditions it is a well-behaved, slowly varying function, so Eq. (4.21) should not introduce large errors. However, when scattered clouds are present or when very detailed analysis is required, a rigorous analysis of Eq. (4.20) with the proper limits of integration may be required.

4.1.3 Reflected Background Radiance

The fraction of the hemisphere above the target that is obscured by background objects (1–*F*) also produces some photon flux onto the target. If we let the spectral radiance reflected from a background toward the target be $L_{b\lambda}(\sigma,\phi)$ [$Wm^{-2}sr^{-1}\mu m^{-1}$], where the angles σ and ϕ denote the direction from which the background radiance comes, then the effects of background radiance are analogous to the downwelled sky radiance. The irradiance onto the target from the portion of the hemisphere above the object that is background (i.e., nonsky) can be expressed as:

$$E_{b\lambda} = \int\int L_{b\lambda}(\sigma,\phi)\cos\sigma\sin\sigma d\sigma d\phi \approx (1-F)L_{b\lambda avg}\pi \tag{4.22}$$

where we have assumed that the radiance from the background is approximately constant with a mean value of $L_{b\lambda avg}$, and the limits of integration are set to encompass the fraction of the hemisphere above the target that is background. Where the constant background assumption introduces too large an error, the full integral form of Eq. (4.22) can be used. We have not yet described a functional expression for $L_{b\lambda}$ and will defer that until we have introduced reflected radiance in the next section (4.2.2).

In summary, we have identified three sources of irradiance onto the target that yield an expression for spectral irradiance onto the target of

$$E_{\tau\lambda} = E'_{s\lambda}\cos\sigma'\tau_1(\lambda) + FE_{d\lambda} + (1-F)L_{b\lambda avg}\pi \tag{4.23}$$

4.2 REFLECTED SOLAR IRRADIANCE AND BIDIRECTIONAL REFLECTANCE

In order to convert the irradiance onto the target into radiance toward the detector, we need to consider the reflectance properties of materials. In general, the reflectance properties are a function of wavelength, illumination angle, and viewing angle. We need to develop a means to express the full impact of these dependencies on the reflected radiance and to develop a simpler expression for cases when full angular reflectance data are not available.

4.2.1 Ways to Characterize Reflectance

In Chapter 3, we encountered what we will now define to be the total spectral reflectance as

$$r(\lambda) = \frac{M_\lambda}{E_\lambda} \tag{4.24}$$

This expression for reflectance, while perfectly valid, fails to provide us with any information about the directional distribution of the reflected flux. Our experience tells us that the directional characteristics of reflectance vary considerably from mirrorlike surfaces (*specular*) to surfaces that appear to have little or no directional character to their reflectance (Lambertian or *diffuse*). Figure 4.7 depicts the angular reflectance characteristics of several idealized surfaces, as well as a more complex object. A perfectly specular object behaves like a mirror with flux only reflected into the direction exactly opposite to the incident ray. A nearly specular object will

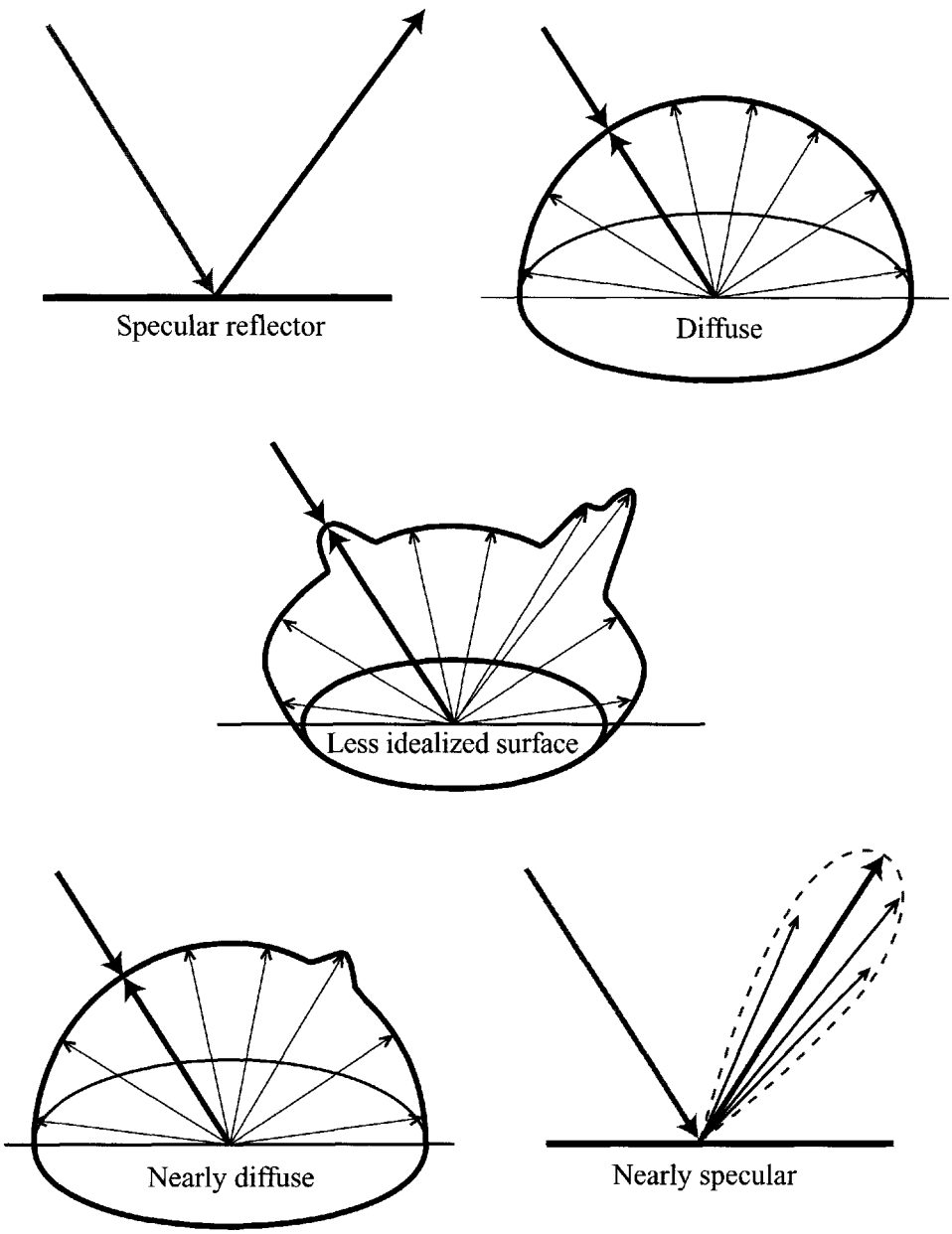


Figure 4.7 Reflectance characteristics of idealized surfaces.

appear to have most of the reflected energy concentrated in a cone about the specular ray. A perfectly diffuse surface appears to have the same amount of reflectance in all directions, while a nearly diffuse object will generally appear brighter in the specular direction. A less-idealized surface may appear brighter in the specular and backscatter directions and darker when viewed at grazing angles. Rough surfaces, such as tree canopies, will typically have a strong return from the backscatter direction, which is referred to as the "hot spot" (cf. Fig. 1.4). Recall from Chapter 3 that visual brightness or apparent reflectance is directly proportional to radiance, so that the vectors in Figure 4.7 can be treated as the magnitude of the radiance in each direction. The surface that encloses the vectors can be thought of as a probability distribution function for the radiance in any direction. More formally (cf. Fig. 4.8), we can define the bidirectional reflectance to be the ratio of the radiance scattered into the direction described by the orientation angles θ_o and ϕ_o to the irradiance from the θ_i, ϕ_i direction, i.e.,

$$r_{\text{BRDF}} = \frac{L(\theta_o, \phi_o)}{E(\theta_i, \phi_i)} [sr^{-1}] \tag{4.25}$$

The *bidirectional reflectance distribution function* (*BRDF*) describes these bidirectional reflectance values for all combinations of input-output angles. The BRDF values will also change as a function of wavelength, so a complete characterization would include the wavelength-dependent BRDF, using the spectral values for

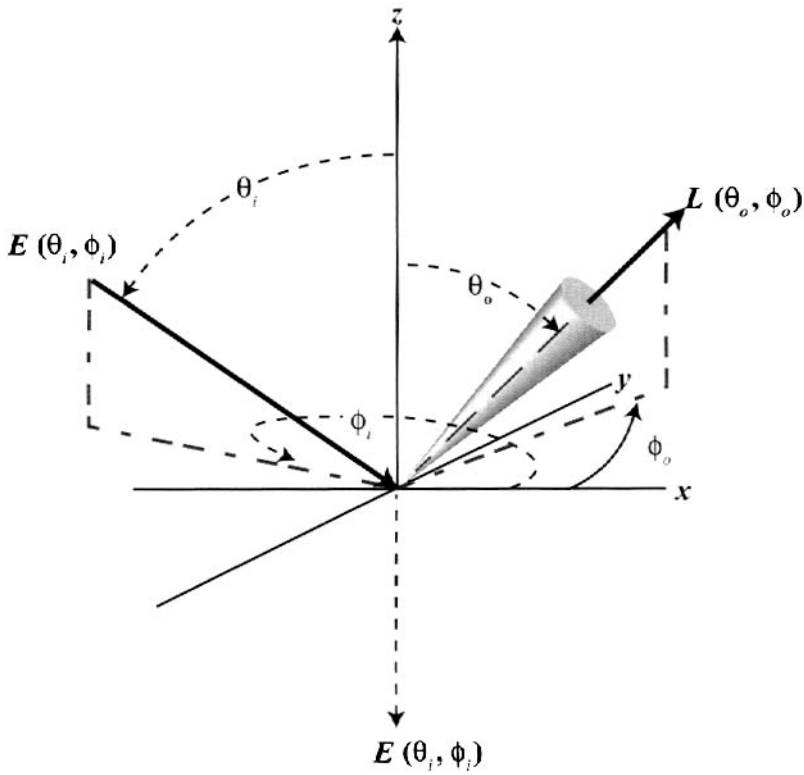


Figure 4.8 Bidirectional reflectance concept.

irradiance and radiance in Eq. (4.25). The BRDF is actually a scattering function analogous to the angular scattering coefficient, $\beta_{sca}(\lambda,\theta)$, introduced in Chapter 3 to describe atmospheric scattering.

BRDFs can be measured in the laboratory [cf. Feng et al. (1993)] or in the field [cf. Deering (1988)]. However, because of the large numbers of angles needed to fully characterize all possible combinations of source and sensor orientations, it is a cumbersome process. As a result, BRDFs are available for only a relatively restricted number of materials and land-cover types, and robust BRDF sample sets (i.e., BRDF values for different samples of the same land cover) are quite scarce. Another alternative when more complete BRDF data are required is to use a model to estimate the BRDF based on measurement or estimates of material properties. Torrance and Sparrow (1967) model the spectral BRDF of a material based on estimates of the complex index of refraction and the surface roughness. Priest and Germer (2002) and Wellems et al. (2000) indicate how the Torrance and Sparrow approach can be extended to describe a full polarimetric treatment of reflectance. A limitation of the Torrance and Sparrow approach and most other BRDF models is the need for detailed knowledge of the optical and structural properties of the surface. In many cases, these properties cannot be directly measured (e.g., the index of refraction of a soil) so they must be estimated by fitting modeled parameters to measurements. Since we turned to models to avoid endless measurements, this is not a particularly satisfying conclusion. However, in certain cases it may be acceptable if we can solve for the model parameters with only a limited number of measurements. Regrettably, neither modeled or measured data are widely available. Consequently, much of the analytical work in remote sensing (cf. Chap. 7) must rely on approaches that do not require a full knowledge of BRDF values.

It is often more convenient to describe directional reflectance in a unitless form. This is accomplished by introducing the *bidirectional reflectance factor* (r_{rF}). This is the ratio of the radiance reflected into a particular direction to the radiance that would be reflected into the same direction by a perfect Lambertian radiator illuminated in an identical fashion. The perfect Lambertian radiator is defined to have a total reflectivity of unity, and because it is Lambertian (cf. Chap. 3), it will have the same radiance in all directions. Feng et al. (1993) show how the reflectance factor (r_{rF}) is related to the bidirectional reflectance (r_{BRDF}) through a simple factor of π steradians (that magic π again), i.e.,

$$r_{\mathrm{BRDF}}[sr^{-1}] = \frac{r_{rF}}{\pi[sr]} \tag{4.26}$$

In general, we will omit the *rF* subscript and use reflectance factors for the remainder of our discussion.

In addition to the bidirectional reflectance terms, two other terms describing reflectivity are commonly used for remote sensing purposes. The first is the directional hemispheric reflectance. This is the ratio of the exitance from a target to the irradiance onto the target from a particular direction. Conversely, according to reciprocity, it is the ratio of the radiance into a particular direction to the radiance (uniform from all angles) irradiating a target.

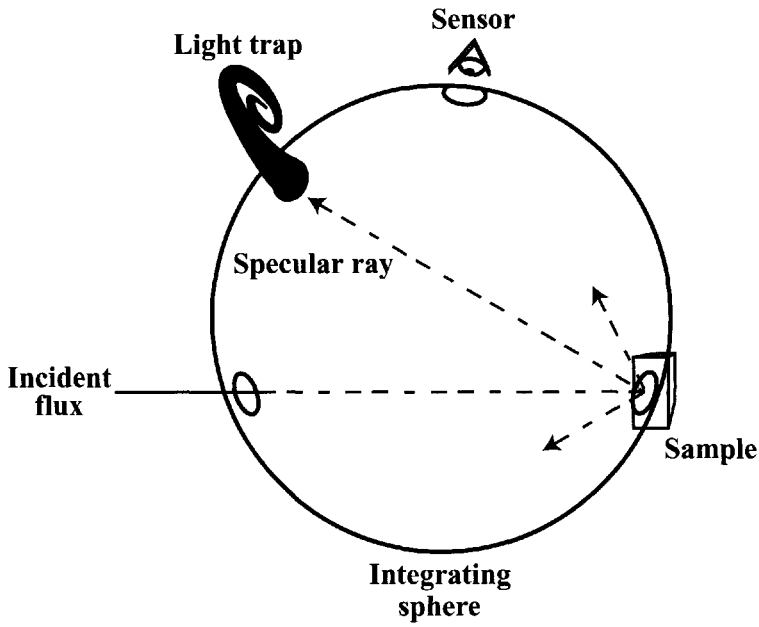


Figure 4.9 Schematic concept for measuring total and diffuse reflectance.

The final reflectance term of interest is the *diffuse reflectance*. This is best described in reference to the instrumentation commonly used to measure it (cf. Fig. 4.9). The instrument is set up so that all the flux from the sample is collected by an integrating sphere, except that which flows into a narrow cone about the specular direction. This is compared to a reading from a "perfect" Lambertian reflector measured in the same manner. The diffuse reflectance can then be defined as the hemispheric reflectance with the specular component removed. If the light trap is not used, we measure the total hemispheric reflectance. The ratio of the diffuse reflectance (r_d) to the total reflectance (r_{tot}) provides a measure of the diffuseness (d) of the sample, i.e.,

$$d = \frac{r_d}{r_{tot}} \tag{4.27}$$

A Lambertian reflector would be perfectly diffuse with a diffuseness of one. The more specular (less diffuse) a sample becomes, the lower its diffuseness, with a mirror having a value of zero. The specularity is often defined as one minus the diffuseness.

4.2.2 Reflected Solar Radiance

Using the reflectance terminology just introduced, we can develop an expression for the spectral radiance reflected toward the sensor associated with each of the irradiance terms in Eq. (4.23).

Using reflectance factor concepts, the spectral radiance $L_{sr\lambda}$ into the sensor direction (θ',ϕ_d') due to irradiance from the solar direction (σ',ϕ_s') can be expressed as

$$L_{sr\lambda}(\theta,\phi_d) = E'_{s\lambda}\tau_1(\lambda)\cos\sigma' \frac{r(\sigma',\phi'_s,\theta',\phi'_d,\lambda)}{\pi} [W m^{-2} sr^{-1} \mu m^{-1}] \tag{4.28}$$

The reflected downwelled radiance L_{dr} is more difficult to express in the general case, so we will first consider the more restrictive case where we can assume that the target is approximately diffuse (Lambertian). In this case, we draw on the derivation of Eq. (3.76) for radiance from a Lambertian reflector to yield

$$L_{dr\lambda} = E_{d\lambda} \frac{r_d(\lambda)}{\pi} \tag{4.29}$$

where r_d is the reflectance of the target if it is assumed diffuse (constant with angle). For a target with shape factor effects, this would be

$$L_{dr\lambda} = FE_{d\lambda} \frac{r_d(\lambda)}{\pi} \tag{4.30}$$

For cases where the reflectance cannot be assumed Lambertian, the reflected downwelled radiance can be expressed as

$$L_{dr\lambda}(\theta',\phi'_d) = \int_{\phi=0}^{2\pi} \int_{\sigma=0}^{\frac{\pi}{2}} L_{d\lambda}(\sigma,\phi)\cos\sigma \frac{r(\sigma',\phi',\theta',\phi'_d,\lambda)}{\pi} \sin\sigma d\sigma d\phi \tag{4.31}$$

For a target with shape factor effects, the limits of integration of Eq. (4.31) would be appropriately adjusted (cf. Fig. 4.5 and related discussion).

For the reflected background radiance (L_{br}), the procedure is similar to the reflected downwelled radiance if the target can be assumed to be approximately Lambertian, i.e.,

$$L_{br\lambda} = (1-F)L_{b\lambda avg} r_d(\lambda) \tag{4.32}$$

When the target is not Lambertian, we have

$$L_{br\lambda}(\theta',\phi'_d) = \int_{1-F} L_{b\lambda}(\sigma,\phi)\cos\sigma\, r_{BRDF}(\lambda) d\Omega$$

$$= \int_{1-F} \int L_{b\lambda}(\sigma,\phi)\cos\sigma \frac{r(\sigma',\phi',\theta',\phi'_d,\lambda)}{\pi} \sin\sigma d\sigma d\phi \tag{4.33}$$

where the limits of integration are over any background objects in the hemisphere above the target.

The reflected spectral radiance $L_{r\lambda}(\theta',\phi_d')$ toward the sensor can then be expressed as

$$L_{r\lambda}(\theta',\phi'_d) = E'_{s\lambda}\cos\sigma'\tau_1(\lambda)\frac{r(\lambda)}{\pi} + \int_F L_{d\lambda}(\sigma,\phi)\cos\sigma\frac{r(\lambda)}{\pi}d\Omega$$
$$+ \int_{1-F} L_{b\lambda}(\sigma,\phi)\cos\sigma\frac{r(\lambda)}{\pi}d\Omega \tag{4.34}$$

where the functional dependence of the reflection factors $r(\lambda)$ on angle has been suppressed for more compact presentation, i.e., $r(\lambda) = r_{rF}(\lambda)$.

When the downwelled radiance or the reflectance can be assumed to be reasonably diffuse, a good approximation to Eq. (4.34) can be achieved by

$$L_{r\lambda}(\theta',\phi'_d) = E'_{s\lambda}\cos\sigma'\tau_1(\lambda)\frac{r(\lambda)}{\pi}$$
$$+FE_{d\lambda}\frac{r_d(\lambda)}{\pi} + (1-F)L_{b\lambda\text{avg}}r_d(\lambda) \tag{4.35}$$

Note that we still use the reflectance factor in the solar irradiance term. This term will vary a great deal with sun-target-sensor geometry, unless the target is quite Lambertian (diffuse).

In Section 4.1, we encountered the reflected background radiance without fully defining it. This can be done now, recognizing that the radiance reflected from the background onto the target results from the same factors that cause reflected radiance from the target, i.e.,

$$L_{b\lambda}(\sigma,\phi) = E'_{s\lambda}\cos\sigma'_b\tau_1(\lambda)\frac{r_b(\lambda)}{\pi} + F_bE_{d\lambda}\frac{r_{bd}(\lambda)}{\pi}$$
$$+(1-F_b)L_{ba\lambda}r_{bd}(\lambda) \tag{4.36}$$

where r_b is the reflectance factor for the background, r_{bd} is the diffuse reflectance of the background, σ_b' is the angle from the normal to the background to the sun, F_b is the shape factor for the background, and L_{ba} is the mean radiance onto the background from the hemisphere above the background that is nonsky (i.e., the background's background). L_{ba} is generally assumed to be the radiance from a horizontal Lambertian reflector with the mean reflectance (albedo) of the earth for the scene being studied. In most cases, it is not necessary to solve Eq. (4.33) for each background point in the hemisphere above the target. In general, only negligible errors will be introduced by solving for the reflected background radiance at a central point in each background object and applying that value over the entire object. (In most cases, r_b can be set equal to r_{bd}, and F_b set equal to F, with no significant error introduced.)

4.3 SOLAR RADIANCE REACHING THE SENSOR

After all this effort, we've finally got the photon flux turned around and headed toward the sensor, ideally carrying the information we are interested in. However, the reflected radiance (L_r) headed toward the sensor is attenuated by absorption and scattering along the path. The transmission losses along the target-sensor path (τ_2)

can be described in the same fashion as the losses along the sun-target path (τ_1). The optical depth can be computed in a piecewise fashion and summed to yield the overall optical depth according to Eq. (3.109). This transmission loss results in an overall loss in signal level at the sensor. A competing effect is introduced through scattering by the atmosphere into the target-sensor path. This upwelled radiance will increase the overall radiance reaching the sensor in an additive fashion. The competition between these two interrelated effects results in radiance values at the sensor that may be larger or smaller than the surface-leaving radiance values. In this section, we will consider how these interactions affect the radiance reaching the sensor.

4.3.1 Solar Scattered Upwelled Radiance (Path Radiance)

The *upwelled radiance* is characterized in a manner analogous to the directional downwelled radiance calculations illustrated in Figure 4.4. In this case, we are concerned about the radiance from the sun scattered upward into the sensor's line of site along the sensor-target path. By analogy to the downwelled radiance case, this can be expressed as

$$L_{u\lambda}(\theta,\phi) = E'_{s\lambda}\int \tau_{L1}(\lambda)\tau_{L2}(\lambda)\beta_{sca}(\lambda,\theta_v)dr \tag{4.37}$$

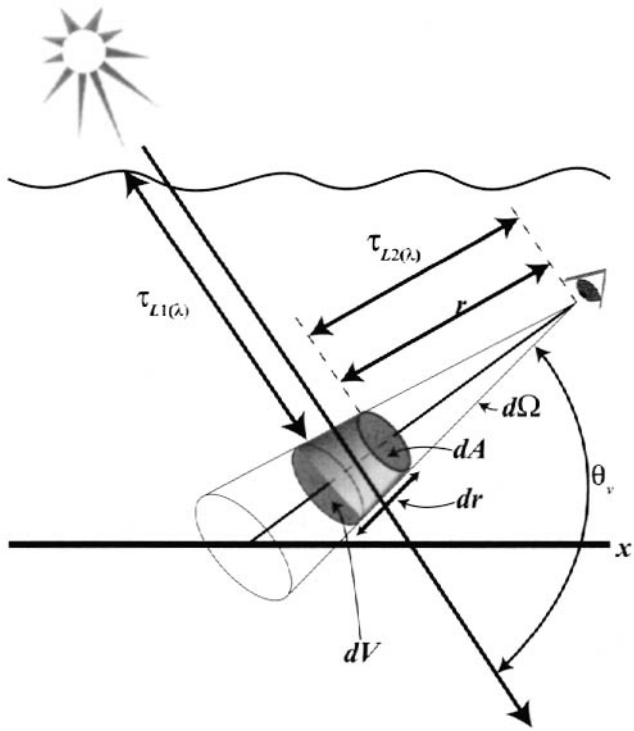


Figure 4.10 Procedure for calculation of upwelled radiance.

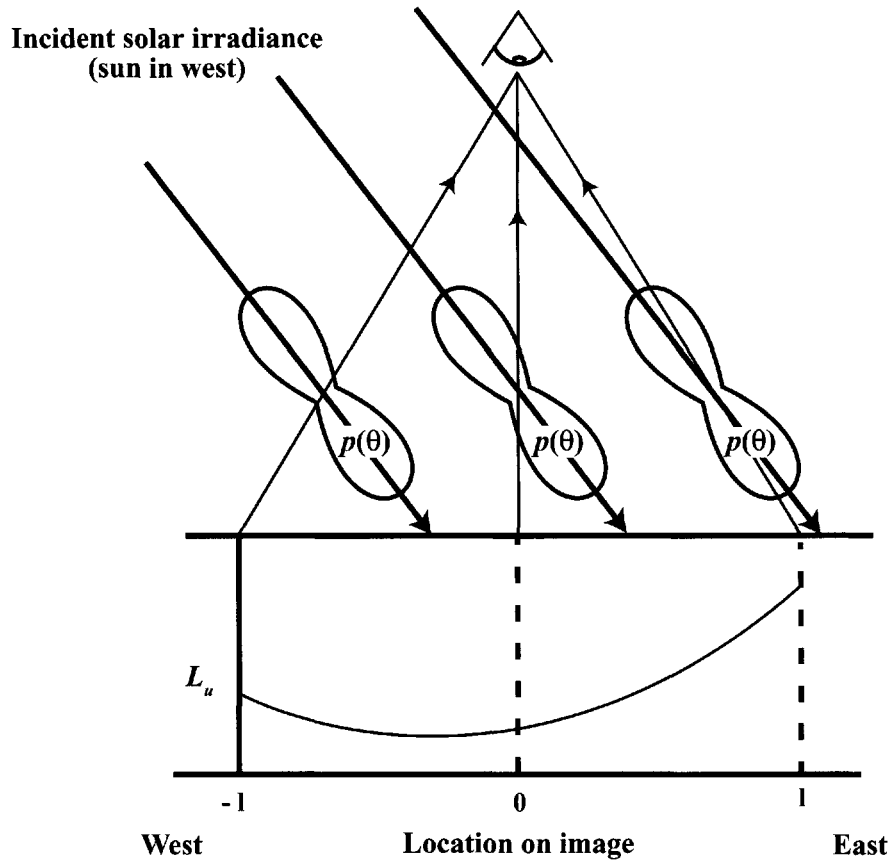


Figure 4.11 Variation in path radiance with view angle. A Rayleigh phase function is shown for reference; however, the upwelled radiance plot is for an actual atmosphere in the visible spectral region. This is the atmospheric upwelled radiance that would be seen in each line of data when flying a north-south line with a line scanner.

where the terms are interpreted as illustrated in Figure 4.10, the integral is along the ray from the sensor to the ground, and the angular scattering coefficient is considered to vary along that path.

The magnitude of the upwelled radiance is seen from Eq. (4.37) to result from a gain due to the angular scattering function and loss due to the transmission τ_{L2} from the scattering volume to the sensor. As a result, it can be difficult to predict whether a change in the atmospheric makeup will increase or decrease the path radiance. In general, if the number of scattering centers increases, path radiance will increase. However, these same scattering centers will scatter the flux scattered toward the sensor out of the target-sensor path, reducing path radiance. Most of the variation in scattering centers occurs in the lower atmosphere where scattering acts mostly as an energy source. The upper atmosphere, where extinction is important, is less variable. Thus, the net effect of increasing scattering centers in the atmosphere is usually an increase in path radiance.

Analysis of Eq. (4.37) also indicates that the path radiance must monotonically increase with path length along a particular line of sight. The variation in

upwelled radiance with view angle is more complicated, as illustrated in Figure 4.11. The scattering phase function plays a very important role in controlling how upwelled radiance varies with view angle. This is coupled to the increase in path length with view angle from nadir, which in general will increase the path radiance. The overall effect is that path radiance will tend to increase with view angle. However, the minimum path radiance will not necessarily occur at nadir.

4.3.2 Cumulative Solar Effects

If we combine the effects due to the transmission along the target-sensor path τ_2 with the upwelled radiance L_u, we have an overall expression for the radiance reaching the sensor due to solar photons of

$$\begin{aligned} L_{s\lambda}[Wm^{-2}sr^{-1}\mu m^{-1}] &= L_{r\lambda}\tau_2(\lambda) + L_{u\lambda}(\theta,\phi) \\ &= \left[E'_{s\lambda}\cos\sigma'\tau_1(\lambda)\frac{r(\lambda)}{\pi} + FE_{d\lambda}\frac{r_d(\lambda)}{\pi}\right. \\ &\left. +(1-F)L_{b\lambda avg}r_d(\lambda)\right]\tau_2(\lambda) + L_{u\lambda} \end{aligned} \tag{4.38}$$

where the angular dependencies are not explicitly indicated for clarity of presentation. At times we will find it convenient to express the transmission to the sensor as a simple function of view angle. For relatively thin atmospheres (predominantly single scattering) and restricted view angles ($\theta < 60$), the functional dependency of optical depth on view angle is nearly directly proportional to the increase in path length, so we can write the transmission to the sensor as

$$\tau_2(\lambda) \approx e^{-\delta'(\lambda)\sec\theta} \tag{4.39}$$

where δ' is the optical depth along the vertical path to the sensor altitude. Note that for a sensor in space $\delta = \delta'$. When this approximation is used, Eq. (4.38) can be represented as

$$\begin{aligned} L_{s\lambda} &= \left[E'_{s\lambda}\cos\sigma' e^{-\delta\sec\sigma}\frac{r(\lambda)}{\pi} + FE_{d\lambda}\frac{r_d(\lambda)}{\pi}\right. \\ &\left. +(1-F)L_{b\lambda avg}r_d(\lambda)\right]e^{-\delta'\sec\theta} + L_{u\lambda} \end{aligned} \tag{4.40}$$

For a fully exposed horizontal surface, this reduces to

$$L_{s\lambda} = \left[E'_{s\lambda}\cos\sigma'\tau_1(\lambda)\frac{r(\lambda)}{\pi} + E_{d\lambda}\frac{r_d(\lambda)}{\pi}\right]\tau_2(\lambda) + L_{u\lambda} \tag{4.41}$$

The relative impact of path radiance on radiance at the sensor is shown in Figure 4.12. This plot shows radiance leaving the ground versus radiance reaching the sensor for several atmospheres with different optical depths. It is clear that the combined effects of path radiance and transmission loss will increase the radiance from most targets, but will have a much more pronounced additive effect on low-reflectance targets. For highly reflecting targets, the transmission loss becomes more important and can significantly decrease the radiance at the sensor. Furthermore, if we compare the case of the less dense to more dense atmosphere, we see

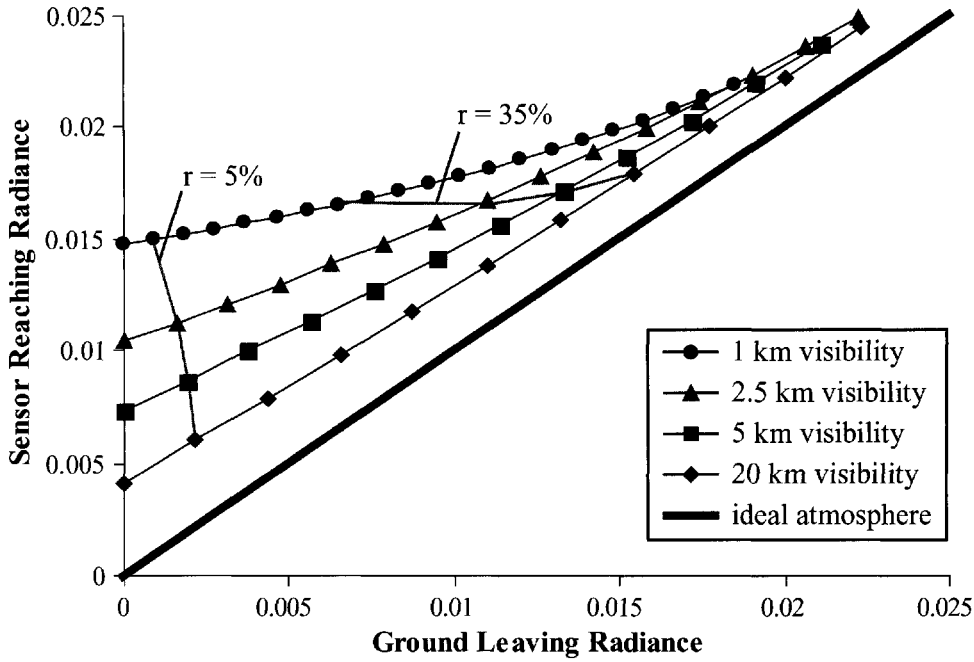


Figure 4.12 Effects of transmission and path radiance on radiance reaching the sensor. (Data derived from the MODTRAN atmospheric propagation model run in the visible spectral region.)

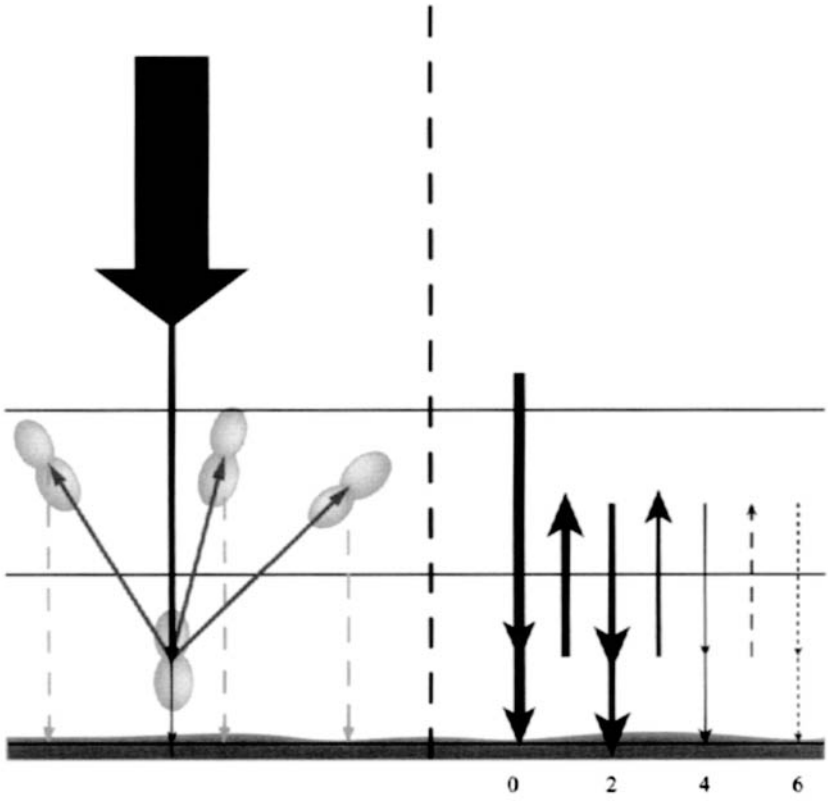
that for low-reflectance targets, the radiance increases with increasing turbidity (decreasing visibility and optical depth) in the atmosphere, as shown for the 5% reflector. Just the opposite occurs for high-reflectance objects (i.e., radiance will decrease with increasing turbidity). For this particular atmosphere, the 35% reflector shows both effects with its observed radiance, first decreasing with turbidity and then increasing. These effects are discussed in more detail and as a function of wavelength in Turner et al. (1975). They point out that while the scattering effects are reduced as you go from the blue toward the near infrared, the same basic effects occur.

4.3.3 Multiple Scattering and Nonlinearity Effects

Careful inspection of Figure 4.12 reveals a disconcerting curvature in the relationship between surface-leaving radiance and at sensor radiance that becomes apparent as the atmosphere becomes very turbid (i.e., very low visibility). Our analysis to this point would suggest that each line in Figure 4.12 should have an equation of the form

$$L_\lambda = \left(\frac{E'_{s\lambda} \tau_1(\lambda) \cos\theta}{\pi} + L_{d\lambda} \right) \tau_2(\lambda) r(\lambda) + L_{u\lambda} = L^{\downarrow}_{surf\lambda} \tau_2(\lambda) r(\lambda) + L_{u\lambda} \quad (4.42)$$

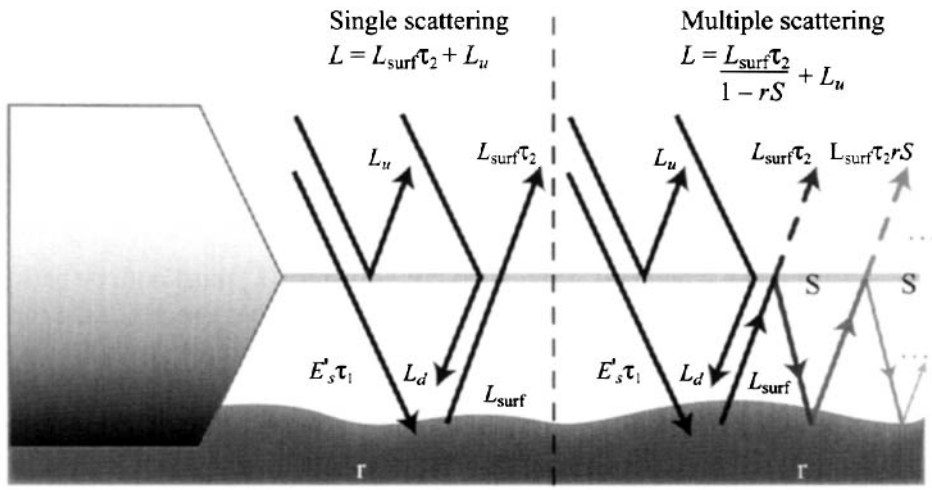
where $E_S'\tau_1(\lambda)\cos\theta/\pi + L_{d\lambda}$ is the radiance reaching the surface that when multiplied by the reflectance factor $[r(\lambda)]$ yields the surface-leaving radiance ($L_{\text{surf }\lambda}$). This suggests that the slope of each line should be the transmission (τ) and the intercept the upwelled radiance (L_u). Since neither τ or L_u should vary with reflectance, we would expect linear behavior. The answer comes when we recognize that our derivations to this point are rigorously valid only for the case of single scattering. The MODTRAN radiative transfer model used to generate Figure 4.12 incorporates multiple scattering effects. This means it attempts to account for the fact that photons propagating along a beam can be scattered out and then back into the beam. This is illustrated in Figure 4.13 for the simple case of a two-layered atmosphere. The beam traveling vertically down through the atmosphere will be attenuated by the first layer, and then some of it will be backscattered by the second layer back into the first. The amount in any direction is governed by the angular scattering coefficient (note the scattering phase function is shown in Figure 4.13 depicting



(a) Illustration of how angular scattering coefficients predict how flux will be back-scattered from the lower layer and then rescattered back into the beam direction by the upper layers.

(b) Illustration of how the net scattering illustrated in figure (a) can be combined into a twice-scattered beam in the original direction, a small fraction of which will in turn be twice scattered back into the original beam direction and so on.

Figure 4.13 Multiple scattering concepts.



$$L_{surf} = \left(\frac{E'_s\tau_1\cos\sigma}{\pi} + L_d\right)r$$

Figure 4.14 Incorporation of multiple scattering effects in the governing equation for radiation propagation.

the relative amount of flux scattered in each direction). The flux scattered into the upper layer is then in turn subject to be scattered back into the original beam direction as governed by the angular scattering coefficient of the upper layer. Figure 4.14 illustrates how we can conceptually aggregate all of the flux scattered back into the original direction into a new beam that will be subject to exactly the same scattering process as the original beam. This process will continue ad infinitum with significantly reduced magnitude after each iteration. Clearly, for an atmosphere conceptually made up of many layers, we will have to consider the iteration between each pair of layers. The MODTRAN model used to generate the data in Figure 4.12 incorporates a numerical approximation of this multiple scattering process for a many-layered atmosphere. For our purpose at the moment, we will use an approximation to the multiple scattering problem by treating the world as though it is made up of two entities: an opaque reflecting surface beneath a partially transmissive surface that reflects both upward and downward. Here we have conceptually aggregated the entire sky dome into a single layer with the effective properties of the sky dome (cf. Fig. 4.14). With this approach, we see that the single scattered radiance can be treated essentially as we have to this point, but that we have to account for the multiple scattering from the sky dome. To accomplish this, we introduce the *spherical scattering albedo,* $S(\lambda)$, which can be thought of as the effective diffuse reflectivity of the sky to upwelling radiation. Then some of the radiance from the earth (captured in the term $L_{surf\lambda}$) will be reflected by the sky $[L_{surf\lambda}S(\lambda)]$ back to the Earth to be reflected once more back toward the sensor $[L_{surf\lambda}r(\lambda)S(\lambda)]$. This process will repeat through an infinite series. Note that at each reflection from the sky, some fraction is transmitted, contributing to our

sensed radiance. The radiance reaching the sensor due to multiple scattering would then be expressed as

$$L_\lambda = L_{surf\lambda}\tau_2(\lambda) + L_{surf\lambda}\tau_2(\lambda)r(\lambda)S(\lambda) + L_{surf\lambda}\tau(\lambda)\left(r(\lambda)S(\lambda)\right)^2 + \ldots + L_{u\lambda}$$
$$= \frac{\left(\frac{E_s'\tau_1(\lambda)\cos\sigma}{\pi} + L_d(\lambda)\right)r(\lambda)\tau_2(\lambda)}{1 - r(\lambda)S(\lambda)} + L_{u\lambda} \tag{4.43}$$

where we have taken advantage of the fact that $r(\lambda)$ and $S(\lambda)$ must both be between 0 and 1 so their product must also be between 0 and 1 and that a series of the form $1 + c + c^2 + c^3 \ldots$ can be reexpressed as $1/(1-c)$ under these circumstances. Eq. (4.43) can reasonably approximate the nonlinear results shown in Figure 4.12. However, we need to recognize that in cases where multiple scattering is important, the values of $L_{u\lambda}$ and $L_{d\lambda}$ in Eq. (4.43) will be larger than in Eq. (4.42) due to multiple scattering between the layers we so conveniently aggregated in Figure 4.14. Generation of values for the spherical scattering albedo [$S(\lambda)$] generally involves the use of radiative transport codes, but it is only dependent on the atmospheric transmission and angular scattering terms already introduced. We will look at these radiative transfer codes in more detail in Chapter 7.

Inspection of Eq. 4.43 indicates that the relationship between radiance and reflectance should remain linear at the 1% level or better as long as the product of $r(\lambda)$ and $S(\lambda)$ is <0.02. In practice, for most reflectors and wavelengths, this criteria is met, so we will often use the linear relationship. However, for bright targets [large $r(\lambda)$] and high scattering condition (e.g., low atmospheric visibility and short wavelengths), we will need to incorporate the nonlinear effects captured in Eq. (4.43).

4.4 THERMAL RADIANCE REACHING THE SENSOR

In Chapter 3, we identified a number of energy paths that self-emitted photons could traverse to produce radiance at the sensor (cf. Fig. 3.3). In this section, we will develop radiometric expressions to describe the radiance associated with those paths and combine them with the expression developed in the previous section for the solar terms.

4.4.1 Self-Emission

The energy path that is most often of interest when sensing in the thermal infrared is associated with the emission from an object due to its temperature (path *D* in Figure 3.3). This is the only path that carries information about the object's temperature. The radiance headed toward the sensor due to the target's temperature will be a function of the Planck equation modified by the wavelength-dependent emissivity of the target. This radiance will be attenuated by the transmission along the target sensor path. Combining these effects with the solar terms yields

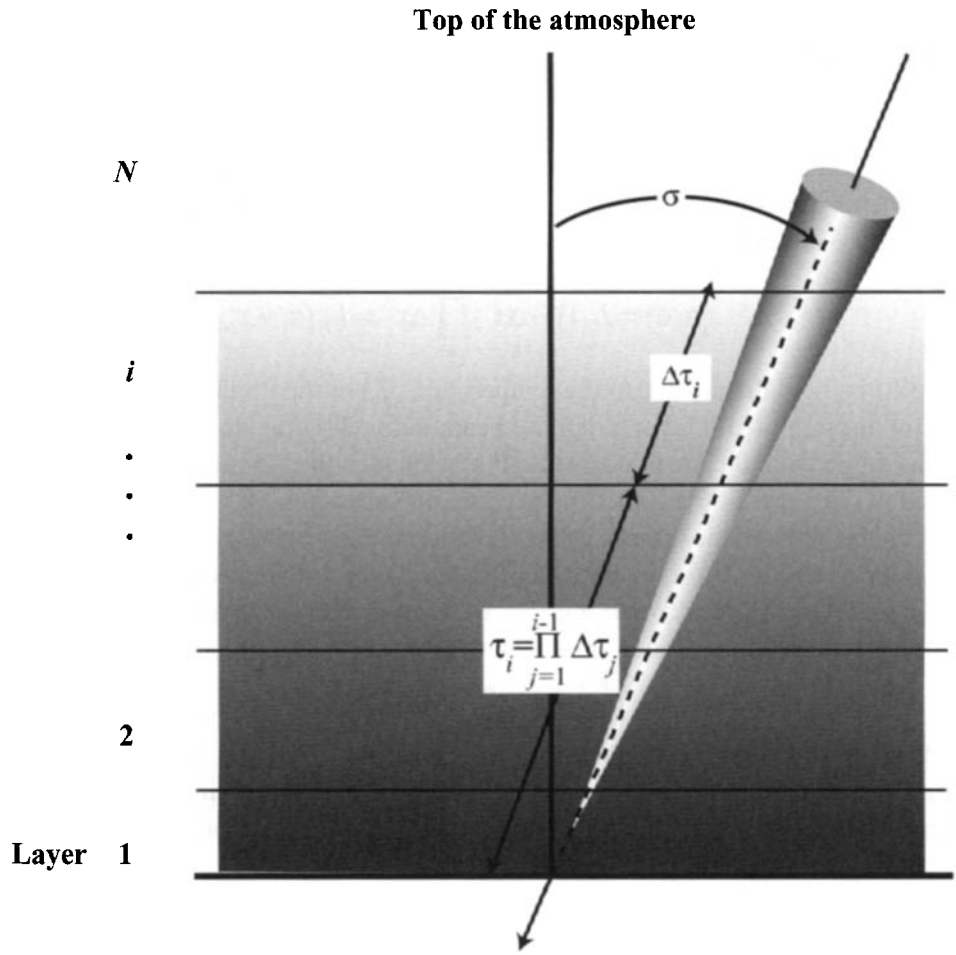


Figure 4.15 Downwelled radiance due to self-emission.

$$L_\lambda = E'_{s\lambda}\cos\sigma'\frac{r(\lambda)}{\pi}\tau_2(\lambda) + E_{d\lambda}\frac{r_d(\lambda)}{\pi}\tau_2(\lambda) + \varepsilon(\lambda)L_{T\lambda}\tau_2(\lambda) + L_{u\lambda} \tag{4.44}$$

where $\varepsilon(\lambda)$ is the wavelength-dependent emissivity—for non-Lambertian objects, ε will also be a function of view angle (θ') and of azimuthal angle (ϕ_d') for azimuthally varying surfaces (e.g., corrugated surfaces) and $L_{T\lambda}$ is the spectral radiance [$Wm^{-2}sr^{-1}\mu m^{-1}$] for a blackbody at temperature T as described by the Planck equation, Eq. (3.78).

4.4.2 Thermal Emission from the Sky and Background Reflected to the Sensor

Recall from Kirchoff's law that an opaque object with emissivity ε will have reflectivity $r = 1 - \varepsilon$. We need to consider reflection by the target of irradiance from the surround due to the temperature of the surround. The atmosphere has some finite temperature and will act as a source of energy to be reflected to the sensor

(type E photons in Fig. 3.3). In order to determine how much radiance is reflected toward the sensor from the sky, we need to determine the contribution from (i.e., the irradiance from) each location in the sky. This is most readily conceptualized by considering an atmosphere made up of many homogeneous layers, as illustrated in Figure 4.15. If we let there be N layers between the ground and the top of the atmosphere, then the contribution to the downwelled radiance from the ith layer from the σ,ϕ direction can be expressed as:

$$\Delta L_{d\varepsilon}(\sigma,\phi) = L_{Ti}(1-\Delta\tau_i)\prod_{j=1}^{i-1}\Delta\tau_j = L_{Ti}(\tau_i - \tau_{i+1}) \tag{4.45}$$

where L_{Ti} is the radiance due to the temperature (T_i) of the ith layer, $\Delta\tau_i$ is the transmission through the ith layer along the beam defined by the direction angles (σ,ϕ), σ is the angle from the Earth normal to the beam, ϕ is the azimuthal angle about the normal in the plane of the Earth, the product on j accounts for the transmission loss along the beam from the bottom of the ith layer to the ground, and τ_i is the transmission along the beam from the bottom of the ith layer to the ground.

To make intuitive sense out of Eq. (4.45), we need to recall that conservation of energy requires that for a homogeneous layer i

$$\Delta\tau_i + \Delta r_i + \Delta\alpha_i = 1 \tag{4.46}$$

where Δr_i and $\Delta\alpha_i$ are the reflectance and absorptance of the ith layer along the beam. Since scattering is negligible at most wavelengths of interest for thermal infrared sensing (rigorously speaking, $\Delta\tau_i$ should be defined as transmission loss due solely to absorption), Δr_i is zero, leaving

$$\Delta\tau_i + \Delta\alpha_i = 1 \tag{4.47}$$

or, from Kirchoff's rule,

$$\Delta\tau_i + \Delta\varepsilon_i = 1 \tag{4.48}$$

So, we see that the effective emissivity of the layer can be expressed as

$$\Delta\varepsilon_i = (1-\Delta\tau_i) \tag{4.49}$$

Thus, we can interpret the contribution to radiance from the ith layer in Eq. (4.45) as the effective emissivity of the layer ($1 - \Delta\tau_i$) times the radiance of the layer (L_{Ti}) times the transmission from the bottom of the layer to the ground.

The difference in transmission values on the far right-hand side of Eq. (4.45) is a simplification commonly found in the LOWTRAN/MODTRAN literature [cf. Kneizys et al. (1988)]. It can be derived by recognizing that

$$\prod_{j=1}^{i-1}\Delta\tau_j = \tau_i \tag{4.50}$$

and

$$\tau_{i+1} = \tau_i\Delta\tau_i \tag{4.51}$$

so that

$$(\tau_i - \tau_{i+1}) = \prod_{j=1}^{i-1} \Delta\tau_j - \left(\prod_{j=1}^{i-1} \Delta\tau_j \right) \Delta\tau_i$$
$$= (1 - \Delta\tau_i) \prod_{j=1}^{i-1} \Delta\tau_j \tag{4.52}$$

The total contribution to downwelled radiance from the direction defined by (σ,ϕ) can be expressed as

$$L_{d\varepsilon}(\sigma,\phi) = \sum_{i=1}^{N} L_{Ti}(\tau_i - \tau_{i+1}) \tag{4.53}$$

where the sum is over all the layers in the atmosphere.

The element of irradiance from the σ,ϕ direction $dE_d(\sigma,\phi)$ can be expressed from the definition of radiance as

$$dE_{d\varepsilon}(\sigma,\phi) = L_{d\varepsilon}(\sigma,\phi)\cos\sigma \, d\Omega \tag{4.54}$$

with the total downwelled irradiance expressed as

$$E_{d\varepsilon} = \int_{2\pi} L_{d\varepsilon}(\sigma,\phi)\cos\sigma \, d\Omega \tag{4.55}$$

For a Lambertian reflector, the spectral radiance reflected to the sensor from downwelled emission from the sky would then be

$$L_{r\varepsilon\lambda} = E_{d\varepsilon\lambda} \frac{r_d(\lambda)}{\pi} \tag{4.56}$$

where the subscript ε denotes dependence on self-emission, and the spectral dependency, which was neglected for readability, has been reintroduced. In the case of a non-Lambertian target, the element of reflected radiance can be expressed as

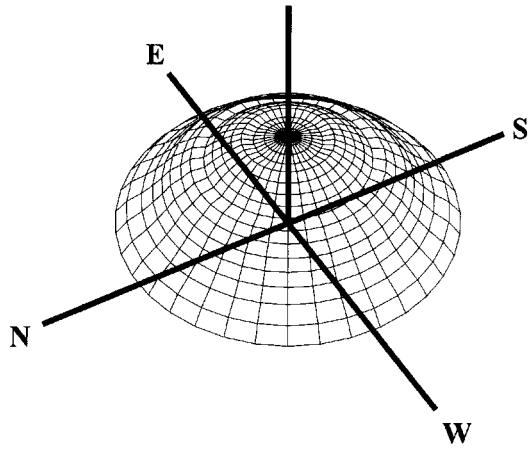
$$dL_{r\varepsilon\lambda} = dE_{d\varepsilon\lambda}(\sigma,\phi)\frac{r(\lambda)}{\pi} \tag{4.57}$$

or, on substituting in Eq. (4.54) and integrating, we have

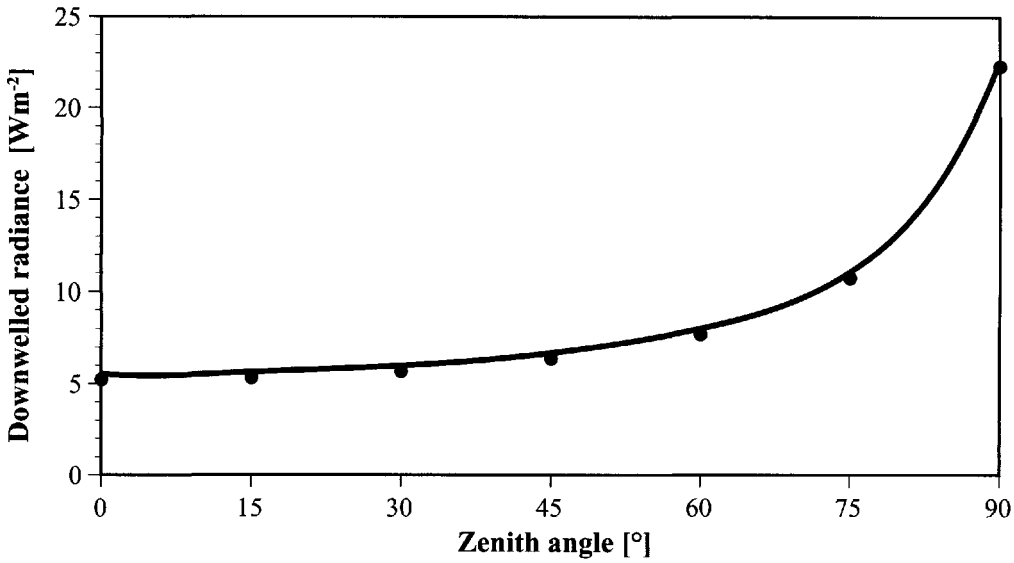
$$L_{r\varepsilon\lambda} = \int_{2\pi} L_{d\varepsilon\lambda}(\sigma,\phi)\cos\sigma \frac{r(\lambda)}{\pi} d\Omega \tag{4.58}$$

If we now introduce the case of nonhorizontal or obscured surfaces, the analysis proceeds in the same fashion as used for scattered downwelled radiance with the use of the shape factor. The radiance reflected toward the sensor due to self-emitted downwelled sky radiance and background self-emission can be expressed as

$$L_{r\varepsilon\lambda}(\theta',\phi_d') = \int_F L_{d\varepsilon\lambda}(\sigma,\phi)\cos\sigma \frac{r(\sigma,\phi,\theta',\phi_d',\lambda)}{\pi} \sin\sigma \, d\sigma \, d\phi$$
$$+ \int_{1-F} L_{Tb\lambda}(\sigma,\phi)\cos\sigma \frac{r(\sigma,\phi,\theta',\phi_d',\lambda)}{\pi} \sin\sigma \, d\sigma \, d\phi \tag{4.59}$$



(a) Polar plot of downwelled radiance.



(b) Plot of downwelled radiance as a function of view angle θ for a fixed azimuth angle ϕ_d.

Figure 4.16 Angular variation in downwelled radiance in the 8-14 μm region due to self-emission by the sky.

where $L_{Tb}(\sigma,\phi)$ is the radiance onto the target due to the temperature of the background (T_b) coming from the σ,ϕ direction, and we have explicitly expressed the dependence of the reflectance factor $r(\lambda)$ on the incident and exitant angles. If the BRDF is slowly varying with angle, or the downwelled radiance and background radiance do not vary greatly, Eq. (4.59) can be expressed to a good approximation as

$$L_{r\varepsilon\lambda} \cong FE_{d\varepsilon\lambda}\frac{r_d(\lambda)}{\pi} + (1-F)L_{b\varepsilon\lambda}r_d(\lambda) \tag{4.60}$$

where $L_{b\varepsilon\lambda}$ is the mean spectral radiance from the background due to self-emission. In most cases, background objects will all have similar temperatures that will be near earth's ambient temperature. The downwelled radiance varies with angle (cf. Fig. 4.16); however, it is a gradual function, so Eq. (4.60) is a good approximation for any target that exhibits reasonably diffuse behavior. Note in Figure 4.16(b) that under clear-sky conditions, the downwelled radiance due to self-emission is always greater near the horizon due to the longer path through a warmer, denser atmosphere.

If we multiply Eq. (4.60) by the transmittance along the target-sensor path and combine it with Eq. (4.40) and the target self-emission term, we have a new expression for the radiance reaching the sensor:

$$\begin{aligned} L_\lambda = {} & E'_{s\lambda}\cos\sigma'\frac{r(\lambda)}{\pi}\tau_1(\lambda)\tau_2(\lambda)+\varepsilon(\lambda)L_{T\lambda}\tau_2(\lambda) \\ & +F\left[E_{ds\lambda}+E_{d\varepsilon\lambda}\right]\frac{r_d(\lambda)}{\pi}\tau_2(\lambda) \\ & +(1-F)[L_{bs\lambda}+L_{b\varepsilon\lambda}]r_d(\lambda)\tau_2(\lambda)+L_{us\lambda} \end{aligned} \tag{4.61}$$

where we have introduced the subscript s on the solar reflected background radiance ($L_{bs\lambda} = L_{b\lambda\text{avg}}$), the solar scattered downwelled irradiance ($E_{ds\lambda} = E_{d\lambda}$), and the upwelled radiance $L_{us\lambda} = L_{u\lambda}$. Explicitly expressing the dependence on solar photons differentiates the solar terms (subscript s) from the self-emissive terms (subscript ε). Consideration of the symmetry of Eq. (4.61) or of the energy paths in Figures 3.1 and 3.3 points us to the one remaining self-emissive term yet to be considered.

4.4.3 Self-Emitted Component of Upwelled Radiance

Equation (4.61) lacks a contribution due to the self-emitted radiance from the atmosphere along the line of site. This upwelled radiance term $L_{u\varepsilon}$ can be expressed in the same fashion as the angular contribution to the self-emitted downwelled radiance $L_{d\varepsilon}(\sigma,\phi)$, i.e.,

$$\begin{aligned} L_{u\varepsilon\lambda}(\theta,\phi) &= \sum_{i=1}^{N} L_{Ti\lambda}[\tau_i(\lambda)-\tau_{i+1}(\lambda)] \\ &= \sum_{i=1}^{N} L_{Ti\lambda}[1-\Delta\tau_i(\lambda)]\prod_{j=1}^{i-1}\Delta\tau_j(\lambda) \end{aligned} \tag{4.62}$$

where we have redefined the indexing scheme for the atmospheric layers to begin at the sensor and end with the layer just above the ground. Thus, τ_i is the transmission along the sensor-target path from the sensor to the top of the ith layer.

Inspection of Eq. (4.62) indicates that for an atmosphere composed of uniform homogeneous layers, the upwelled radiance due to self-emission $L_{u\varepsilon}$ should be independent of the azimuthal view angle and symmetric about the view angle from the nadir. As view angle increases from the nadir, two competing processes take place. The increased path through each layer increases the number of emitters

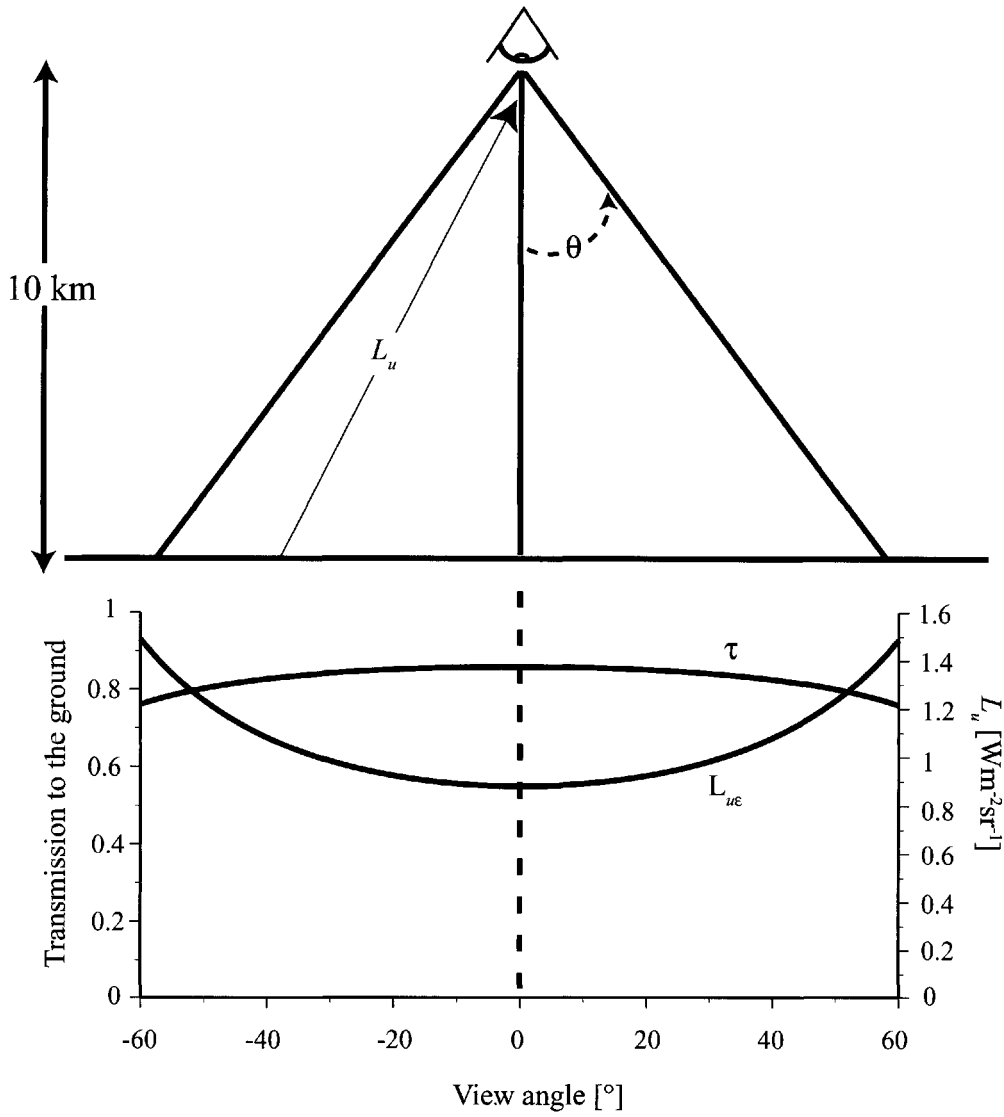


Figure 4.17 Variation in transmission and upwelled radiance with view angle due to self-emission in the 10–12 μm region.

in the beam, causing the effective emissivity of the layer $(1 - \Delta\tau_i)$ to increase. At the same time, the transmission along the path decreases due to the increased path length. In general, for reasonably clear atmospheres, the additive effects win out, and $L_{u\varepsilon}$ increases with view angle (cf. Fig. 4.17).

Adding the upwelled radiance term due to self-emission $L_{u\varepsilon}$ into our expression for the radiance at the sensor [Eq. (4.61)], we have a final expression (what my students over the years have taken to calling the *big equation*) for the spectral radiance reaching the sensor:

$$L_\lambda = [E'_{s\lambda} \cos\sigma' \tau_1(\lambda) \frac{r(\lambda)}{\pi} + \varepsilon(\lambda) L_{T\lambda} + F(E_{ds\lambda} + E_{d\varepsilon\lambda}) \frac{r_d(\lambda)}{\pi} + (1-F)(L_{bs\lambda} + L_{b\varepsilon\lambda}) r_d(\lambda)] \tau_2(\lambda) + L_{us\lambda} + L_{u\varepsilon\lambda} \quad (4.63)$$

$$L = L_A + L_D + L_B + L_E + L_G + L_H + L_C + L_F$$

$$L_\lambda = [\, E_{s\lambda} \cos(\sigma)\, \tau_1(\lambda) \frac{r(\lambda)}{\pi} + \varepsilon(\lambda) L_{T\lambda} + F\,(E_{ds\lambda} + E_{d\varepsilon\lambda}) \frac{r_d(\lambda)}{\pi}$$

$$+ (1 - F)(L_{bs\lambda} + L_{b\varepsilon\lambda}) r_d(\lambda)\,]\, \tau_2(\lambda) + L_{us\lambda} + L_{u\varepsilon\lambda}$$

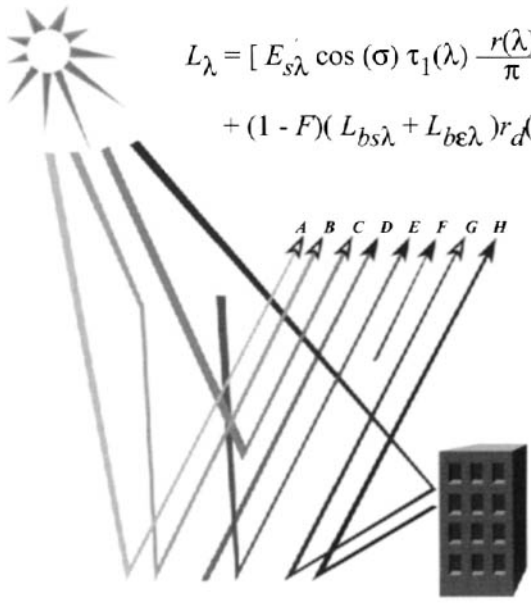


Figure 4.18 Relationship between terms in "the *big equation*" and energy paths associated with the photon flux onto the sensor.

The relationship between each of these terms and the energy path diagrams in Figures 3.1 and 3.3 is illustrated in Figure 4.18. For clarity, we will generally use a form of Eq. (4.63) in the remainder of the book. However, if a more rigorous treatment employing full BRDF data is required, the shape factor simplifications can be replaced with the integrals over the hemisphere [cf. Eqs. (4.34) and (4.59)].

4.5 INCORPORATION OF SENSOR SPECTRAL RESPONSE

The expression we have developed for spectral radiance reaching the sensor Eq. (4.63) must be cascaded with the sensor spectral response function to determine how much of the radiance is "sensed" by the system. Using the concept of effective radiance from Eq. (3.18), we can describe the effective radiance reaching the sensor as

$$L_{\text{eff}} = \int_0^\infty L_\lambda R'(\lambda) d\lambda \tag{4.64}$$

where L_λ is defined by Eq. (4.63), and $R'(\lambda)$ is the normalized spectral response function of the sensor.

In many operational situations, it is desirable to use average or effective bandpass values to simplify the calculation or representation of terms in equations such as Eq. (4.63). This can be done by approximating each term in the following fashion:

$$L_{\Delta\lambda} = \int_0^\infty \varepsilon(\lambda) L_{T\lambda} \tau_2(\lambda) R'(\lambda) d\lambda \approx \varepsilon_{\text{avg}} L_{T\text{eff}} \tau_{2\text{avg}} = \varepsilon L_T \tau_2 \tag{4.65}$$

where $L_{\Delta\lambda}$ is the effective radiance due to the target self-emission reaching the sensor, ε_{avg} is the mean emissivity across the bandpass, $\tau_{2\text{avg}}$ is the mean transmission across the bandpass, and $L_{T\text{eff}}$ is the effective radiance (due to the target's temperature) over the sensor's spectral response. For convenience, the subscripts avg and eff will be assumed, except where needed for clarity. From freshman calculus, we know that Eq. (4.65) is not true in general. It is, however, often a reasonable approximation if the wavelength-dependent terms are continuous over the bandpass and approximately constant. Figure 4.19 shows a case in the LWIR where this approximation might hold reasonably well. However, the MWIR case shown is an example where the approximation will break down. Recognizing these limitations, we will proceed to use the approximation method for clarity of presentation. The reader is cautioned that when there is substantial spectral character to the data, the wavelength-by-wavelength solution of Eq. (4.64) is required for accuracy. Using the simplified form, the effective radiance in the bandpass reaching the sensor can be expressed as:

$$\begin{aligned} L = \{ & E_s' \tau_1 \cos\sigma' \frac{r}{\pi} + \varepsilon L_T + [F(L_{ds} + L_{d\varepsilon}) \\ & + (1-F)(L_{bs} + L_{b\varepsilon})] r_d \} \tau_2 + L_{us} + L_{u\varepsilon} \end{aligned} \tag{4.66}$$

Even in this simplified form, the number of parameters needed to solve Eq. (4.66) is somewhat daunting. In the next section, we will evaluate what conditions might allow us to further simplify "the big equation."

4.6 SIMPLIFICATION OF THE BIG EQUATION AND RELATIVE MAGNITUDE ASSESSMENT

We've now come to the stage in our analysis I always loved as a physics student: where the professor drew a big line through expressions setting them equal to zero so that we no longer needed to think about them. In this section, we want to determine what the governing equation, Eq. (4.66), will look like in different portions of the spectrum and for different targets. In terms of the image chain approach, we want to determine what the governing equation for the radiometric portion of the image should be for each atmospheric transmission window.

4.6.1 Simplification

When we are trying to measure reflectance, we will seldom be interested in making measurements to better than 0.1 reflectance units (e.g., from 10.6% to 10.7% reflectance). Therefore, terms in Eq. (4.66) that result in changes significantly less than one part in several hundred will be considered negligible. Similarly, when try-

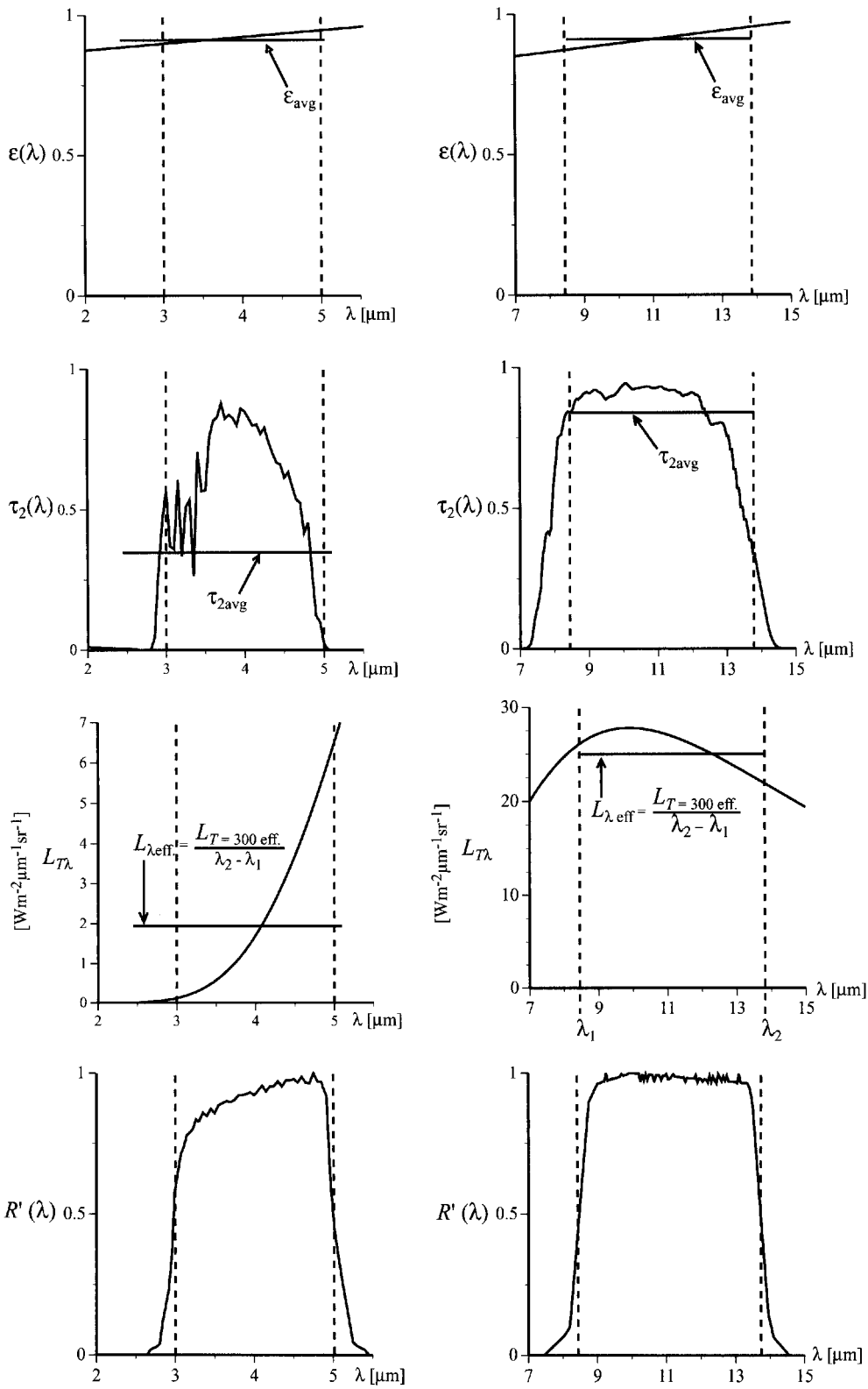


Figure 4.19 Spectral character of the parameters impacting the radiance reaching the sensor in the MWIR (left) and LWIR (right) spectral regions.

ing to measure temperature, we will seldom be interested in measuring temperature to better than 0.1 K. At 300 K, a change of 0.1 K amounts to a change in radiance of about one part in 800 in the LWIR and one part in 300 in the MWIR. Clearly, the required precision will change with both the target and the application. A small change will be much more important when measuring low reflectance (e.g., water in the VIS-NIR) or low temperature (e.g., nighttime winter scenes). For any conditions that deviate substantially from those identified here, a similar analysis should be performed employing the specific parameters and precisions required. To determine the importance of the terms in the "big equation," we can perform a simple relative magnitude computation for a typical case. This involves estimating the magnitude of each parameter in Eq. (4.63), computing the value of each term by cascading in the spectral response of the sensor, Eq. (4.64), and summing to yield the total radiance at the sensor. If each term in turn is then set to zero and the sum of all the other terms compared to the total radiance, we can determine the relative

Table 4.1 Sensitivity Analysis of Terms in the Big Equation.
Each entry in the matrix represents the magnitude of the radiance. The entries for apparent reflectance and temperature represent the apparent reflectance or temperature if the term were set equal to zero.

L_A $E'_s \cos\sigma \frac{r}{\pi}\tau_1\tau_2$	L_D $\varepsilon L_1\tau_2$	L_B $FL_{ds}r\tau_2$	L_E $FL_{d\varepsilon}r\tau_2$	L_G $(1-F)L_{Bs}r\tau_2$	L_H $(1-F)L_{b\varepsilon}r\tau_2$	L_F $L_{u\varepsilon}$	L_C L_{us}	L_{total}
Integrated Radiance [$Wcm^{-2}sr^{-1}$], $\Delta\lambda = 0.4 - 0.7$ μm								
$\varepsilon = 0.8$	2.28E-03	9.10E-28	6.36E-04	6.92E-31	3.63E-05	4.55E-29	1.14E-29	1.75E-03	4.70E-03
$\varepsilon = 0.9$	1.14E-03	1.02E-27	3.18E-04	3.46E-31	1.81E-05	2.56E-29	1.14E-29	1.75E-03	3.22E-03
$\varepsilon = 0.986$	1.60E-04	1.12E-27	4.45E-05	4.48E-32	2.54E-06	3.93E-30	1.14E-29	1.75E-03	1.95E-03
Integrated Radiance [$Wcm^{-2}sr^{-1}$], $\Delta\lambda = 3 - 5$ μm								
$\varepsilon = 0.8$	9.51E-05	1.28E-04	1.15E-06	4.94E-06	4.98E-07	6.61E-06	3.93E-05	6.55E-07	2.77E-04
$\varepsilon = 0.9$	4.75E-05	1.44E-04	5.73E-07	2.47E-06	2.49E-07	3.71E-06	3.93E-05	6.55E-07	2.39E-04
$\varepsilon = 0.986$	6.66E-06	1.58E-04	8.03E-08	3.46E-07	3.48E-08	5.67E-07	3.93E-05	6.55E-07	2.06E-04
Integrated Radiance [$Wcm^{-2}sr^{-1}$], $\Delta\lambda = 8 - 14$ μm								
$\varepsilon = 0.8$	6.92E-06	5.19E-03	4.11E-08	1.86E-04	3.45E-08	2.67E-04	1.56E-03	2.39E-08	7.21E-03
$\varepsilon = 0.9$	3.46E-06	5.84E-03	2.06E-08	9.29E-05	1.73E-08	1.50E-04	1.56E-03	2.39E-08	7.65E-03
$\varepsilon = 0.986$	4.85E-07	6.39E-03	2.88E-09	1.30E-05	2.42E-09	2.29E-05	1.56E-03	2.39E-08	7.99E-03
Apparent Reflectance, $\Delta\lambda = 0.4 - 0.7$ μm								
$\varepsilon = 0.8$	0.0760	0.1478	0.1278	0.1478	0.1476	0.1478	0.1478	0.0929	0.1478
$\varepsilon = 0.9$	0.0655	0.1013	0.0913	0.1013	0.1008	0.1013	0.1013	0.0465	0.1013
$\varepsilon = 0.986$	0.0564	0.0614	0.0600	0.0614	0.0613	0.0614	0.0614	0.0065	0.0614
Apparent Temperature (Kelvin), $\Delta\lambda = 3 - 5$ μm								
$\varepsilon = 0.8$	282.12	277.62	291.63	291.31	291.68	291.16	288.19	291.67	291.72
$\varepsilon = 0.9$	283.30	267.86	288.29	288.11	288.32	287.99	284.26	288.28	288.35
$\varepsilon = 0.986$	284.21	253.67	284.94	284.91	284.95	284.89	280.19	284.88	284.95
Apparent Temperature (Kelvin), $\Delta\lambda = 8 - 14$ μm								
$\varepsilon = 0.8$	280.58	222.35	280.64	279.13	280.64	278.46	267.15	280.64	280.64
$\varepsilon = 0.9$	284.03	218.31	284.06	283.34	284.06	282.90	271.07	284.06	284.06
$\varepsilon = 0.986$	286.64	214.09	286.64	286.55	286.64	286.48	274.12	286.64	286.64

importance of each term. If zeroing the term has negligible impact based on our required precision, we can neglect that term. Since this is not a robust assessment, we should include any terms that are close to our rejection threshold and perform a more robust test if our application varies substantially from the test presented here.

The example case we considered is presented in Table 4.1. The parameters were generated by manipulating the output from the MODTRAN radiation propagation model using the equations developed in this chapter [cf. the discussion of MODTRAN in Sec. 7.3.3 or Salvaggio et al. (1993)]. By comparing the total radiance (L) in Table 4.1 to the radiance when each term is subtracted, we can determine the relative importance of the term in each spectral interval. To simplify this assessment, we have introduced the apparent reflectance and apparent temperature concepts. These terms are often used to provide a more intuitive interpretation of radiance values in a spectral bandpass. The apparent reflectance (also referred to as the *reflectance at the top of the atmosphere*) is the reflectance a Lambertian reflector, located directly in front of the sensor, would have to have, in the absence of any atmosphere, to produce the observed radiance. It would be solved for according to

$$r_{ap} = L_{eff}\left(\frac{E_s' \cos\sigma'}{\pi}\right)^{-1} \tag{4.67}$$

where L_{eff} is the effective radiance that we wish to express in terms of apparent reflectance.

The *apparent temperature* is the temperature a blackbody would have to have to produce the effective radiance observed at the sensor; i.e., it is the temperature T_{ap} that would make the following expression true:

$$L_{eff} = \int_0^\infty L_{T_{ap}\lambda} R'(\lambda) d\lambda \tag{4.68}$$

The quantitative data in Table 4.1 for the case studied verify the graphical data of Figure 3.23. In the VIS-NIR region, the solar energy is so many orders of magnitude higher than the self-emitted energy that all the thermal energy paths (D, E, F, and H) are negligible. So in the reflective region, we can correctly approximate the effective radiance at the sensor as

$$L = L_A + L_B + L_G + L_C$$
$$L = \left[E_s' \cos\sigma' \frac{r}{\pi}\tau_1(\lambda) + [FE_{ds} + (1-F)E_{bs}]\frac{r_d}{\pi}\right]\tau_2 + L_{us} \tag{4.69}$$

Furthermore, we can see that the solar irradiance term (A) and the path radiance term (C) are the largest contributors with the reflected skylight (B) and the reflected background (G) terms being significantly smaller, but not necessarily negligible. It's also clear that as the shape factor approaches 1, the reflected background radiance term will become a negligible contributor to the total radiance. This will be the case for any nearly horizontal surface that does not have taller objects or

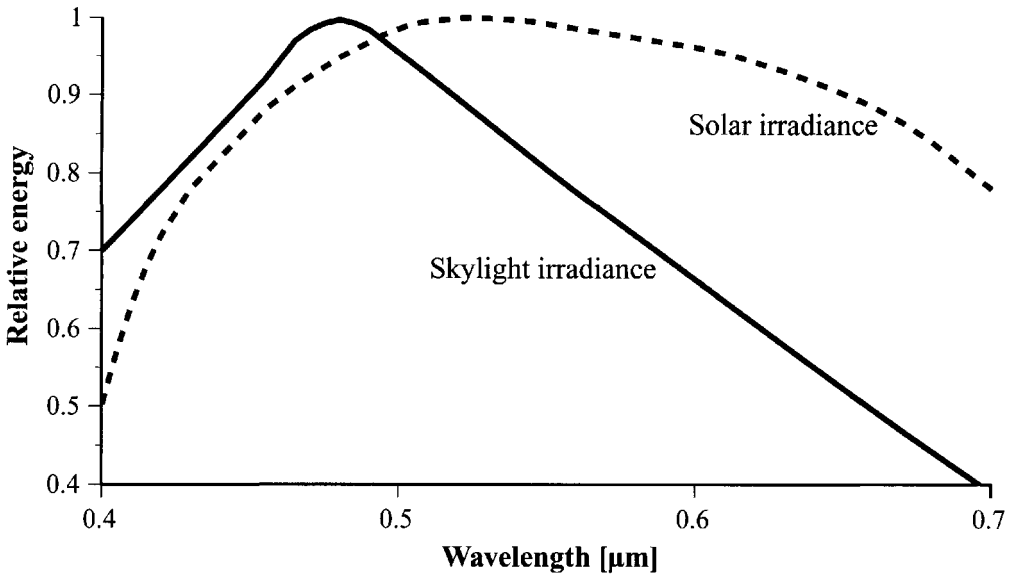


Figure 4.20 Comparison of peak normalized direct solar irradiance and skylight irradiance (after Piech and Walker, 1971).

high terrain in its vicinity. It's also clear from inspection of Table 4.1 that the relative magnitude of the reflected radiance ($A + B + G$) and the path radiance (C) will vary considerably depending on the reflectance of the target. For dark targets (1–3% reflectors), the path radiance may represent 50% or more of the total radiance. In this case, a small error in computation of upwelled radiance would result in a very large error in the computed reflectance. The importance of the upwelled radiance term never vanishes as reflectance increases, but the overall radiance (and computed reflectance) becomes less sensitive to errors in the path radiance. The radiance data for the 10% reflectors in Table 4.1 are reproduced as percent of total radiance in Table 4.2 to highlight the relative importance of each term.

The case presented in Table 4.1 is for a relatively clear atmosphere. The relative importance of both the skylight term and the path radiance term will increase as the atmosphere becomes more turbid and as we shift to shorter wavelengths. Piech and Walker (1971), for example, show the peak normalized spectral shape of the sunlight and skylight irradiance in the visible spectrum as plotted in Figure 4.20. They also observed that in the visible spectrum, the direct solar irradiance term could range from approximately seven times the downwelled irradiance under clear sky conditions to nearly equal values for the two terms under very hazy conditions.

Table 4.2 Illustration of the Percentage of Total Sensor-Reaching Radiance Contributed by Each Term in the Governing Equation. The examples shown are for the 10% reflectors ($\varepsilon = 0.9$) from Table 4.1.

$\Delta\lambda$	L_A	L_D	L_B	L_E	L_G	L_H	L_F	L_C
0.4-0.7 μm	35.4	0	9.9	0	0.6	0	0	54.3
3-5 μm	19.9	60.3	0.2	1.0	0.1	1.6	16.4	0.3
8-14 μm	0	76.3	0	1.2	0	2.0	20.4	0

Switching to the other end of the spectrum, we see from Table 4.1 that in the LWIR region the solar photons become a vanishingly small portion of the total radiance. Even the direct solar irradiance produces too few photons to make a significant contribution to the total irradiance. As a result, in the LWIR region the effective radiance reaching the sensor can usually be expressed with acceptable error as

$$L = L_D + L_E + L_H + L_F$$
$$L = \left[\varepsilon L_T + (FE_{d\varepsilon} + (1-F)E_{b\varepsilon}) \frac{r_d}{\pi} \right] \tau_2 + L_{u\varepsilon} \tag{4.70}$$

The target self-emission term (D) dominates this expression with significant contribution from upwelled radiance (F) (cf. Fig. 4.18). The reflected downwelled radiance (E) and reflected background radiance (H) terms are typically much smaller, though still significant contributors if measurement accuracies of tenths of a degree are desired. The relative importance of these reflected terms will decrease with increasing emissivity (decreasing reflectivity), but in general they will not be negligible until emissivity values approach 0.99. The relative importance of the downwelled radiance and the background radiance is controlled by the shape factor just as in the VIS-NIR region. For nearly horizontal unobstructed surfaces, the shape factor F approaches 1.0 and the background term becomes negligible. In Chapter 7, we will use this phenomenon to select calibration targets without background effects to simplify the equations used in computing the values of atmospheric parameters.

Finally, when we consider the MWIR window region for the case treated in Table 4.1, we see the effects of the intersection of the solar irradiance and self-emission curves of Figure 3.23. The significant contribution to the photon flux from solar effects and terrestrial self-emission results in all of the terms in Eq. (4.66) being potentially significant. The solar terms will be larger under high solar elevation conditions and for cold or low emissivity targets. The thermal terms will be more important for low solar elevation angles and higher temperature targets. However, in most cases the direct solar insolation term will be negligible only at night. In many cases, for low solar elevation conditions and targets with temperatures of 290 K or higher, the reflected solar downwelled radiance (B) and solar background effects (G) may be negligible. Because of the complexity of the radiometry in this window, quantitative analysis is usually only attempted on nighttime imagery. Even with nighttime imagery, care must be taken because of the strong spectral character of several of the radiometric terms, $L_{T\lambda}$, $\tau_2(\lambda)$, in this window. The spectral integral form of the radiance equation, Eq. (4.64) must be used rather than the simplification of Eq. (4.66) if quantitative results are required.

This simple relative magnitude analysis will serve us well in determining which terms can be eliminated from our analysis. This approach can be used to obtain an intuitive sense of the relative importance of the terms in the radiance equation. It is a useful exercise whenever you begin work in a new spectral region or on targets that are substantially different from those for which you've already developed an intuitive feel. This approach does not tell us which terms or param-

eters will generate the dominant sources of error or how an error in one parameter will affect the radiance observed or the temperature or reflectance measured. To obtain this type of error information, a more detailed error or sensitivity analysis is required.

4.6.2 Sensitivity Analysis—Error Propagation

Before beginning this discussion, we should remind the reader of the often-neglected distinction between accuracy and precision, which are both measures of error. Precision describes the repeatability of a measurement. It is often characterized by the standard deviation from the mean of many measurements. For example, if we measured the reflectance of a target 20 times and computed a mean reflectance of 0.18 with a standard deviation of 0.02, we could claim that the precision of the measurement to one standard deviation was 0.02 or two reflectance units. Accuracy, on the other hand, describes how closely an instrument or procedure can match some standardized value or what we have defined to be truth. It is often characterized by the deviation between the mean of several measured values and the true value. In the case just cited, if the true reflectance value of the sample was 0.17, we would have an accuracy associated with the measurement process of 0.01. The individual measurement error that describes how closely any individual measurement comes to truth is often taken to be the root sum square error value, i.e.,

$$s_m = (s_p^2 + s_i^2)^{1/2} \tag{4.71}$$

where s_p is the precision of the measurement, s_i is the accuracy of the measurement instrument or approach, and s_m is the total error and can be thought of as the error associated with the individual measurement (i.e., 2.2 reflectance units in our example). Note that, in many cases, calibration procedures can generate unbiased errors such that the average of many readings is a very good estimate of the true value (i.e., $s_i \approx 0$). In this case, the precision of the measurement approach becomes a good estimate of the error.

In general, the error (precision, accuracy, or total) of a measurement approach is the result of errors in the procedures or values that go into that measurement. For the case where a governing equation can be used to describe a parameter of interest, a relatively simple expression can be written to describe the relationship between the errors [cf. Beers, (1957)]. In the simplest case, if we can define a dependent variable Y in terms of one or more independent (i.e., uncorrelated) variables X_i, i.e.,

$$Y = f(X_1, X_2 \ldots X_N) \tag{4.72}$$

then we can express the error in Y (s_Y) as:

$$S_y = \left[\left(\frac{\partial Y}{\partial X_1} S_{X1} \right)^2 + \left(\frac{\partial Y}{\partial X_2} S_{X2} \right)^2 \cdots + \left(\frac{\partial Y}{\partial X_N} S_{XN} \right)^2 \right]^{1/2} \tag{4.73}$$

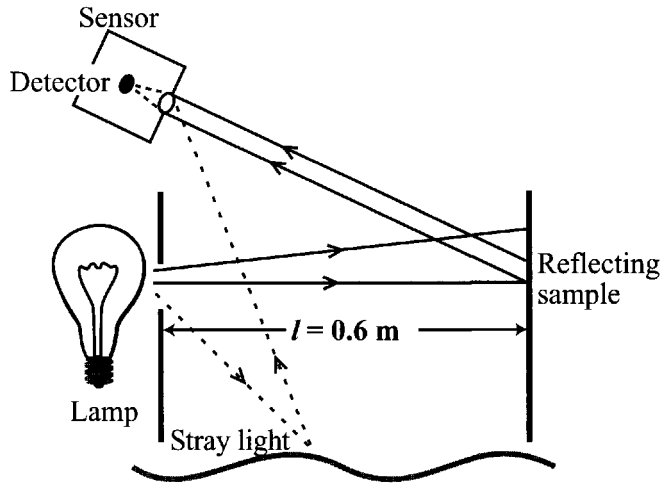


Figure 4.21 A device for measurement of reflected radiance.

where s_{Xi} is the error in the individual input variables. The partials of the dependent variable (Y) with respect to the input variables describe the sensitivity of Y to small changes in X. Multiplying the partial derivative by the error on the input variable X_i generates the error in Y (s_{Yi}) associated with an error in X_i. The total error is just the square root of the sum of the squared values because independent errors tend to add in quadrature. Beers (1957) also points out that for the case where the input variables X_i values are correlated, Eq. (4.73) must be modified to reflect how an error in one input variable may be exaggerated or compensated for by the correlation with a second input variable. In this case, Beers expresses the error as

$$s_Y = \left[\left(\frac{\partial Y}{\partial X_1} s_{X1} \right)^2 + \left(\frac{\partial Y}{\partial X_2} s_{X2} \right)^2 \cdots + \left(\frac{\partial Y}{\partial X_N} s_{XN} \right)^2 + \Sigma 2\rho_{ij} \frac{\partial Y}{\partial X_i} \frac{\partial Y}{\partial X_j} s_{Xi} s_{Xj} \right]^{1/2} \tag{4.74}$$

where ρ_{ij} is the standard correlation coefficient between variables X_i and X_j, and the sum is over all combinations of correlated variables. Note that ρ_{ij} can take on values from -1 to 1, so the inclusion of the correlation term will reduce the error computed for negatively correlated variables.

This type of error propagation analysis not only lets us assess the overall error, but also, by inspecting the individual error terms (s_{Yi}), tells us what parameters contribute the most to the error. This helps to prioritize error reduction efforts, since in most cases reducing the error in a small error source will have little or no impact on the total error [i.e., we first want to work on the weak links (large error sources) in the radiometric image chain].

A complete treatment of the error analysis of the "big equation" is beyond the scope of this book. However, throughout the book, we will refer to the relative im-

portance of variables or error terms that are the result of this type of error analysis. To see how this analysis is performed, we will look at one simple case.

Consider a laboratory instrument that measures the radiance reflected from a sample irradiated as shown in Figure 4.21. The lamp is a calibration standard of known radiant intensity that produces an irradiance (E) on the sample of 110 [Wm^{-2}]. The expression governing the observed radiance (L) for a Lambertian reflector (r_d) can be expressed as

$$L[W\mathrm{m}^{-2}\mathrm{sr}^{-1}] = E\frac{r_d}{\pi} + L_p \tag{4.75}$$

where L_p is any radiance scattered to the sensor through a process other than reflection from the target. We can rearrange Eq. (4.75) to provide the governing equation for using this device to measure reflection, i.e.,

$$r_d = \frac{(L - L_p)\pi}{E} \tag{4.76}$$

From Eq. (4.74), we can express the error in the reflectance (s_r) as

$$\begin{aligned}(s_r) &= \left[\left(\frac{\partial r}{\partial L}s_L\right)^2 + \left(\frac{\partial r}{\partial L_p}s_{L_p}\right)^2 + \left(\frac{\partial r}{\partial E}s_E\right)^2 + \left(\frac{\partial r}{\partial \pi}s_\pi\right)^2\right]^{1/2} \\ &= \left[\left(\frac{\pi}{E}s_L\right)^2 + \left(-\frac{\pi}{E}s_{L_p}\right)^2 + \left(\frac{(L_p - L)\pi}{E^2}s_E\right)^2\right]^{1/2}\end{aligned} \tag{4.77}$$

where we have assumed s_π to be approximately zero and that all the variables are uncorrelated. For a particular measurement, we observe the radiance (L) to be 3.50 [$Wm^{-2}sr^{-1}$], and we know that the instrument error (s_L) is ±0.17 [$Wm^{-2}sr^{-1}$]. The scattered radiance term $L_p \pm s_{Lp}$ is observed to be 0.18 ± 0.03 [$Wm^{-2}sr^{-1}$]. However, the lamp error is not expressed in terms of irradiance. The standards laboratory specifies the lamp assembly to have a radiant intensity ($I \pm s_I$) of 39.6 ± 1.7 [Wsr^{-1}] from which we computed the irradiance onto the sample, 0.6 [m] away, to be

$$E = \frac{I}{l^2} = \frac{39.6}{(.6)^2} = 110[W\mathrm{m}^{-2}] \tag{4.78}$$

The error in irradiance can then be expressed as

$$\begin{aligned}s_E &= \left[\left(\frac{\partial E}{\partial I}s_I\right)^2 + \left(\frac{\partial E}{\partial l}s_l\right)^2\right]^{1/2} \\ &= \left[\left(\frac{1}{l^2}s_I\right)^2 + \left(\frac{-2I}{l^3}s_l\right)^2\right]^{1/2}\end{aligned} \tag{4.79}$$

If we estimate the error in our distance measurement s_l to be 0.002 m, we can generate a value for the error in irradiance using Eq. (4.79):

$$s_E = \left[\left(\frac{1}{(0.6)^2}1.7\right)^2 + \left(\frac{-2\cdot 39.6}{(0.6)^3}0.002\right)^2\right]^{1/2}$$
$$= [22.3 + 0.54]^{1/2} = 4.8[W\text{m}^{-2}] \quad (4.80)$$

where we notice that the distance measurement error contributes very little to the total error.

We can now solve Eq. (4.77) for the error in reflectance. First we solve for the reflectance of the sample using Eq. (4.76), i.e.,

$$r = \frac{(3.50 - 0.18)\pi}{110} = 0.095 \quad (4.81)$$

The error in reflectance is then given by Eq. (4.77) to be

$$s_r = \left[\left(\frac{\pi[sr]\cdot 0.17[W\text{m}^{-2}sr^{-1}]}{110[W\text{m}^{-2}]}\right)^2 + \left(-\frac{\pi\cdot 0.03}{110}\right)^2 + \left(\frac{(0.18-3.5)\pi\cdot 4.8}{110^2}\right)^2\right]^{1/2}$$
$$= \left[235.7\cdot 10^{-7} + 7.3\cdot 10^{-7} + 171.2\cdot 10^{-7}\right]^{1/2} = 0.006 \quad (4.82)$$

We can then express the observed reflectance to be 0.095 ± 0.006. Furthermore, we see that the largest contributor to the total error is the sensor error, closely followed by the error due to the lamp standard. It is important to realize that, for this type of analysis to yield reasonable results, all of the input errors must be of a common form (e.g., one standard deviation or RMS deviation from truth). This simple case study serves to illuminate the basic concept of error propagation. Schott et al. (2004) extensively use this approach to assess the errors associated with thermal infrared remote sensing systems. In cases where a governing equation cannot be written (e.g., when iterative solutions are used), this same type of sensitivity analysis can be performed using computer simulation of the process and Monte Carlo methods [cf. Rubinstein (1981)].

In this chapter, we have developed in detail a governing equation that describes the radiance reaching the sensor. We also analyzed how this equation can be simplified in certain spectral regions. Finally, we have briefly described error analysis methods that are used throughout the book to assess the errors and the importance of parameters. In the next chapter, we will analyze how the radiance reaching the sensor is converted by various sensors to recorded signal levels.

4.7 REFERENCES

Beers, Y. (1957). *Introduction to the Theory of Errors*. Addison-Wesley, Reading, MA.

Chandrasekhar, S. (1960). *Radioactive Transfer*. Dover, Minneola, NY.

Deering, D.W. (1988). "Parabola directional field radiometer for aiding in space sensor data interpretation." Proceedings of the SPIE, pp. 924-933, Orlando, FL.

Feng, X, Schott, J.R., & Gallagher, T.W. (1993). "Comparison of methods for generation of absolute reflectance factor values for BRDF studies." *Applied Optics,* Vol. 32, No. 7, pp.1234-1242.

Kneizys, F.X., Shettle, E.P., Abreu, L.W., Chetwynd, J.H., Anderson, G.P., Gallery, W.O., Selby, J.E.A., & Clough, S.A. (1988). "Users guide to LOWTRAN7," AFGL-TR-88-0177, Environmental Research Papers, No. 1010, Air Force Geophysics Laboratory, Optical/Infrared Technology Division, Hanscom AFB, MA.

Piech, K.R., & Walker, J.E. (1971). Aerial color analysis of water quality. *Journal of Survey and Mapping Division, American Society of Civil Engineers*, Vol. 97, No. SU2, pp. 185-197.

Priest, R.G., & Germer, T.A. (2002). "Polarimetric BRDF in the microfacet model: theory and measurements." Proceedings of the 2000 Meeting of the Military Sensing Symposia Specialty Group on Passive Sensors Vol. 1,: pp. 169-181.

Rubinstein, R.Y. (1981). *Simulation and the Monte Carlo Method.* Wiley, NY.

Salvaggio, C., Sirianni, J.D., & Schott, J.R. (1993). "Use of LOWTRAN derived atmospheric parameters in synthetic image generation models." Proceedings of the SPIE, Recent Advances in Sensors, Radiometric Calibration, and Processing of Remotely Sensed Data, Vol. 1938, Orlando, FL.

Schott, J.R., Brown, S.D., & Barsi, J.A. (2004). Calibration of Thermal Infrared (TIR) Sensors. In J. Lavall & D. Quattrochi, eds. *Thermal Remote Sensing in Land Surface Processing*, pp. 363-404. CRC Press Boca Raton, FL.

Torrance, K.E., & Sparrow, E.M. (1967). Theory for off-specular reflection from roughened surfaces. *Journal of the Optical Society of America* 57, pp 1105-1114.

Turner, R.E., Malila, W.A., Nalepka, R.F., & Thomson, F.D. (1975). Influence of the atmosphere on remotely sensed data. Proceedings SPIE, Vol. 51, pp. 101-114.

Wellems, D., Serna, M., Sposato, S.H., Fetrow, M.P., Arko, S.A. & Caudill, T.R. (2000). "Spectral polarimetric BRDF model and comparison to measurements from isotropic roughened glass." Special Report RD-WS-01-01, U.S. Army Aviation and Missile Command, Workshop on Multi/Hyperspectral Sensors, Measurements, Modeling and Simulation.

CHAPTER 5
SENSING SYSTEMS

In Chapter 4, we derived an expression for the radiance reaching the sensor. In this chapter, one of our goals is to see how we can use imaging sensors to record and measure that radiance.

The treatment of sensors is one of the most involved and exciting aspects of remote sensing. Traditionally, the sensor has been the most complex and expensive part of the image chain. It has been the jewel designed to offset, or blind one to, any flaws in the other links. Today the sensor is still a critical and expensive component in the image chain. However, the user community is placing increasing emphasis on the processing, storage, and information extraction stages of the image chain (cf. Chaps. 7–12). The details of sensor design or a rigorous characterization of even a few sensors are beyond the scope of this treatment. Indeed, the rapid rate of sensor evolution brought on by advances in semiconductor technology makes detailed sensor characterization obsolete almost as fast as it can be published in text form. We will therefore concentrate on a broad treatment of the fundamental principles of sensor technology. This is intended to provide the reader with a basic knowledge of traditional sensor designs and a capability to understand more detailed studies of new systems as they appear in the current literature.

This chapter emphasizes image acquisition and quantitative detection of radiation from a systems design perspective. This addresses the question of how we convert the radiance reaching the entrance aperture of the imaging system into a measurable, recordable, and quantifiable signal. A more specific treatment of various instrument design concepts is given in Chapter 6. The reader unfamiliar with the basic optical terminology referred to in this chapter should consult Hecht (1990).

5.1 CAMERAS AND FILM SYSTEMS

Film offers an inexpensive, high-resolution solution to many remote sensing problems. Much remote sensing is still, and will continue to be, conducted using photographic film camera systems. The images can be acquired with instruments ranging from the conventional 35-mm format cameras, used by many amateurs, to the large-format cameras used on the space shuttle (cf. Fig. 5.1). Film also offers great geometric fidelity for mapping and mensuration at very modest costs, making it very attractive for most photogrammetric applications. Since many readers have a general familiarity with photography, we will use film systems as a point of reference for the more exotic electro-optical systems to follow. Because of our interest in radiometric and computer-based image analysis, this section will emphasize how film systems can be used as quantitative radiometric sensors and as a source of digital image data. For a more complete treatment of film systems, refer to Chapter 6 of the *Manual of Remote Sensing* (Colwell 1983).

5.1.1 Irradiance onto the Focal Plane

As we will see in the next section, film systems are characterized in terms of the *exposure* (H) on the film, i.e.,

$$H[\text{joules m}^{-2}] = \int E(t)dt \cong Et \tag{5.1}$$

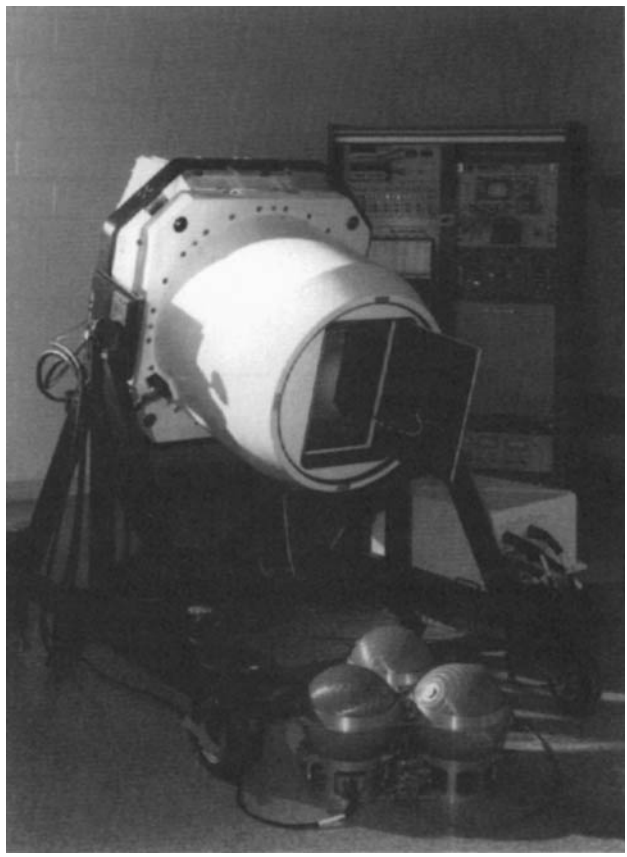
where t [sec] is the exposure time and E [Wm^{-2}] is the irradiance onto the film. In the last chapter, we dealt extensively with the radiance reaching the entrance aperture of a sensing system. In order to determine how that radiance will impact a sensor (the film, in this case), we need to develop a relationship between radiance reaching the front of the sensor and irradiance at the focal plane. This is commonly referred to as the *camera equation* and is written as

$$E = \frac{L}{G\#} \tag{5.2}$$

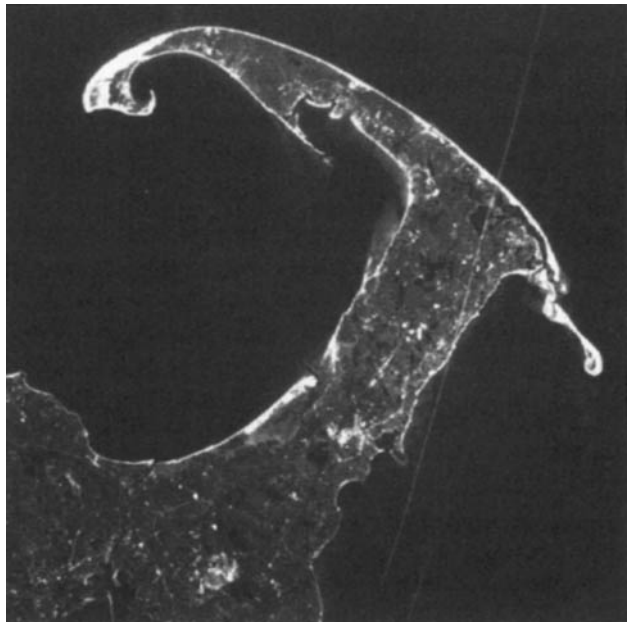
where the $G\#$ [sr^{-1}] defines the throughput of the system in terms of how well it converts radiance to irradiance on the focal plane. For a simple camera, the $G\#$ is usually defined in terms of the value along the optical axis. From Chapter 3, we recall that we can express the element of irradiance due to the radiance from each element of solid angle through an aperture (cf. Fig. 5.2) as

$$dE = \tau L_i d\Omega_i \cos\theta_i \tag{5.3}$$

where τ is the transmittance of the lens assembly, L_i is the radiance at location i in the aperture, $d\Omega_i$ is the solid angle subtended by an area element dA_i in the aperture, and θ_i is the angle from the optical axis to the area element dA_i. If the aperture is small compared to the focal length, then for a field of uniform radiance the aperture can be approximated as a point source with $\theta_i = 0$ and $r_i = r$. This yields



(a) Large-format camera (LFC) flown on board the space shuttle.



(b) Portion of an image from the LFC showing Cape Cod. Note the condensation trails and their shadows.

Figure 5.1 Examples of a specialized camera used for remote sensing.

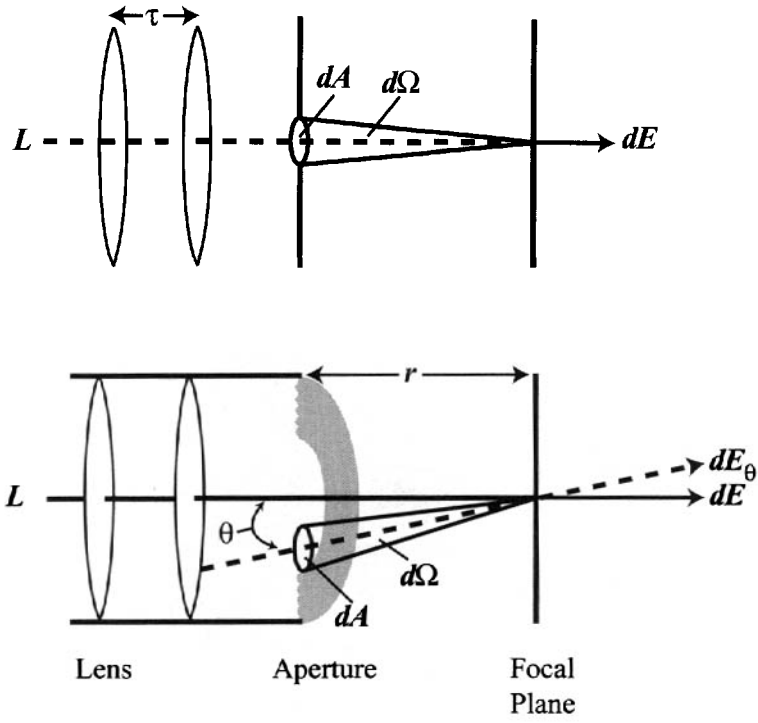


Figure 5.2 Derivation of the camera equation.

$$E = \int_A dE_i = \int L\tau \frac{dA_i}{r_i^2} = \frac{LA\tau}{r^2} \tag{5.4}$$

For a system focused at infinity (which is the case for most remote sensing systems), r is the focal length (f) of the optical system. The irradiance can then be expressed as

$$E = \frac{L\pi d^2\tau}{4f^2} = \frac{L\pi\tau}{4(f\#)^2} = \frac{L}{G\#} \tag{5.5}$$

where d is the diameter of the aperture (i.e., $\pi d^2/4 = A$) and the *F number* ($f\#$) of the system is defined as

$$f\# = \frac{f}{d} \tag{5.6}$$

Rearranging Eq. (5.5), we have an expression for the $G\#$ of

$$G\# = \frac{L}{E} = \frac{4(f\#)^2}{\tau\pi} \tag{5.7}$$

which is valid for large values of $f\#$. Grum and Becherer (1979) indicate that Eq. (5.7) is valid to better than 1% for $f\#$ values greater than 10, and they solve the integral of Eq. (5.3) rigorously for a simple lens to yield

$$E = \frac{\tau\pi L \frac{d^2}{4}}{\frac{d^2}{4} + f^2} \tag{5.8}$$

which is valid for all $f\#$ values, i.e.,

$$G\# = \frac{1 + 4(f\#)^2}{\tau\pi} \tag{5.9}$$

From our analysis of radial lens falloff in Chapter 3, i.e., Eq. (3.91), we recognize that the off-axis irradiance or exposure will be further reduced by approximately $\cos^3 \theta$. In the next section, we will analyze how this exposure is recorded by the film.

5.1.2 Sensitometric Analysis

In order to describe photographic systems quantitatively, we need to introduce some terminology for numerically describing the photographic image. In general, we will be dealing with photographic transparencies (the equivalent of black -and-white negatives or color slides in amateur photography), and the common way to quantify them is in terms of the film density or opacity. *Density* is defined as

$$D = -\log_{10} \tau \tag{5.10}$$

where τ is the fraction of incident radiation transmitted through the film (transmission). Density can be defined as a function of wavelength, but it is more commonly defined over a spectral bandpass corresponding to how the data are to be presented to the viewer or instrument. For example, for color film we would have red, green, and blue density values corresponding to film transmission measurements made through filters that only transmit in the red, green, and blue spectral regions, respectively.

The density (D) of developed film is related to the exposure (H) of the film through what is called the D log$H$ curve after Hurter and Driffield (1890), who developed the method of describing film response shown in Figure 5.3. The curve shown is for one spectral emulsion layer of a positive working color transparency. The slope [gamma (γ)] of the straight-line portion is negative for film positives because the greater the exposure, the thinner (less dense) the film and, therefore, the brighter it will appear when viewed or projected in transmission mode. Note that the curve is plotted in terms of log exposure and that, at low exposures (high densities on a positive), the film becomes rapidly less responsive until changes in exposure have no impact on the film. The same effect occurs at the toe of the curve for high exposures. The approximately linear region between the shoulder and the

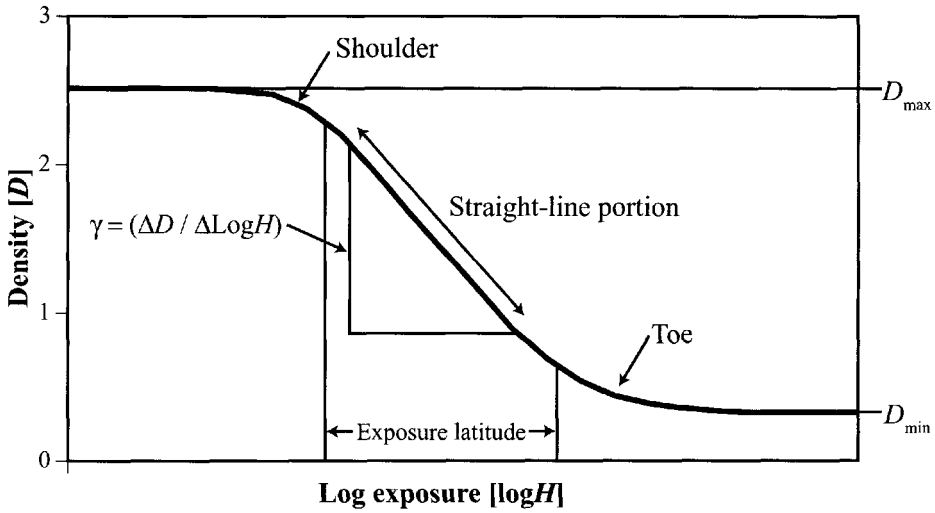
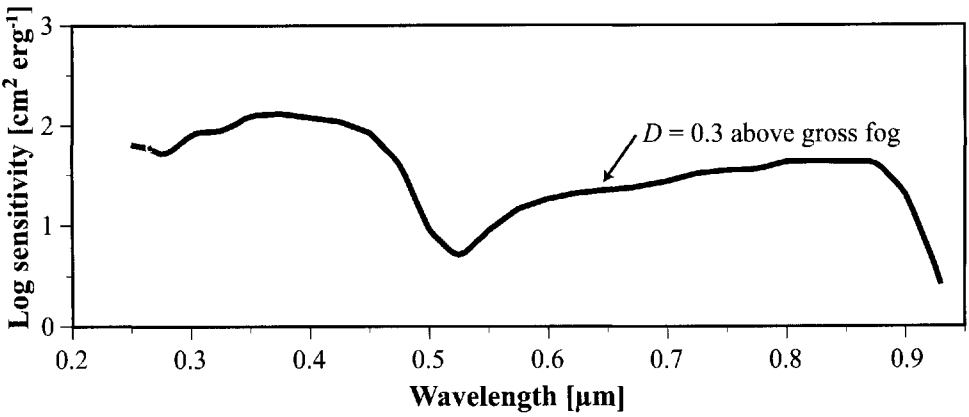


Figure 5.3 Characteristics of an H and D curve.

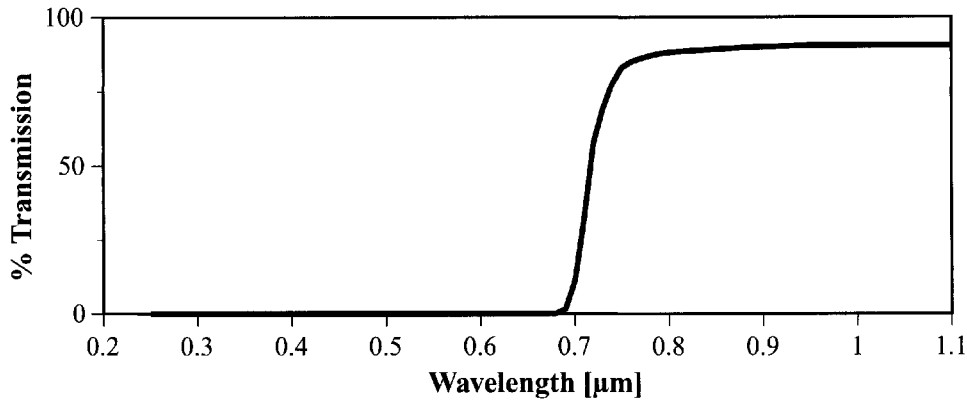
toe of the curve is where one normally wants the image information to fall. The projection of this linear region on the log exposure axis defines the film's exposure latitude. Objects with exposure values much beyond this region will be over- or underexposed. The gamma of the film defines the contrast. Since the D_{max} of a film is largely limited by the film type, there is an inverse relationship between contrast and exposure latitude. The D log$H$ curve is also referred to as the *characteristic curve* since it is indicative of the response characteristics of a particular film type. It is not, however, fixed for each film type. It will vary considerably between batches of film, with the physical and chemical characteristics of the film development and with the exposure time. Most films exhibit some time hysterisis (reciprocity law failure) effects such that identical exposures with significantly different exposure times will produce different density values.

The spectral sensitivity of film is controlled by the chemical composition of the film. A thorough treatment of film chemistry and analysis can be found in James (1977). The effective sensitivity of a film system is often controlled by cascading the film's intrinsic spectral sensitivity with a filter of a chosen spectral transmission. This is illustrated in Figure 5.4, where visible and IR-sensitive film is filtered with a visible-blocking (longpass) filter to achieve a film filter combination sensitive just to near infrared (NIR) wavelengths. The sensitivity of film is commonly defined as the reciprocal of the exposure needed to cause a chosen film density for the processing conditions specified. The units of sensitivity are cm^2/erg.

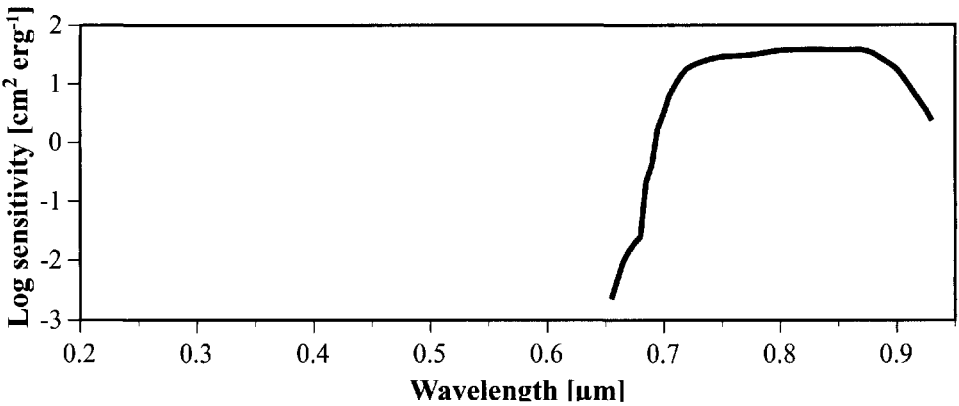
For color film, the film is generally composed of three absorbing layers sensitive to three different spectral regions. The density of the developed film in the red, green, and blue spectral regions is controlled by the amount of exposure in each of the three absorbing layers. This is illustrated in Figure 5.5 for the color infrared film commonly flown for vegetation studies. When flown with a minus blue (yellow) filter, the absorption layers roughly isolate green, red, and NIR energy. The spectral transmission of the corresponding dye layers in the developed film are also



(a) Film sensitivity.

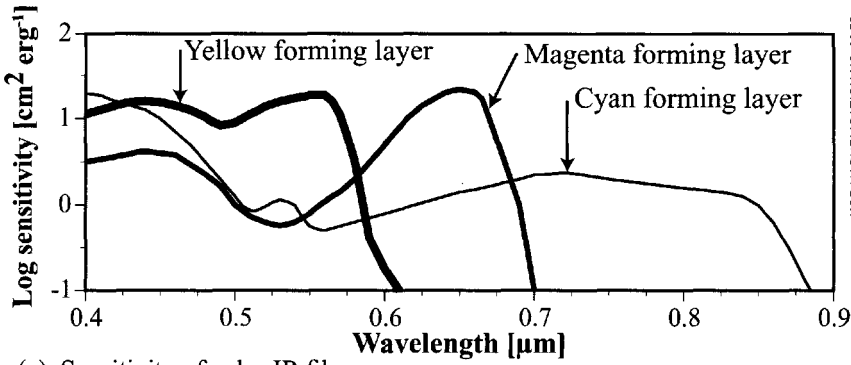


(b) Kodak Wratten #89B filter.

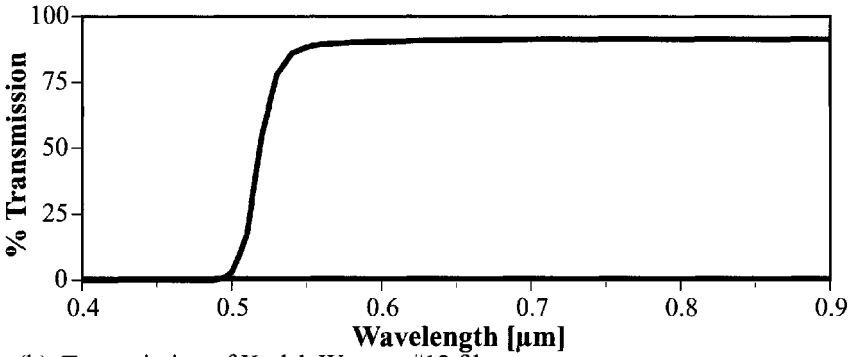


(c) Effective sensitivity of film/filter combination.

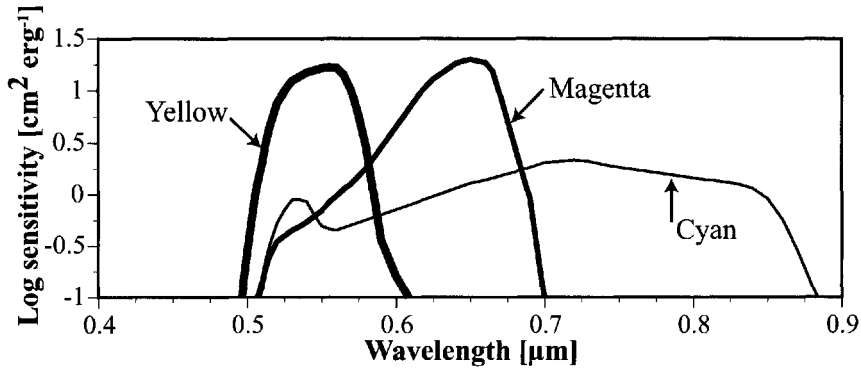
Figure 5.4 Spectral sensitivity of Kodak type 2424 film and its effective sensitivity when flown with a Kodak Wratten #89B filter.



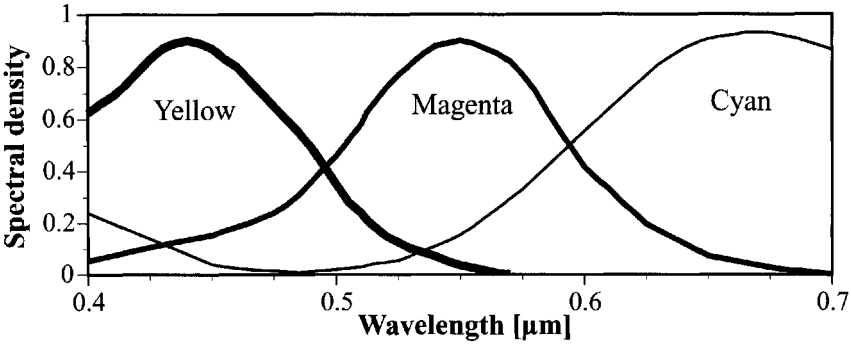
(a) Sensitivity of color IR film.



(b) Transmission of Kodak Wratten #12 filter.



(c) Effective sensitivity of film/filter combination.



(d) Transmission of dye layers in developed image.

Figure 5.5 Spectral characteristics of color infrared film (Kodak infrared 2443 when flown with a Kodak Wratten #12 filter).

shown in Figure 5.5. From this we see that the density of the red layer (i.e., through a red filter) of the developed film depends on the NIR exposure, the density through a green filter carries red exposure information, and the blue density carries green information.

From a practical standpoint, the quantitative user of film cannot count on the chemical system remaining stable for all the variables that affect the D log$H$ curve. Instead, the characteristic curve must be produced for each roll of film developed (ideally for each frame on long rolls). This is accomplished using a device called a *sensitometer*, which places a step wedge (or step tablet) on a portion of the film, before (header) or after (trailer) the images. The step tablet contains a sequence of known exposure steps. By plotting the density versus the log of the exposure used to create the step, the D log$H$ curve can be produced. For color film, densities are measured with red, green, and blue filters to characterize the cyan, magenta, and yellow dye layers. In making the step wedges, care must be taken to ensure that the spectral shape of the lamp used to expose the film approximately matches the spectral shape of the flux incident on the sensor (usually the source is filtered to about 5500–6000 K) and that any filters to be used with the film are also used in the sensitometer. Finally, to avoid reciprocity law failure (time hysterisis effects), the exposure time used in putting the step wedge on the film should approximately match the exposure time used with the film.

After processing, the film density (for each layer if color film) is measured for each step with the same type of densitometer to be used on the image data. The plot of density versus log exposure can then be produced as shown in Figure 5.3. Film transmission and, therefore, density are dependent on the scattering properties of the film, as well as the illumination and collection optics of the densitometer (transmissometer) used (cf. Fig. 5.6). While procedures for correlating densities between instruments can be developed, it is advisable where possible to analyze the step wedge and the image data with the same instrument.

The density versus log exposure data can be fit to a polynomial of the form

$$D = b(0) + b(1)logH + b(2)(logH)^2 + b(3)(logH)^3 \tag{5.11}$$

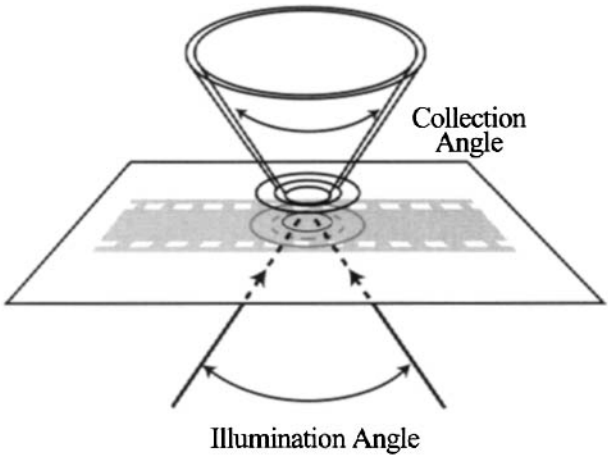
where the constants $b(0)$, $b(1)$, etc., are determined by a least-squares regression. Alternately, the density and log exposure values can be tabulated and a piecewise linear interpolation used to convert between density and log exposure.

The radiance corresponding to any density value can be found by combining Eqs. (5.7), (5.1), and (3.91) to yield

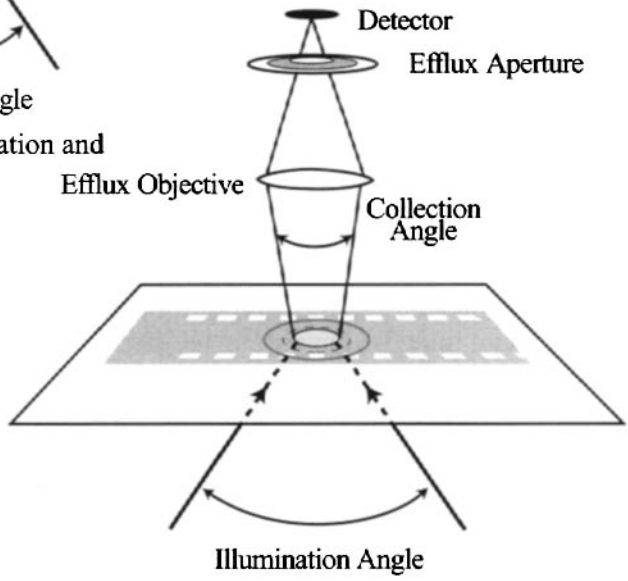
$$L = \frac{HG\#}{t\cos^n \theta} \tag{5.12}$$

where the value of H must be measured from the D log$H$ curve for each density value. The value of n can be calculated by calibration of the performance of the lens, as discussed in Section 3.3. The view angle θ can be expressed as

$$\theta = \tan^{-1}\left(\frac{r}{f}\right) \tag{5.13}$$



(a) Double diffuse; wide illumination and collection angles.



(b) Diffuse; wide illumination and narrow collection angles.

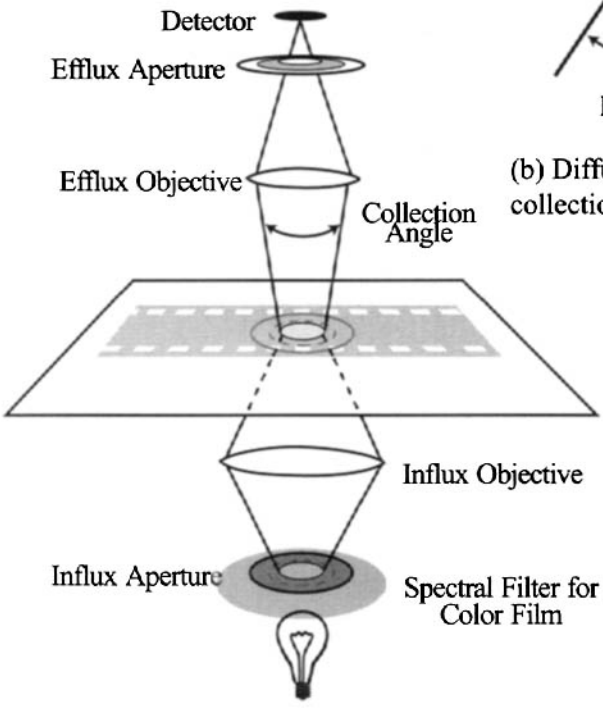


Figure 5.6 Types of density reading and instrumentation.

where r is the radial distance from the center of the film format to the point where the density measurement is made. The value of L is the effective radiance reaching the sensor corresponding to the value derived in Chapter 4. The normalized sensitivity is used in place of the normalized responsivity to compute the effective radiance for film systems according to

$$L = L_{\text{eff}} = \int L_{\lambda} \left(\frac{R_s(\lambda)}{R_{s\max}} \right) d\lambda \tag{5.14}$$

where $R_s(\lambda)$ is the sensitivity, $R_{s\max}$ is the maximum sensitivity, and the subscript eff is generally dropped from the effective radiance term for convenience.

Using the procedures described above, we can use a camera as a two-dimensional radiance meter. When many points on an image are to be radiometrically analyzed, or when we need to process the entire image, it is often useful to digitize the image. This involves dividing the image into a two-dimensional array of picture elements (pixels) and assigning to each pixel a digital value corresponding to the density or transmission through the film at that pixel location. Traditionally, the digitizing process has been done with scanning microdensitometers. The film is scanned in a raster fashion with density values sampled at equal intervals in x and y and converted to scaled digital counts. By scanning the step wedge at the same time, a relationship between digital count and exposure can be developed for the image data and used in place of the density log exposure curve to characterize and analyze the film. With the recent advances in digital image scanning technology associated with the electronic-publishing industry, there are many low-cost image scanning systems available. Many of these can be used for quantitative remote sensing if proper care is taken to ensure the integrity of the data. (The user should test for dynamic range, spatial uniformity, geometric fidelity, and spectral separability of the film layers if color film is used.) By scanning in the step tablet at the same time as the image data, the digital count to exposure relationship can be directly measured from the tablet and applied to the digitized image data (cf. Fig. 5.7).

In this subsection, we have seen how we can use cameras and photographic film to produce imaging radiometers. For relatively broad spectral windows between 0.4 and 0.9 μm, film cameras can often provide a simple low-cost sensing solution. When narrow-band spectral sampling is required, film may not be fast enough. In addition, when spectral regions beyond about 0.9 μm are of interest or for unmanned satellite sensors, other approaches must be considered. This brings us to electro-optical (EO) systems, which will be the topic of the remainder of this chapter.

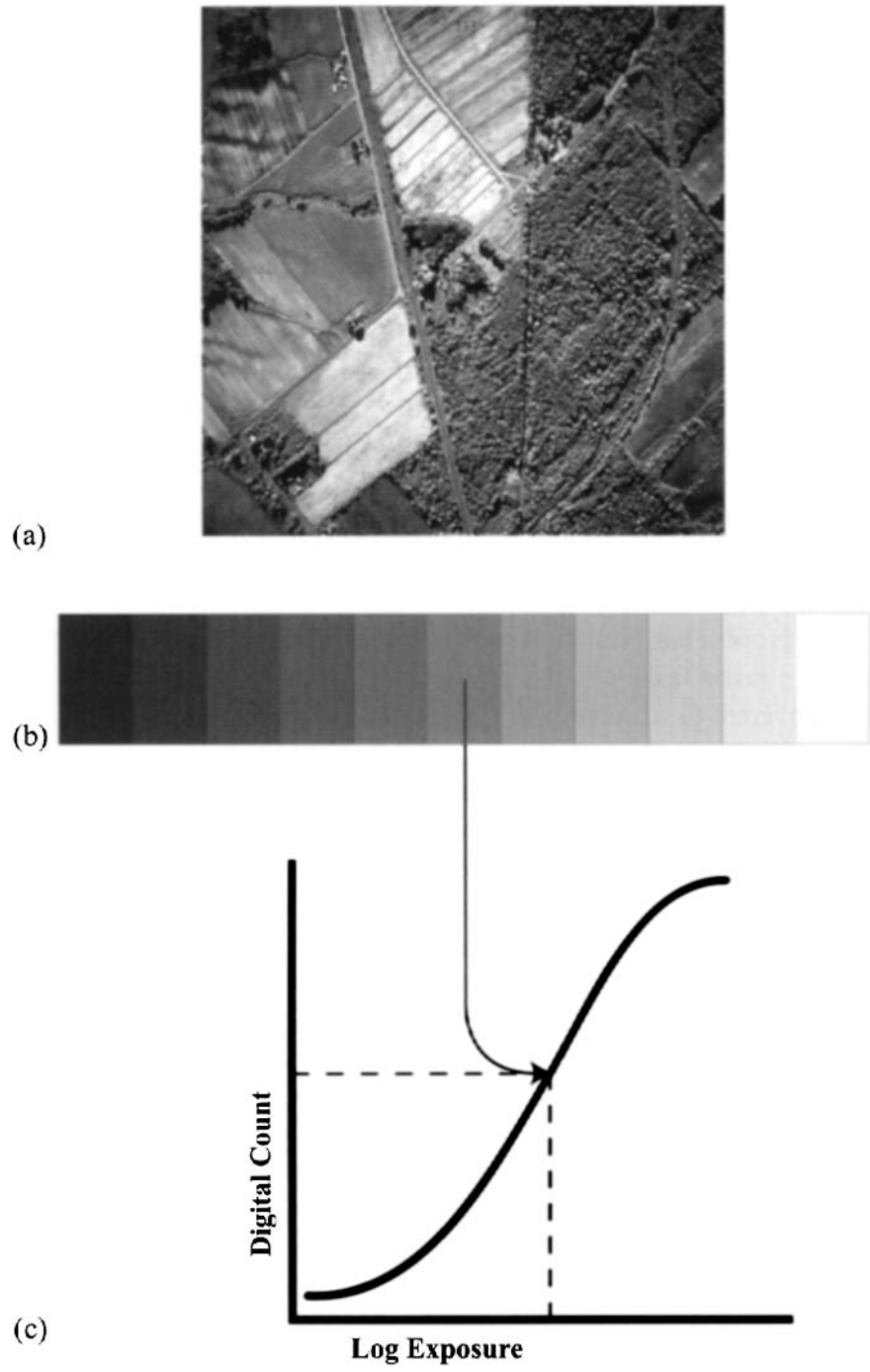


Figure 5.7 Digitized color infrared film image (a), an illustration of a step tablet from the film header (b), and the digital count versus relative log exposure for the red film layer (c). See color plate 5.7.

5.2 SIMILARITIES BETWEEN SIMPLE CAMERAS AND MORE EXOTIC ELECTRO-OPTICAL IMAGING SYSTEMS

Remote sensing film camera systems seem fairly simple to us because of their similarity to the amateur cameras we all use. In many ways, even the most exotic space-based film camera systems are just very large versions of conventional cameras. When we begin to look at electro-optical cameras (or *imagers* as they are more commonly called), many of the similarities to conventional cameras are still there, but they are less obvious. In this subsection, we will briefly examine common optical designs used in EO cameras and a basic breakdown of the components of EO imaging systems.

5.2.1 Optics and Irradiance at the Focal Plane

Conventional cameras, and some EO cameras, use refractive optics (e.g., glass lenses) to focus an image of the scene onto the focal plane. However, many EO systems used for remote sensing are designed to operate over a broad spectral range, which generally makes the use of refractive optics impossible or at least very expensive. Conventional optical glass, for example, does not transmit in the LWIR, and the most convenient lens materials for use in the LWIR (e.g., germanium) do not transmit in the visible wavelengths. To overcome this limitation, many EO systems use reflective optics (mirrors). This is particularly true of space-based systems because the reflective optics are much lighter than comparable refractive optics. Even for sensing in the VIS-NIR from space, reflective optics are generally used for the larger optical elements. While even an introductory treatment of optical systems is beyond the scope of this study, there are a few optical designs commonly used in EO systems that are briefly described. A more complete treatment of EO optical design can be found in Accetta and Shumaker (1993).

The function of reflective optics is the same as for conventional refractive optics (i.e., to focus the image of the Earth onto the focal plane). Because reflective optics reflect all wavelengths the same amount, they have another advantage over refractive optics in bringing all spectral bands to focus in the same plane. A disadvantage of reflective optics is that it is difficult to use the full aperture of the system because the optical elements are opaque. Figure 5.8 shows several designs for the long focal length optical systems commonly used in remote sensing. The off-axis parabola uses only a portion of a parabolic shape to bring the image to focus alongside the primary beam where detectors or film can be introduced. This design is very long (i.e., the imaging system must be as long as the focal length of the lens), making it particularly unattractive for many space applications. It is, however, commonly used in laboratory collimators for calibration and testing of other optical systems. Test targets are placed at the focal plane and projected onto the sensor's optics simulating targets effectively located at infinity. The Newtonian design uses a simple fold to make the image accessible. This has the disadvantage of obscuring some of the incident flux, and still makes for a long optical system. Another way to make the image accessible is by placing a hole in the center of the primary mirror

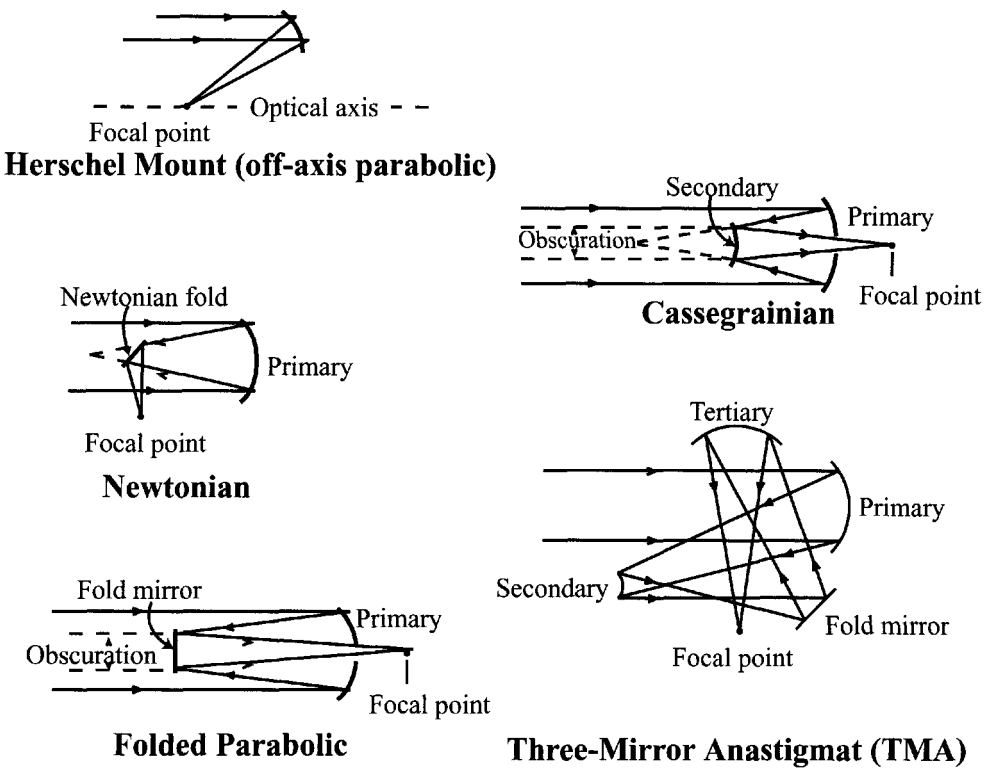


Figure 5.8 Optical designs commonly used for remote sensing. Adapted from Wolfe and Zissis (1985.)

and folding the energy back through the hole (folded parabolic). This design shortens the overall system but causes considerable obscuration. This limitation is reduced with the Cassegrainian design that uses a parabolic primary and a hyperbolic secondary to fold the energy back through a hole in the primary. This approach has less obscuration than the simple folded configuration and a much shorter overall length to achieve the same focal length. As a result, the *Cassegrain* and its variants are among the most commonly used designs, particularly for satellite systems. The conventional Cassegrain design has considerable aberration. This can be largely eliminated by making both the primary and the secondary aspheric. This more complex (i.e., expensive) design is called a *Ritchey-Chretian*. A somewhat simpler variation of the Cassegrain called the *Dall-Kirkham* uses an aspheric primary and a spherical secondary. This design still has some aberration, but it is often acceptable in systems with a narrow field of view through the telescope. For systems where wide fields of view (i.e., several degrees in one dimension) are required, a three-mirror anastigmatic (TMA) design is widely used. A Z-fold variation of the TMA achieves a relatively wide field of view in one dimension with no obscuration.

Whatever the optical configuration, we will be interested in computing the irradiance onto the focal plane. For systems using Cassegrainian-type optics with a centrally obscured aperture, the $G\#$ of Eq. (5.9) must be corrected for the obscuration. The equation relating irradiance on the focal plane to the incident radiance can be written as

$$E = \frac{L\tau_l\tau_s\pi d^2}{(d^2 + 4f^2)} \tag{5.15}$$

where τ_l is the transmission loss due to less-than-perfect reflection or transmission by all of the optical elements, τ_s is the transmission loss due to obscuration by the secondary, d is the diameter of the entrance aperture (usually the diameter of the primary), and f is the overall system focal length. If we assume that the spider web holding the secondary in place has negligible size (i.e., it doesn't add to the obscuration), then we can approximate the obscuration loss as

$$\tau_s = 1 - \frac{\frac{\pi d_s^2}{4}}{\frac{\pi d^2}{4}} = 1 - \frac{d_s^2}{d^2} \tag{5.16}$$

where d_s is the diameter of the secondary. The $G\#$ can then be redefined to be approximately

$$G\# = \frac{1 + 4(f\#)^2}{\tau_l\tau_s\pi} \tag{5.17}$$

and used in the same way as we did for film-camera systems.

5.2.2 System Characterization

In many cases we will find it useful to describe EO imagers in terms of the components that make up the entire system. These components are outlined in Figure 5.9. The opto-mechanical elements are the optical elements that focus the image onto the focal plane and the mechanical elements that control what the focal plane "sees." This is the equivalent of lens, shutter, and film advance in a conventional camera system. The detectors in EO systems convert the flux incident on them into electronic signals in a fashion similar to the way film converts the flux into the latent image. The *preamplifier* is a critical component in an EO system that applies a fixed-magnitude, low-noise gain to the very low signals from the detector. This raises the signals up to where they can be more readily processed without concern about adding significant noise. The next stage of processing, called *signal conditioning*, is done in order for the output signal to span the range needed for ease of recording, analog-to-digital conversion, or transmission to a ground station. Several alternative sequences are shown in Figure 5.9 for the route the conditioned signals may follow. Any one of these, or even more than one sequence, may be followed, depending on the type of system. Nearly all signals today are converted to digital form at some point in the processing. For example, a signal from 0 to 1 V will be converted in equal steps to span from 0 to 255 (2^8) digital counts by an 8-bit system, or signals from -5 to +5 V may be converted to digital values from 0 to 1023 (2^{10}) by a 10-bit system (cf. Chap. 8 for a more complete treatment of digitizing issues). Data recording may be analog or digital depending on the system. However, improvements in the speed, capacity, and cost of digital systems have

moved most systems to digital recording. Similarly, most data are transmitted in digital form. The reason for this is that with error correction routines, the likelihood of corruption of digital data during recording, playback, or data transmission is very low. Digital processing/transmission also allows for the use of digital compression techniques.

The digital signals are then preprocessed to reconstruct the image. This may involve correcting for sensor geometric distortion, registration of different spectral bands, noise suppression, digital signal conditioning, and radiometric corrections. This entire process from the preamplifier to the end of the digital preprocessing chain is the equivalent of the chemical processing of the latent film image to a developed image. The final digital EO image may be displayed on a CRT (soft copy) or written out to a film image (hard copy). The digital EO image will most likely

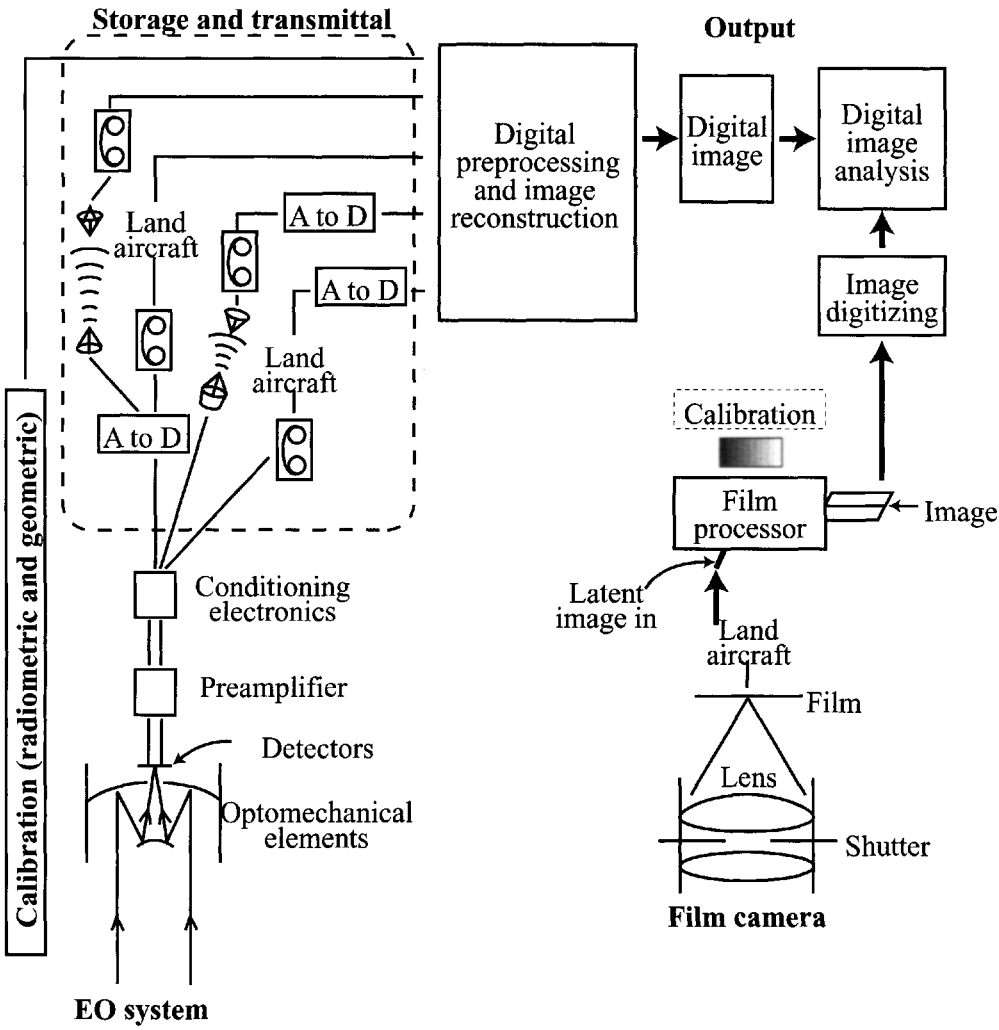


Figure 5.9 Components of an EO imaging system compared to film cameras.

be sent to a digital image processing system for further image analysis and information extraction. For convenience, we are treating the digital image processing system as separate from the sensor system (cf. Chap. 8), although at times the data analysis systems are packaged as part of the sensor system.

Operating in parallel with, and sometimes through, the opto-mechanical and signal processing electronics are the calibration elements of the sensing system. These are typically sources of known radiance for radiometric calibration and orientation and timing data needed for geometric reconstruction of the data and projection into ground coordinates. For the film-camera systems, the radiometric calibration elements would be sensitometric control wedges. Global positioning and gyroscopic data would provide the x, y, z location of the sensor in global coordinates and the camera line-of-site angles for projection of the camera image to ground coordinates. In the discussions that follow, we will generally treat only the fundamental elements of the opto-mechanical systems, the detector characteristics, and relevant calibration data. For brevity's sake, the electronics and digital system characteristics are generally lumped together with only a few critical parameters addressed. A much more complete treatment of detector electronics, amplification, and signal-to-noise issues is contained in Rogatto (1993).

5.3 DETECTORS AND SENSOR PERFORMANCE SPECIFICATIONS

This section will introduce the basic terminology of *detectors* with an emphasis on providing some simple concepts for making rough calculations of the performance characteristics of detectors and detector-noise-limited systems. A more rigorous treatment, including discussion of noise types, amplifier noise, and details of various detector materials, can be found in Budde (1983). Regrettably, a proper understanding of this topic requires nearly a book of its own, so we must restrict our discussion to operational terminology, with the underlying semiconductor physics and noise theory left to the reader's own inquiry.

5.3.1 Detector Types

There are three fundamentally different types of EO detectors we wish to consider. They can be divided into the following categories:

A. *Thermal detectors*: These detectors absorb incident flux and undergo a temperature change. They have a high rate of change in electrical resistance with changes in temperature. As a result, when externally biased they can be made to exhibit a change in voltage across a reference resistor corresponding to changes in incident flux. (Examples—bolometer: blackened metal flake with a high rate of change of resistance with temperature; thermistor: blackened thermally sensitive semiconductor with properties similar to the bolometer.)

B. *External photo-effect detectors*: The photosensitive materials used in these detectors have sufficiently low work functions that incident photons with adequate energy can free electrons from the surface of the material and

produce a current in an external circuit. (Example—photomultiplier tube (PMT): electrons are emitted by the photoelectric effect at the cathode of a vacuum tube and are accelerated by an external voltage onto a metal surface. The collision of the electrons with the metal produces more free electrons, which are accelerated to a second metal surface where the process is repeated. This is continued through several stages until the multiplied electrons reach the anode and produce a current in an external circuit.)

C. *Internal photo-effect detectors*: These detectors are semiconductors in which the electrons undergo internal energy level transitions when they absorb a photon. Two types of interaction are of interest:

1. *Photoconductive detectors*: In these detectors the photon is absorbed by an electron in the valence band and excited to the conduction band of the semiconductor, where it can be observed as a change in the resistance. This can be monitored by changes in the current induced through the detector using an externally supplied voltage. (Examples—cadmium sulfide, CdS, and indium antimonide, InSb).

2. *Photovoltaic detectors*: These detectors take advantage of the internal potential difference that can develop at the junction between dissimilar materials in a semiconductor. Photons incident on this junction produce charge carriers that migrate under the internal bias and produce a voltage difference in an external circuit. (Examples—Silicon, Si, photo diode, mercury cadmium telluride, HgCdTe).

The last two general categories are both referred to as *photon detectors* since discrete interactions of photons with electrons cause the observed electrical signals.

In some cases, some of the performance characteristics of a photovoltaic detector can be enhanced by operating it in a photoconductive mode. In this case, an externally supplied reverse bias voltage is applied to the junction. Incident photons still release charge carriers; however, they travel under the influence of the induced potential in the opposite direction, producing a current that can be observed in the external circuit.

Thermal detectors have the attractive feature of being nearly uniformly sensitive over all wavelengths, with the spectral sensitivity largely governed by the method used to blacken the surface. On the other hand, because they must undergo an actual temperature change to produce a measurable signal, they tend to be somewhat slow (response times of 10^{-3} sec are typical), and as a result, they traditionally were not used extensively in imaging devices. However, their simplicity and stability make them attractive for reference instruments, and they are widely used in radiometers for laboratory and ground truth studies in the MWIR and LWIR. Figure 5.10 shows an example using a thermal detector in a radiometer with a thermally controlled reference cavity and a reflective chopper. The detector first "sees" the world and then the chopper. The chopper is a nearly perfect diffuse reflector, so

that the detector is exposed to a reference flux level due only to the temperature of the reference cavity. The output signal is proportional to the ratio of the unchopped to the chopped signal.

Advances in semiconductor design and fabrication are improving the time response of thermal detectors. An example is the silicon microbolometer design shown in Figure 5.11. In this design, the tiny blackened detector element (e.g., tens of micrometer on a side) is suspended over a cavity approximately one-quarter wavelength deep by a support/readout structure. The entire assembly is in a vacuum to prevent convective heat exchange. Incident photons are absorbed by the blackened detector element and change the detector temperature, thereby changing the conductive properties, which can be sensed by the readout electronics. Because the detector element is so thin (to reduce its thermal inertia), some of the photon flux is transmitted and reflected from the unblackened substructure. The reflected wave is out of phase with the incident wave, promoting destructive interference and hence

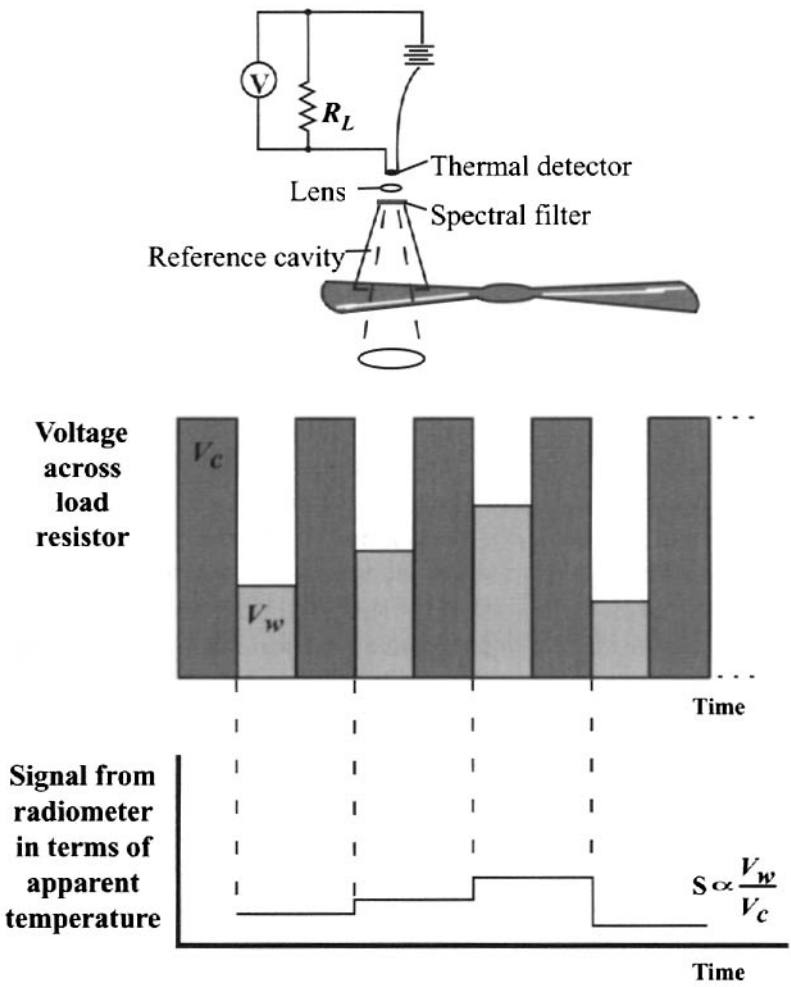


Figure 5.10 Radiometer schematic using thermal detector.

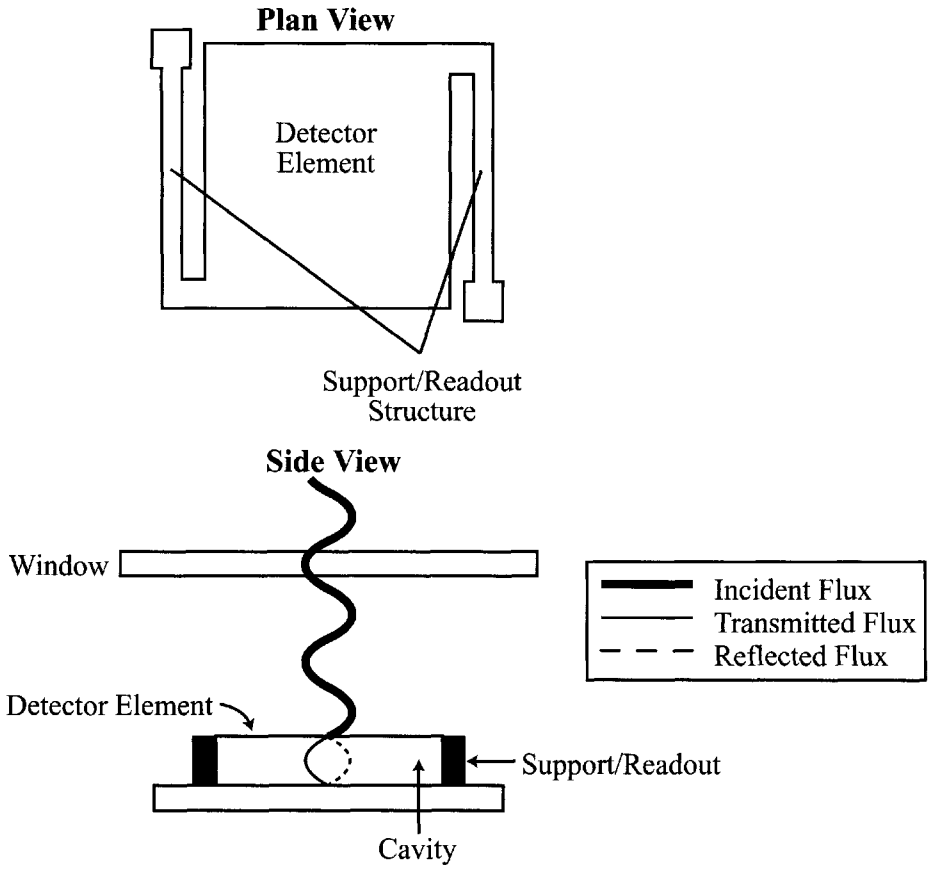


Figure 5.11 Illustration of an uncooled silicon bolometer thermal detector.

absorption at the detector element of any transmitted flux. The small size and thermal isolation of designs such as this are generating thermal detectors with response times and noise characteristics suitable for some imaging applications. These new thermal sensor designs (cf. Kruse 2001) are particularly intriguing because unlike many of the photon detectors described below for use in the thermal infared, the thermal detectors do not need to be cooled to cryogenic temperatures.

Unlike thermal detectors, photon detectors are sensitive to the amount of energy associated with each incident photon. In particular, if the amount of energy per photon (hv) does not exceed the work function for a PMT or the bandgap energy for the internal photo-effect electrons, then no signal will be produced. Since the energy needed to free an electron from the conduction band (the work function) is quite large, the external photoelectric effect (PMTs) can only be used for high-energy (short-wavelength) flux. The use of PMTs is therefore restricted to the VIS-NIR region, where their multiplicative effect makes them attractive for sensing low signal levels. However, the large size and high voltage requirements of the PMTs restrict their use to designs where only a limited number of discrete detectors are required.

As a result, the internal photo-effect detectors have become the most popular class of detectors in modern imaging systems. By changing material types, the entire spectrum from 0.2 to 20 μm can be sampled using these detectors. Advances

in semiconductor-based technology are providing a host of small, highly sensitive devices capable of meeting most design requirements. Indeed, as we will see in the remainder of this section, the advances in detector technology, coupled with the ability to mate the detectors to semiconductor electronic circuits, have opened the door to a revolution of new sensor designs (cf. in particular Sec. 6.2 on imaging spectrometers). To talk more specifically about the various semiconductor detectors, we need to introduce some terms for characterizing detectors.

5.3.2 Detector Figures of Merit

A number of terms have been developed over the years to describe the performance characteristics of detectors. These *figures of merit* are used to perform tradeoff studies between detector types, materials, and manufacturers and to evaluate what the expected performance of a system will be when a particular detector is employed. Table 5.1 lists several of the detector parameters and figures of merit we will use to characterize detectors. This is only a small sampling of the many terms EO engineers and detector manufacturers have developed to characterize specific performance features of detectors (and to confuse the innocent). The reader is referred to Rogatto (1993) for a more complete treatment.

Referring to Table 5.1, we have already introduced the wavelength-dependent responsivity, $R(\lambda)$, to be the signal S produced by the detector in volts or amps per unit incident flux, Φ, at wavelength, λ. Similarly, recall that the total effective *responsivity* is the integral of the wavelength-dependent responsivity weighted by the incident spectral flux according to

$$R = \int_0^\infty R(\lambda) \frac{\Phi_\lambda}{\int_0^\infty \Phi_\lambda d\lambda} d\lambda \tag{5.18}$$

making the responsivity a function of the spectral shape of the incident flux.

The signal out of the detector can then be expressed as

$$S[\mathrm{V}] = \int R(\lambda)\Phi_\lambda d\lambda = R\Phi[\mathrm{VW}^{-1}][\mathrm{W}] \tag{5.19}$$

as long as the total flux (Φ) has the same spectral distribution as was used in computing the total effective responsivity (R) in Eq. (5.18). This process of weighting by the source spectral shape can be used to convert any of the spectral terms in Table 5.1 to their total effective values.

Due to thermal variation in the detector (Johnson noise), random occurrence of photon events (shot noise), variation in the thermal exchange with the surround (temperature noise), and the random variation in signal with input frequency, there will be random variations in the signal level even when the detector is exposed to a constant flux level. These variations about the mean signal level are referred to as *noise* (N) and are usually characterized by the root mean square (RMS) variation in the instantaneous signal level (S_i) according to

Table 5.1 Detector Terminology for Figures of Merit

Term	Symbol	Definition	Units	Comments
Signal	S	Output of Detector	[A or V]	
Spectral responsivity	$R(\lambda)$	$R(\lambda) = \frac{dS}{d\Phi}$	$\frac{[\text{A or V}]}{\text{W}}$	
Noise	N	$\left[\frac{\sum_{i=1}^{n}(S_i - S_{avg})^2}{n}\right]^{\frac{1}{2}}$	[A or V]	RMS deviation in signal at fixed input (often zero flux)
Signal-to-Noise ratio		S/N	units	
Spectral Noise equivalent power	NEP(λ)	$\frac{N}{R(\lambda)}$	[W]	Incident flux equivalent needed to yield a signal-to-noise ratio of 1 (dark)
Spectral detectivity (λ)	$D(\lambda)$	$\frac{1}{\text{NEP}(\lambda)}$	[W^{-1}]	
D "star" or specific detectivity	$D^*(\lambda)$	$(Af)^{1/2}D(\lambda)$	[W^{-1}cm $Hz^{1/2}$]	A is detector area; f is the electronic bandwidth of the sensor
Noise equivalent temperature increment	NEΔT			The change in temperature of a blackbody needed to produce S/N of unity (i.e., to produce a power level equal to that of the NEP)

$$N_{\text{RMS}} = \left[\frac{\sum_{i=1}^{n}(S_i - S_{\text{avg}})^2}{n}\right]^{1/2} \quad [\text{V}] \tag{5.20}$$

when S_{avg} is the mean signal level and n is the number of samples. In many cases, the noise level will be a function of the signal level, so it is necessary to define noise relative to some flux level. A simple way to do this is to define the noise when no flux is incident on the detector (*dark noise*). However, for many systems, even the minimum flux levels are well above the dark level, so this noise is not indicative of what will be experienced operationally. For this reason, noise levels are often specified relative to some incident flux level. For example, the noise might

be defined as the RMS variation about the signal that would be generated by a flux onto the detector corresponding to the sensor viewing a 5% reflector at the top of the atmosphere for some set of solar conditions. In the thermal region, it could be the noise about the signal associated with viewing a 300 K blackbody.

While noise is a measure of the quality of a signal and less noise is better, noise really only takes on meaning when viewed relative to the corresponding signal as expressed by the *signal-to-noise* ratio (*S/N*). For example, a system with a signal-to-noise ratio of 20 for signals corresponding to a 10% reflector would have approximately 10 discernible levels in signal (2*N*) between 0 and the level of a 10% reflector (i.e., steps in reflectance of less than one reflectance unit would be difficult to separate from the noise).

In many cases, it is useful to express the concept of noise in radiometric input units [W] rather than in output signal units [*V*]. This can be accomplished using the wavelength-dependent *noise-equivalent power*, NEP(λ), which is the amount of incremental flux at wavelength λ required to change the signal level by an amount equal to the noise, i.e.,

$$\mathrm{NEP}(\lambda) = \frac{N}{R(\lambda)}[\mathrm{W}] \tag{5.21}$$

This term tells us that flux levels or variations in flux must be above the NEP level to have any chance of being detected. The *detectivity* term, $D(\lambda)$, is simply 1/NEP(λ) and has the dubious value of increasing with the quality of the detector. Of greater interest is the wavelength-dependent *specific detectivity*, $D^*(\lambda)$. This term makes performance comparison between detectors more valid by adjusting the detectivity by its correlation with $(Af)^{1/2}$ according to

$$D^*(\lambda) = D(\lambda)\sqrt{Af} = \frac{\sqrt{Af}}{\mathrm{NEP}(\lambda)}[\mathrm{W}^{-1}\mathrm{cmHz}^{1/2}] \tag{5.22}$$

where A is the area of the detector and f is the signal bandwidth. This specific dependence on temporal frequency is useful in that it reminds us that all the figures of merit are dependent on the time frequency or bandwidth of the incident signal. It is also important to realize that the response characteristics of detectors are dependent on the operating temperature of the detector. A small sampling of detector types and variation in D^* values with wavelength, temperature, and bandwidth are shown in Figures 5.12–5.15. Note in particular the poor temporal response of the bulk thermistor shown in Figure 5.12 and the improved performance of the InSb detector at cryogenic temperatures.

To this point, we have been implicitly talking about discrete detectors. Increasingly, linear array and two-dimensional array detectors are being used for imaging. These arrays can be thought of as a row of discrete detectors that accumulate charge (signal) and transfer the signal to a charge-coupled device (CCD). The CCD is an electronic device that can hold charge packages in discrete elements and then transfer the charge in a bucket brigade fashion through the elements and into a conventional electronic circuit at the end. It is also possible to make CCDs that are themselves photosensitive. This is particularly attractive in the VNIR region

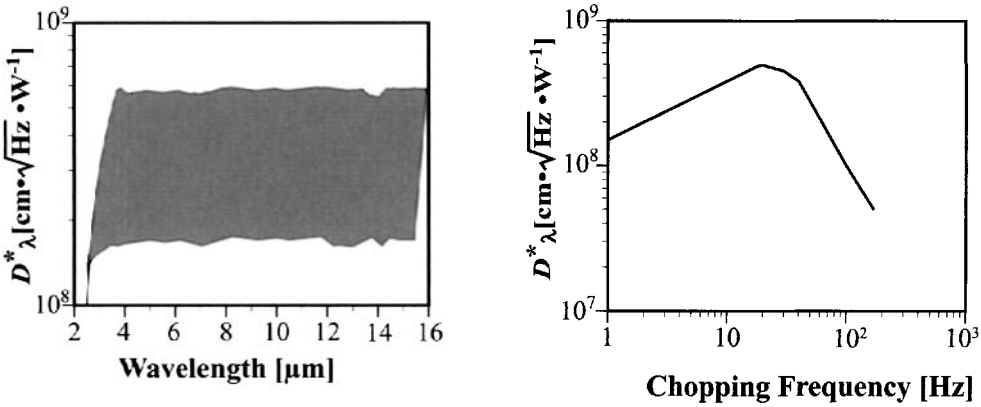


Figure 5.12 Sample performance characteristics for a thermistor at 300 K. Shaded region shows typical range.

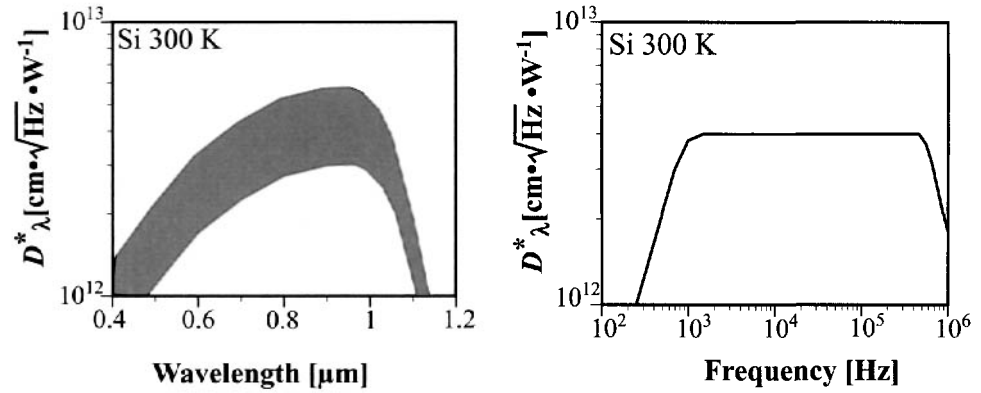


Figure 5.13 Sample performance characteristics for a silicon detector at 300 K.

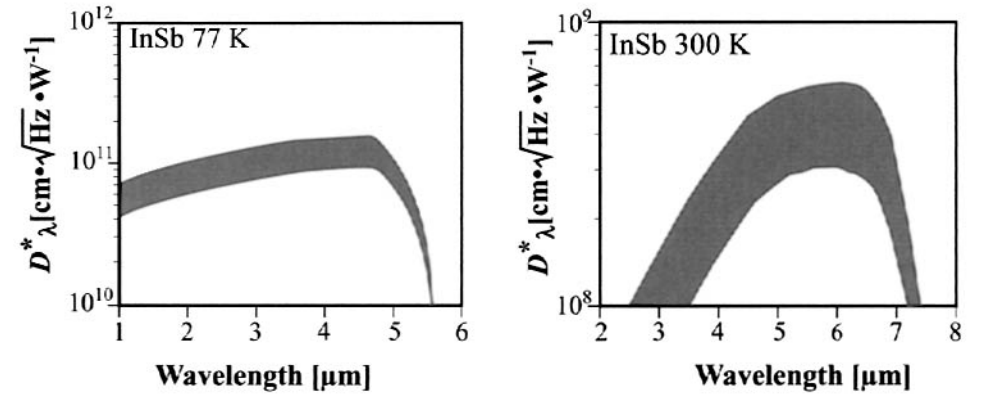


Figure 5.14 Sample performance characteristics for an InSb detector, photovoltaic mode at 77 K and photoconductive mode at 300 K. (Adapted from Wolfe and Zissis 1985.)

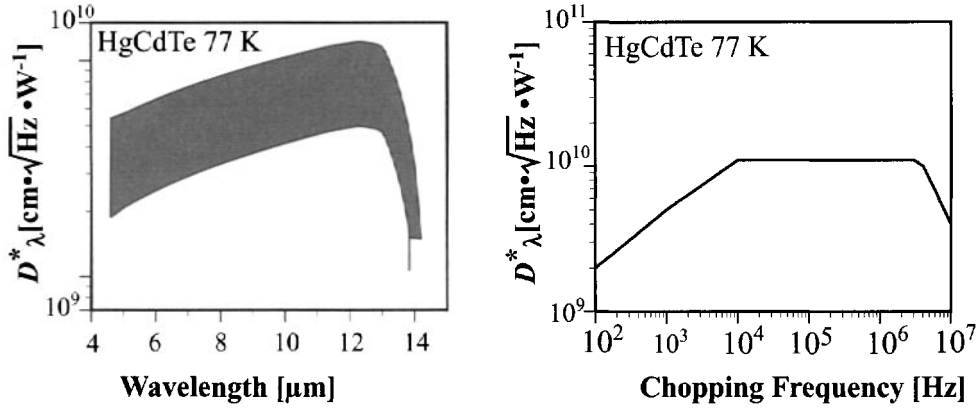


Figure 5.15 Sample performance of an HgCdTe detector operator at 77 K. (Adapted from Wolfe and Zissis 1985.)

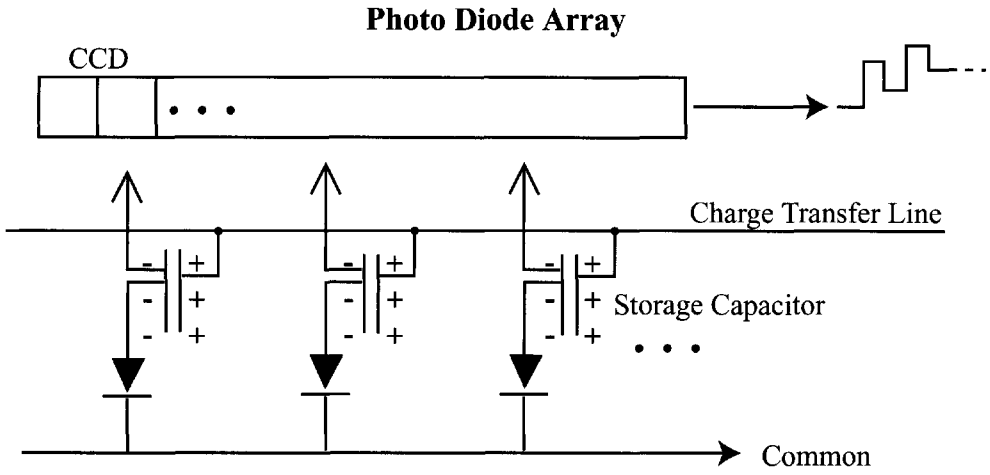
where silicon is photosensitive, since most solid-state electronics are silicon based. Figure 5.16 shows a linear array using photo diodes as the sensing elements and a CCD as an electronic charge transfer element, as well as a linear array where the CCD is used directly as the sensor. In either case, either through a charge transfer or through the internal photo effect, charges are produced in a conductive layer of the material and kept in the conductive layer by an insulating layer. Discrete elements are produced by forming localized potential wells that inhibit migration of charge from where it is introduced or formed. When the signal is ready for processing, electrodes on the opposite side of the insulating layer from the conductor are set to a positive level relative to the charges, causing a migration of the charge toward the electrode. By properly sequencing the voltage pulse trains to the electrodes, the charges can be passed from element to element to the end of the CCD and into a circuit. In a two-dimensional array, the signal is usually read out a line at a time.

Array detectors offer the great advantage in imaging and spectrometry of being able to collect many lines or spectral levels simultaneously. This lets the individual detector element collect data longer, increasing the dwell time and improving the signal-to-noise ratio. These detectors have some additional noise due to inefficiencies in the charge transfer process. However, in most applications this is not sufficiently large to reduce their utility.

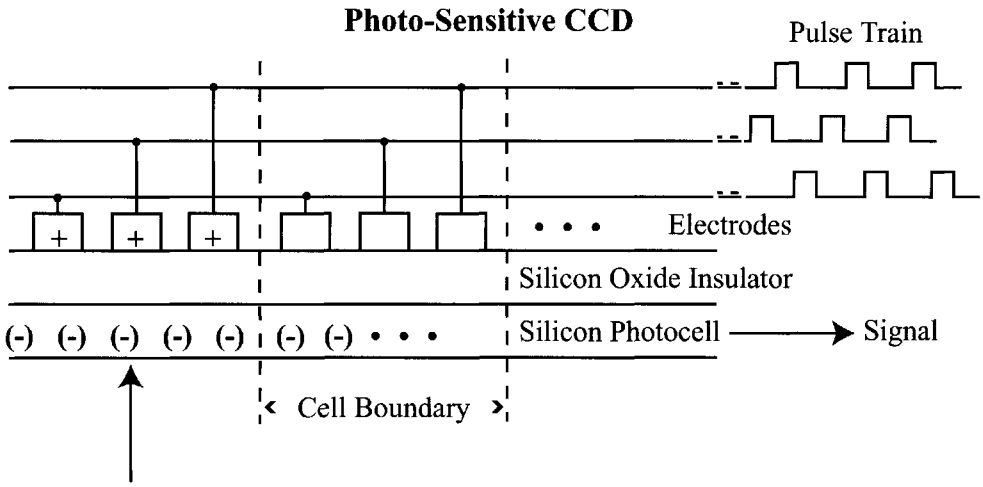
Dereniak and Crowe (1984) suggest that the use of figures of merit employing photon energy and electronic charge quanta are more useful for many detectors than the figures of merit discussed thus far. Many CCD arrays have their noise levels expressed in terms of numbers of *noise electrons*. In terms of the units we have introduced, we can express the signal (S) out of the detector-preamplifier in terms of quantum units as

$$S[\mathrm{V}] = \Phi_p t_{\mathrm{int}} \mathrm{QE} e \mathrm{CE} \left[\frac{\text{photons}}{\text{sec}} \sec \frac{\text{electrons}}{\text{photon}} \frac{\text{coul}}{\text{electron}} \frac{\text{V}}{\text{coul}} \right] \tag{5.23}$$

where Φ_p [photons/sec] is the incident flux expressed in terms of numbers of photons per second, t_{int} [sec] is the *integration time* of the detector element, QE is the



(a) Diode array using CCD only for charge transfer.



(b) CCD used as the photo-sensitive device.

Figure 5.16 Linear array concepts.

quantum efficiency of the detector defined as the average number of free electrons produced per incident photon [electrons/photon], e [coul per electron] is the charge on an electron, and CE is the *conversion efficiency* of the preamplifier [V/coul] which defines how many volts are produced per unit of charge. We can then define the photon responsivity in terms of photon flux as

$$R_p = \frac{S}{\Phi_p} = t_{\text{int}} \text{QE} e \text{CE} \left[\frac{\text{V sec}}{\text{photon}} \right] \tag{5.24}$$

and the noise equivalent photon power as

$$\text{NEP}_p[\text{photons sec}^{-1}] = \frac{N}{R_p} \tag{5.25}$$

where N [V] is the detector-preamplifier noise, including charge transfer noise. The NEP_p is interpreted in the same way as NEP, except that we are now in units of photons per second required to produce a signal equal to the noise. Dereniak and Crowe (1984) point out that the responsivity expressions in Eqs. (5.24) and (5.18) are for spatially uniform signals and must be modified for spatially varying signals by the spatial frequency response of the sensor (cf. Sec. 13.2).

5.3.3 Sensor Performance Parameters

Detectors are a critical element in a sensor's performance, and many sensors are *detector limited* in terms of their noise limitations (i.e., the detector is often the weakest link in the radiometric portion of the image chain in that it is the largest source of noise). However, there are many other factors that may be the weak link in the detector-electronics signal processing chain.

In some systems, the overall sensor performance may not be limited by the detector's performance but by noise in the preamplifier electronics or detector readout noise. The preamplifier components generate most of the same types of noise as the detector, and since the incoming signal is small, the preamplifier noise can become the limiting factor in sensor performance. From the sensor user's point of view, the source of the limiting noise is not critical, only its overall effect on the system performance. It is often useful, therefore, to specify system performance characteristics (i.e., rather than detector performance characteristics) such as the sensor NEP. From the system NEP and knowing the optical throughput ($G\#$), the *noise-equivalent radiance* (NER) of the sensor can be computed according to

$$\text{NER}(\lambda) = \frac{\text{NEP}(\lambda)}{A_d} G\# = \text{NEI}\, G\# \tag{5.26}$$

where NER is the amount of radiance or change in radiance on the front of the sensor required to produce a change in sensor output equal to the sensor's noise level, NEI is the *noise-equivalent irradiance*, NEP is the sensor noise equivalent power (equal to the detector NEP only if the system is detector noise limited), and A_d is the area of the detector.

From the noise equivalent radiance, the *noise-equivalent change in reflectance* (NE$\Delta\rho$) or the *noise-equivalent change in temperature* (NEΔT) can be computed from

$$\text{NE}\Delta\rho = \text{NER} \cdot \frac{\Delta\rho}{\Delta L} \tag{5.27}$$

or

$$\text{NE}\Delta T = \text{NER} \cdot \frac{\Delta T}{\Delta L} \tag{5.28}$$

where $(\Delta\rho/\Delta L)^{-1}$ represents the rate of change in radiance at the sensor corresponding to a unit change in reflectance (this is often expressed in terms of a top of the atmosphere value at set solar conditions), $(\Delta T/\Delta L)^{-1}$ represents the rate of change in radiance at the front of the sensor corresponding to a unit change in the temperature of a blackbody in front of the sensor, and all the terms on the right-hand sides of Eqs. (5.27) and (5.28) are the effective values in the sensor bandpass [cf. Eq. (5.18)]. These are typically the most meaningful terms for the sensor designer and user, since they provide a quick assessment of the radiometric performance in intuitive units. Recognize, however, that these values must still be adjusted for atmospheric effects to predict expected operational performance.

Two final factors need to be considered in evaluating noise-limited performance of imaging sensors. In some cases, the recorder, transmitter, or playback electronics may have worse noise specifications than the detector-preamplifier and become a limiting factor. Finally, in many systems the analog-to-digital converter that quantizes the signal may be the limiting factor. End-to-end performance assessment of the radiometric strand of the image chain is treated in greater detail in Section 13.2.

5.4 DETECTOR-SENSOR PERFORMANCE CALCULATIONS

A full treatment of detector performance issues calls for a text of its own. Indeed, Dereniak and Crow (1984) and Budde (1983) do just that. In this section, we will simply look at some simple, often best-case, scenarios to expose the reader to some of the issues associated with detector performance calculations. To begin with, let's consider the performance we might expect from a relatively high-spectral-resolution system. We will focus on one band located near the peak of the solar radiation spectrum that takes advantage of high-performance silicon detector array technology. The hypothetical instrument under consideration looks like the line scanning types of imaging spectrometers introduced in Chapter 6 (e.g., AVIRIS, MISI). The specifications for the instrument (cf. Fig. 5.17) and the spectral channel of interest (centered at 0.5 μm) are listed in Table 5.2. In this exercise, we will compute the upper limitation to the performance of such a system (i.e., with a perfect detector) based on the noise limitations induced by the variation in photon flux sampled over a discrete time interval (i.e., *shot noise*). The number of photons emitted and/or detected in a finite interval can be modeled with a Poisson distribution when the numbers are large. The standard deviation of the number of discrete events occurring in a sample window for samples with a Poisson distribution is simply the square root of the number of samples. Thus, if a detector is expected to observe 10^4 photons in a discrete time interval, then the one standard deviation error in the estimate due to photon or shot noise would be $\sqrt{10^4} = 100$ photons. In practice, the samples we actually observe are the electrons generated at the detectors, and the noise of interest is the variation in the number of electrons per discrete time sample. Since the signal due just to shot noise will always be the noise squared, we can generate a best case for the signal to noise of

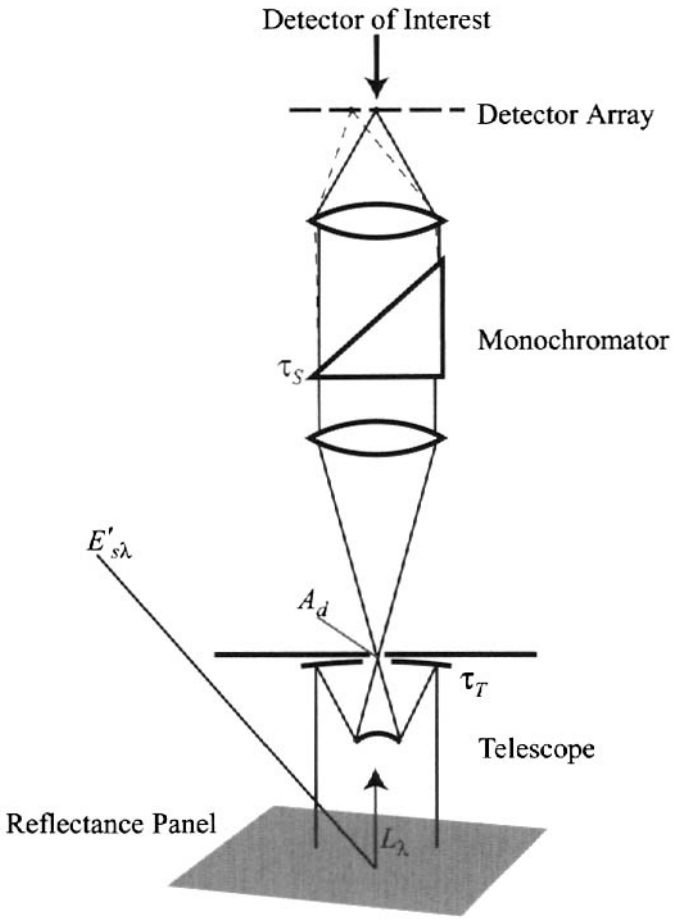


Figure 5.17 Illustration of a hypothetical sensor used to collect spectral data with the specifications listed in Table 5.2.

Table 5.2 Key parameters used in sample performance calculations for a reflective band sensor.

Central wavelength	0.5	λ_{test}
Spectral band width	0.01 μm (10 nanometers)	$\Delta\lambda$
System *F#*	2.6	*F#*
Effective detector size	250 μm × 250 μm	A_d
Telescope transmission	0.82	τ_T
Spectrometer transmission	0.7	τ_S
Quantum efficiency	0.95 electrons/photon	QE
Integration time	2.5×10^{-5} sec	t_{int}
Exoatmospheric irradiance	1828 W/m^2 μm	E_s
Solar zenith angle	40°	σ

$$S/N = \frac{S}{\sqrt{S}} = \sqrt{S} \tag{5.29}$$

In practice, there will be additional electronic noise sources that will often dominate over the signal-dependent photon noise. However, since in general there is no way around the shot noise, we can treat it as a best-case scenario.

To compute the shot noise, we need to calculate the number of signal-dependent electrons generated by an observation. Considering the instrument described in Table 5.2, we will compute the number of signal-related electrons produced by an 18% Lambertian reflector above the Earth's atmosphere. The reflected spectral radiance can be expressed as

$$L_\lambda = \frac{E'_{s\lambda} r \cos\sigma}{\pi} \left[\mathrm{Wm^{-2}sr^{-1}\mu m^{-1}}\right] \tag{5.30}$$

The spectral irradiance reaching the sensor is then

$$E_\lambda = \frac{L_\lambda}{G\#} = \frac{E'_{s\lambda} r \cos\sigma}{\pi G\#} \left[\mathrm{Wm^{-2}\mu m^{-1}}\right] \tag{5.31}$$

where the $G\#$ for this system is

$$G\# = \frac{1+4F\#^2}{\tau_T \tau_s \pi} = \frac{1+4\cdot(2.6)^2}{0.82\cdot 0.7\pi} = 15.5[sr^{-1}] \tag{5.32}$$

where the system as specified has an $F\#$ of 2.6; the telescope transmission, including obscuration and reflection from mirrors, is 0.82; and the spectrometer used to provide spectral separation causes another loss of 70% over the spectral band of interest. The spectral flux on the detector can then be expressed as

$$\Phi_\lambda = E_\lambda A_d = \frac{E'_{s\lambda} r \cos\sigma A_d}{\pi G\#} [\mathrm{W\mu m^{-1}}] \tag{5.33}$$

where A_d is the effective area of the detector. In this case, A_d would be the entrance aperture to a monochromator, and 70% (τ_s) of the flux in the spectral band of interest would be focused onto the detector element corresponding to the band of interest. If the monochromator used 1-to-1 optics, the actual detector area could also be A_d. The flux over a narrow spectral interval $\Delta\lambda$ can then be expressed as

$$\Phi = \Phi_\lambda \Delta\lambda = \frac{E'_{s\lambda} r \cos\sigma A_d \Delta\lambda}{\pi G\#} [\mathrm{W}] \tag{5.34}$$

where we have assumed that $\Delta\lambda$ is narrow enough that the wavelength-dependent values in Eq. (5.34) are approximately constant over the interval (when this is not true, numerical integration is required). The photon flux can now be expressed as

$$\Phi_p = \frac{\Phi}{hc/\lambda} = \frac{E'_{s\lambda} r \cos\sigma A_d \Delta\lambda\lambda}{\pi G\# hc} \left[\frac{\mathrm{W{=}joules/sec}}{\mathrm{joules/photon}} = \frac{\mathrm{photons}}{\mathrm{sec}}\right] \tag{5.35}$$

The signal, expressed as numbers of photons, for a sensor integrating over a time t_{int} is then

$$S_p = \Phi_p t_{\text{int}} = \frac{E'_{s\lambda} r \cos\sigma A_d \Delta\lambda\lambda t_{\text{int}}}{\pi G\#hc}[\text{photons}] \tag{5.36}$$

This yields the following expression for the number of electrons generated:

$$S_e = S_p QE = \frac{E'_{s\lambda} r \cos\sigma A_d \Delta\lambda\lambda t_{\text{int}} QE}{\pi G\#hc}\left[\text{photons}\frac{\text{electrons}}{\text{photon}}\right] \tag{5.37}$$

For our example, we will assume that all the spectral terms are slowly varying over the 10 nm spectral bandwidth of the spectral channel of interest so that we can use mean values for $E'_{s\lambda}$,r(λ),λ,QE(λ), and G#(λ). If this were not the case, or for more precise calculation, numerical integration can be performed. Using this simplified assumption, the number of signal electrons expected for an 18% reflector would be

$$\begin{aligned} S_e &= \frac{1828\text{Wm}^2\mu\text{m}^{-1}0.18\cos(40°)\left(250\cdot10^{-6}\text{m}\right)^2 0.01\mu\text{m}0.5\cdot10^{-6}\text{m}2.5\cdot10^{-6}\text{sec}0.95\frac{\text{electrons}}{\text{photon}}}{\pi\text{sr}15.5\text{sr}^{-1}6.62\cdot10^{-34}\text{joulessec}3\cdot10^{8}\text{msec}\cdot1/\text{photons}} \\ &= 1.93\cdot10^{5}[\text{electrons}] \end{aligned} \tag{5.38}$$

If the number of electrons generated by the discrete sampling of photons follows a Poisson distribution, then the photon arrival rate or shot noise (N_Φ) can be expressed as

$$N_\Phi = S_e^{\frac{1}{2}} = 439\text{electrons} \tag{5.39}$$

which would yield a best-case (i.e., signal limited) signal-to-noise ratio for this example of

$$S/N_\Phi = \frac{S_e}{N_\Phi} = 4.4\cdot10^2 \tag{5.40}$$

In practice, Dereniak and Crowe (1984) point out that the detector and readout noise for such a system might be of the order of 700–1200 electrons (note that actual electronic noise levels are very design and device specific). If a more detailed analysis showed that our specific system had an aggregate electronic noise N_e (detector and readout) of 720 electrons, then the total noise can be estimated as

$$N_{tot} = \left(N_\Phi^2 + N_e^2\right)^{\frac{1}{2}} = 843 \tag{5.41}$$

for an actual signal to noise ratio of

$$S/N = 230 \tag{5.42}$$

If we assume that the overall instrument noise is dominated by the detector (a good assumption for array detectors in the VNIR), then we can estimate expected performance in apparent or top of the atmosphere reflectance units in the following manner. The change in signal corresponding to a one-unit change in reflectance

can be easily calculated by taking advantage of our calculations for an 18% reflector. The signal from a 19% reflector (S_{e19}) would be

$$S_{e19} = S_{e18}\frac{0.19}{0.18} = 2.037 \cdot 10^5 [\text{electrons}] \tag{5.43}$$

and the change in signal per unit change in reflectance is

$$\frac{\Delta S}{\Delta \rho} = \frac{S_{e19} - S_{e18}}{1} = \frac{2.037 \cdot 10^5 - 1.93 \cdot 10^5}{1} = 1.07 \cdot 10^4 [\text{electrons / reflectance unit}] \tag{5.44}$$

This yields a noise equivalent change in reflectance (for signals corresponding to an 18% apparent reflector) of

$$NE\Delta\rho = \frac{\frac{N_{tot}}{\Delta S}}{\Delta \rho} = \frac{843[\text{electrons}]}{1.07 \cdot 10^4 [\text{electrons/reflectance unit}]} = 0.08 \text{reflectance units} \tag{5.45}$$

which indicates that change from 18% to 18.1% should be just above the system noise level.

In the example above, we see the fairly typical case where noise is dominated by the additive or bias (i.e., signal-independent) noise from the electronics and our best-case estimate of signal-limited noise would be quite optimistic. On the other hand, for bright targets or more "photon-rich" designs, the signal noise can begin to dominate. For instance, if our target in the example above were a 60% reflector instead of the 18% reflector we used in the previous calculations, then the signal would increase to

$$S_{e60} = S_{e18}\frac{0.60}{0.18} = 6.43 \cdot 10^5 [\text{electrons}] \tag{5.46}$$

yielding photon (shot) noise of

$$N_{p60} = (S_{e60})^{\frac{1}{2}} = 802[\text{electrons}] \tag{5.47}$$

At this point, our signal-dependent noise and our bias noise are approximately equal, and the best-case estimates are more in keeping with actual performance. In this situation, the noise will have some signal-dependent (i.e., multiplicative) character. The reader should recognize that this example and the one that follows are designed to provide illustrations of the type of analysis required for each channel of a sensing system, not of common behavior. In practice, most remote sensing systems are photon starved for targets of interest and we will typically assume that bias noise dominates in the image processing algorithms discussed in Chapters 8–11.

We will now consider another case in the thermal region to evaluate other factors that come into play where signal-limited calculations can be used productively. In this case, we assume we have a photovoltaic detector with a 12 μm spectral cutoff and 10 μm long pass filter assembled in an imaging system as shown in Figure 5.18. We assume the detector coldstop and filter are maintained at 77K

and that the telescope and support structure are maintained at 297K. Note that the detector in this design employs a coldstop to limit its field of view to ±10° from the optical axis, so that ideally only flux coming from the scene (i.e., from the secondary mirror) is admitted. There will be a photon flux from the rest of the hemisphere above the detector, but these photons are limited in number by cooling the coldstop and the filter to the detector operating temperature of liquid nitrogen (77K).

The photon flux onto the detector will come primarily from three sources. The first source is the radiance from the scene in the bandpass of interest attenuated by the telescope assembly and the longpass filter. The second source is flux, in the bandpass of interest, coming from the telescope assembly itself. This is due to the nonzero emissivity of each mirrored surface ($\varepsilon = 1\text{-}r$) and the typically near-unit emissivity of the support structure that may be in the field of view of the detector (at a minimum, this usually includes the spider web supporting the secondary and may include small parts of the barrel to allow the primary to serve as the limiting aperture). The third source is from the cold cavity itself, including the filter and photons emitted at all wavelengths to which the detector is sensitive.

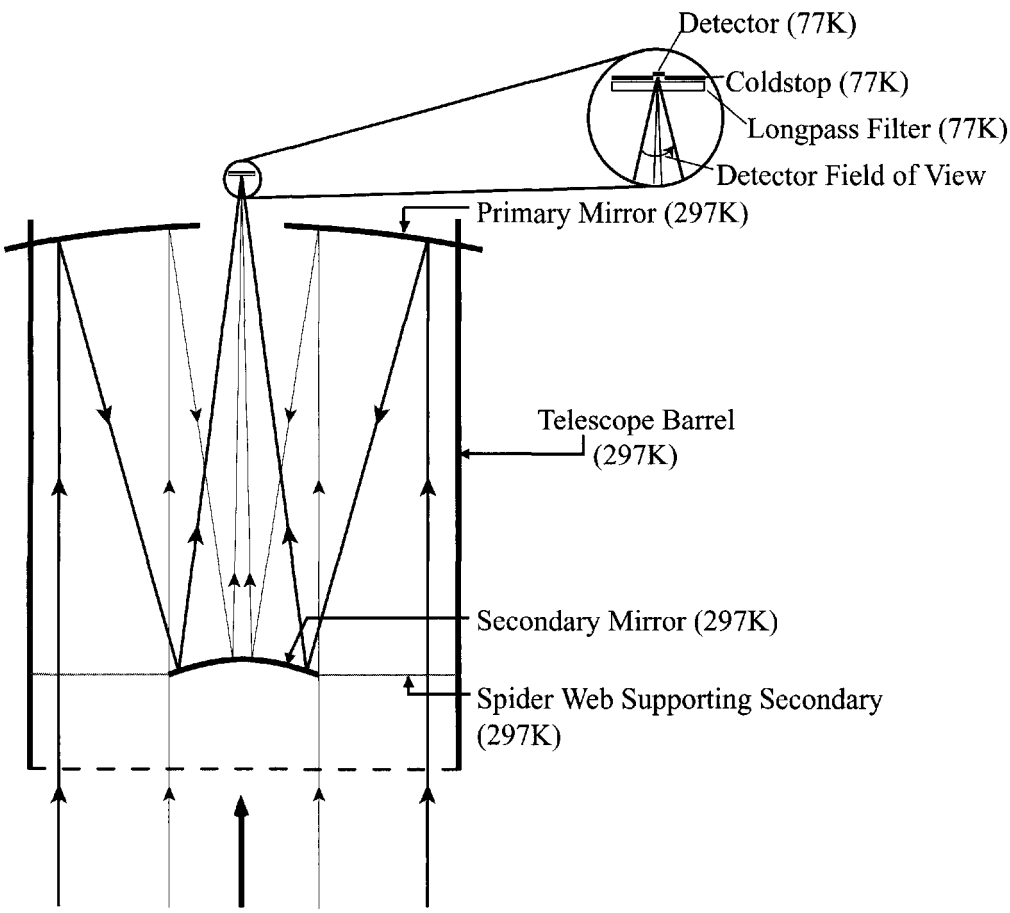


Figure 5.18 Illustration of a hypothetical sensor used to collect data with the specifications listed in Table 5.3.

Table 5.3 Key parameters used in sample noise calculations for a thermal infrared sensor.		
Bandpass of interest	10–12 μm	$\Delta\lambda$
System *F#*	4	*F#*
Detector size	100 μm × 100 μm	A_d
Telescope transmission	0.75	τ_T
Quantum efficiency	0.93	QE
Integration time	2.0×10^{-5} sec	t
Minimun target temperature	273	T_{min}
Detector temperature	77	T_D
Telescope temperature	297	T_T
Effective emissivity of telescope	0.3	ε_T
Filter transmission	0.89	τ_F
Detector cut off wave length	12 μm	λ_{max}
Filter cut on wavelength	10 μm	λ_{min}
Detector acceptance angle	±10°	$\theta/2$

Dereniak and Crow (1984) indicate that the RMS photon-limited noise for such a detector can be computed based on the Bose-Einstein statistics associated with blackbody radiation according to

$$N_p = \left[\frac{S_p e^{\frac{hc}{\lambda kT}}}{\left(e^{hc/\lambda kT} - 1\right)} \right]^{\frac{1}{2}} \tag{5.48}$$

However, for $\lambda < 14$ and $T < 320$ the Boson factor $(e^{hc/\lambda kT})/(e^{hc/\lambda kT}-1)$ can be set to 1 with only a few percent error. This yields an expression for the photon noise of

$$N_p \cong S_p^{\frac{1}{2}} [\text{photons}] \tag{5.49}$$

which is the Poisson approximation introduced earlier. This allows us to compute the best-case performance for a thermal infrared detector, assuming that photon statistical fluctuations dominate over electronic noise sources. The difference between thermal detectors and detectors in the reflective regime is that we need to account for the potentially large numbers of electrons generated by background radiation. To evaluate the magnitude of these sources for our example, we will once again compute the total number of electrons generated for the conditions listed in Table 5.3.

For a blackbody, the spectral flux on the detector can be expressed as

$$\Phi_\lambda = L_\lambda(T) G\#^{-1}(\lambda) A_d \quad [\text{W}\mu\text{m}^{-1}] \tag{5.50}$$

where L_λ is the Planckian spectral radiance for a blackbody at temperature T, A_d is the area of the detector, and the $G\#(\lambda)$ for these systems can be expressed as

$$G\#(\lambda) = \frac{1+4F\#^2}{\tau_T \tau_F \pi} = \frac{1+4\cdot 4^2}{0.75\cdot 0.89\cdot \pi} = 31\left[sr^{-1}\right] \tag{5.51}$$

where we have assumed constant values over the bandpass of interest. The flux over a relatively narrow spectral interval can be approximated as

$$\Phi_{\Delta\lambda} \cong \frac{L_{\lambda 2}(T) + L_{\lambda 1}(T)}{2} \cdot \frac{1}{G\#(\lambda)} \cdot A_d \Delta\lambda [\mathrm{W}] \tag{5.52}$$

where $\Delta\lambda = \lambda_2 - \lambda_1$ represents the spectral interval of interest. The photon flux in an interval can then be expressed as

$$\Phi_{p\Delta\lambda} = \frac{\Phi_{\Delta\lambda}\lambda}{hc} = \frac{[L_{\lambda 2}(T) + L_{\lambda 1}(T)]A_d \lambda \Delta\lambda}{2G\#(\lambda)hc}\left[\frac{\text{photons}}{\text{sec}}\right] \tag{5.53}$$

where for a small interval ($\Delta\lambda$) the mean wavelength over the interval can be used as an estimate of λ.

The total number of photons in a spectral interval can then be expressed as

$$S_{p\Delta\lambda} = \Phi_{p\Delta\lambda} t = \frac{[L_{\lambda 2}(T) + L_{\lambda 1}(T)]A_d \lambda \Delta\lambda t}{2G\#(\lambda)hc}[\text{photons}] \tag{5.54}$$

where t is the effective sample integration time, which can be approximated as

$$t = \frac{1}{2\Delta f}[\text{sec}] \tag{5.55}$$

where Δf is the noise bandwidth of the system.

The number of electrons in an interval is then expressed as

$$S_{e\Delta\lambda} = S_{p\Delta\lambda} QE(\lambda) = \frac{[L_{\lambda 2}(T) + L_{\lambda 1}(T)]A_d \lambda \Delta\lambda t QE(\lambda)}{2G\#(\lambda)hc}[\text{electrons}] \tag{5.56}$$

In practice, it is often convenient when performing numerical integrating over a bandpass to define the spectral photon radiance $L_{p\lambda}$ as

$$L_{p\lambda} = \frac{L_\lambda(T)\lambda}{hc}\left[\text{photons sec}^{-1} m^{-2} sr^{-1} \mu m^{-1}\right] \tag{5.57}$$

The signal generated by radiance from the scene over a bandpass can then be expressed as

$$S_e = \int \frac{L_{p\lambda}(T) A_d t Q E d\lambda}{G\#(\lambda)}[\text{electrons}] \tag{5.58}$$

For convenience, we will assume QE and $G\#$ are approximately constant with wavelength, so they can be removed from the integral, and we can use tabulated values to solve Eq. (5.58) according to

$$S_e = \frac{A_d tQE}{G\#}\left[\int_0^{\lambda_2} L_{p\lambda}(T)d\lambda - \int_0^{\lambda_1} L_{p\lambda}(T)d\lambda\right] \tag{5.59}$$

Appendix A includes solutions to the cumulative blackbody integrals used in Eq. (5.59) for a small range of wavelengths and selected temperatures of interest. Use of Appendix A for a target at 273K, $\lambda_2 = 12$ μm and $\lambda_1 = 10$ μm yields

$$S_e = \frac{\left(100\cdot 10^{-6}\text{m}\right)^2 2\cdot 10^{-5}\ \text{sec}\ 0.93\ \text{electrons/photon}}{31\ \text{sr}^{-1}}\left[1.53\cdot 10^{21} - 8.51\cdot 10^{20}\ \text{photons}\, m^{-2}\,\text{sec}^{-1}\, sr^{-1}\right]$$
$$= 4.07\cdot 10^{6}\ \text{electrons} \tag{5.60}$$

This is the number of electrons we would expect to be generated by the radiance from a scene with an apparent temperature of 273K. However, as we indicated, there are two other sources of flux. The detector is surrounded by a 77K environment that is radiating over the entire spectral range from 0 to its cutoff frequency, so this will yield a photon irradiance of

$$E_{p\lambda} = \int\int L_{p\lambda}(T)\cos\theta\sin\theta d\theta d\phi \tag{5.61}$$

Since for our case the surround is at a fixed temperature of 77K and the integral is over nearly the entire hemisphere above the target, we can use the magic π derivation to approximate the integral as

$$E_{p\lambda} = L_{p\lambda}(T)\pi\left[\text{photons sec}^{-1}\, m^{-2}\mu m^{1}\right] \tag{5.62}$$

Note that a rigorous treatment would recognize that over the small angle subtended by the coldstop aperture, radiance associated with wavelengths larger than the cut on wavelength of the filter should not be included in the integral in Eq. (5.61). However, as we will see, ignoring this effect will result in only a small error in a small term and can be comfortably ignored for first-order calculations.

The electrons generated by the detector's thermal environment can then be expressed as

$$S_{eT} = \int_0^{\lambda co} L_{p\lambda}(77)\pi A_d tQE d\lambda \approx \pi A_d tQE \int_0^{12\mu m} L_{p\lambda}(77)d\lambda \tag{5.63}$$

where λ_{co} is the cutoff wavelength of the detector and we have again, for convenience, assumed that the quantum efficiency is constant over the wavelength range to allow us to use tabulated values for the photon radiance from Appendix A to yield

$$S_{eT} = \left[100\cdot 10^{-6} m\right]^2 \pi\ \text{sr}\ 2\cdot 10^{-5}\ \text{sec}\, 0.93\text{electrons/photon}\ 4.39\cdot 10^{15}\,\text{photons}\,\text{sec}^{-1}\, m^{-2} sr^{-1}$$
$$= 2.6\cdot 10^{3}\,\text{electrons} \tag{5.64}$$

which, as indicated, is small compared to the value of S_e, so our approximations should induce negligible error. The final and often very large source of flux is from the optics. This comes from the fact that the detector has a relatively large acceptance angle to gather flux from the entire telescope (note this is not the detector's instantaneous field of view, as discussed in Chapter 6). The forward optics are

usually operated at relatively high temperatures, and a significant fraction of the acceptance angle may include high-emissivity (i.e., nonmirror) surfaces. The number of electrons generated by the fore optics can be estimated as

$$\begin{aligned} S_{eo} &= \int \varepsilon_{eff} L_{p\lambda}(T)\Omega t QE \tau_F A_d d\lambda \\ &\approx \varepsilon_{eff}\Omega t QE \tau_F A_d \left[\int_0^{\lambda 2} L_{p\lambda}(T)d\lambda - \int_0^{\lambda 1} L_{p\lambda}(T)d\lambda \right] \\ &= 0.3 \cdot \pi[\sin(\theta/2)]^2[sr]2 \cdot 10^{-5}[\text{sec}]0.93\left[\frac{\text{electrons}}{\text{photon}}\right]0.89\left(100 \cdot 10^{-6}[m]\right)^2 \\ &\cdot \left(2.47 \cdot 10^{21} - 1.46 \cdot 10^{21}\right)\left[\text{photons sec}^{-1}m^{-2}sr^{-1}\right] = 4.75 \cdot 10^6 \text{electrons} \end{aligned} \quad (5.65)$$

where $\Omega = \pi(\sin\theta/2)^2$ is the acceptance solid angle of the detector, we have assumed the fore optics are all at a constant temperature of 297K, the filter cut on at 10 μm controls the lower limit of the bandpass, and the detector cutoff at 12 μm controls the upper limit. Furthermore, for convenience, we have again assumed all the spectral terms (ε_{eff}, QE, τ_F) are constant over the bandpass, so we can use the tabulated cumulative radiance values from Appendix A. Finally, we have approximated the fore optics model with a simple effective emissivity term of 0.3. This is the solid-angle weighted sum of the effective emissivities of all of the objects within the detector acceptance angle. Figure 5.19 provides a conceptual illustration of the world from the detector's point of view. In a rigorous fore optics model, the temperature, emissivity, and solid angle of all primary, as well as secondary, tertiary, ect., surfaces must be included in the computation.

The total number of electrons for our example case is then

$$\begin{aligned} S_{etot} &= S_e + S_{eT} + S_{eo} = 4.07 \cdot 10^6 + 0.0026 \cdot 10^6 + 4.75 \cdot 10^6 \\ &= 8.82 \cdot 10^6 \text{electrons} \end{aligned} \quad (5.66)$$

The shot noise for this case would then be

$$N_\Phi = S_{etot}^{\frac{1}{2}} = 2.97 \cdot 10^3 \text{electrons} \quad (5.67)$$

and is background limited (i.e., the bulk of the noise is caused by uncertainties in the signal level caused by random fluctuations in the number of non scene related photons. In many cases, (particularly those employing array detectors) the *background-limited infrared performance* (*BLIP*) is even more pronounced. For cases such as this, when the shot noise exceeds the aggregate electronic noise sources, we still have a noise that is predominately additive. This is because the background flux generates a noise floor that is only slightly impacted by the increase in flux (and therefore shot noise) when the signal increases.

If we assume our example system is limited by shot noise dominated by background flux, we can then evaluate the expected performance in terms of the NEΔT of the system. Using data from Appendix A for 297 K and 300 K blackbodies, we can use Eq. (5.59) to compute the change in signal per unit change in temperature at approximately 300 K to be

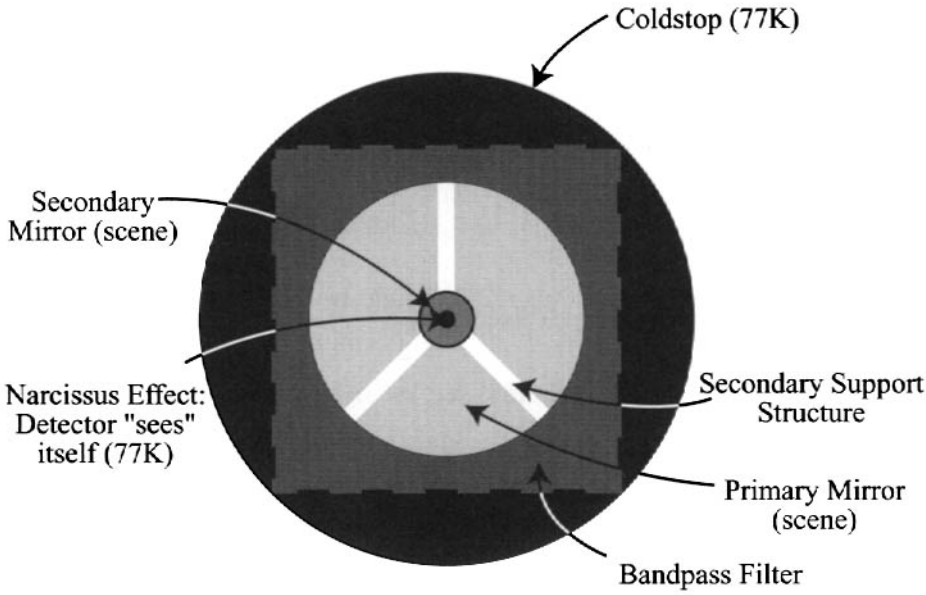


Figure 5.19 Conceptual illustration of the sources of radiance incident on a detector. Gray shades illustrate relative energy levels.

$$\frac{\Delta S_{etot}}{\Delta T} = \frac{2.73 \cdot 10^5 [\text{electrons}]}{3[\text{K}]} = 9.1 \cdot 10^4 [\text{electrons/K}] \tag{5.68}$$

Since we assume the noise is background limited, we can use the noise calculated above for a 273 K blackbody to estimate the NEΔT as

$$\text{NE}\Delta\text{T} = \frac{N_\Phi}{\Delta S_{etot} / \Delta T} \cong \frac{2.97 \cdot 10^3 [\text{electrons}]}{9.1 \cdot 10^4 [\text{electrons/K}]} = 0.033K \tag{5.69}$$

(Note: a useful student exercise involves seeing how the NEΔT changes for the actual shot noise increase calculated for a 300 K blackbody.)

This is the best possible performance that could be expected from this detector. In practice, electronic noise from the detector or processing electronics may reduce the actual observed performance, as discussed in Section 6.5.

In this section, we have reviewed several of the factors that characterize a detector's or a sensor's ability to resolve small changes in the incident signal (radiance), which can be described as the sensor's radiometric resolution. The absolute magnitude of the radiance levels has not yet been considered and will be addressed in Section 6.4. For brevity's sake, we have not addressed many of the components of sensor systems as discrete entities. The next chapter describes several sensor systems and introduces component elements as needed.

5.5 REFERENCES

Accetta, J.S., & Shumaker, D.L., eds. (1993). *The Infrared and Electro-optical Systems Handbook.* SPIE Optical Engineering Press, Bellingham, WA.

Budde, W. (1983). *Optical Radiation Measurements: Vol. 4, Physical Detectors of Optical Radiation.* Academic, NY.

Colwell, R.N., ed. (1983). *Manual of Remote Sensing.* 2d ed., American Society of Photogrammetry and Remote Sensing, Bethesda, MD.

Dereniak, E.L., & Crowe, D.G. (1984) *Optical Radiation Detectors.* Wiley, NY.

Grum, F., & Becherer, R.J. (1979). *Optical Radiation Measurements: Vol. 1, Radiometry.* Academic, NY.

Hecht, E. (1990). *Optics.* 2nd. ed., Addison-Wesley, Reading, MA.

Hurter, F., & Driffield, V.C. (1890). Photo chemical investigations and a new method of determination of the sensitivity of photographic plates. *Journal of the Society of Chemical Industrials*, No. 9, pp. 455-469.

James, T.H. (1977). *Theory of the Photographic Process.* 4th ed., Macmillan, NY.

Kruse, P.W. (2001). Uncooled thermal imaging arrays, systems and applications, tutorial texts. In *Optical Engineering*, Vol. TT51. SPIE, Bellingham, WA.

Rogatto, W.D., ed. (1993). Electro-optical components. *The Infrared and Electro-optical System Handbook*, Vol. 3, J.S. Accetta, & D.L. Shumaker eds. SPIE Optical Engineering Press, Bellingham, WA.

Wolfe, W.L., & Zissis, G.J., eds. (1985). *The Infrared Handbook.* Rev. ed., Environmental Research Institute of Michigan, Ann Arbor, MI.

This page intentionally left blank

CHAPTER 6
IMAGING SENSORS AND INSTRUMENT CALIBRATION

In this chapter, we will explore imaging sensors as complete systems designed to capture radiometric signals and to enable the reconstruction of full two-dimensional geometrically and radiometrically faithful representations of the sampled radiance field.

For the sake of space, we will emphasize airborne and satellite electro-optical (EO) imaging systems concepts, introducing specific systems only as examples of more generic approaches. The reader is referred to Chen (1985) for a more complete treatment of satellite sensing systems, including sounders and microwave systems, and to Kramer (2001) and Morain and Budge (2001) for a comprehensive listing of the specifications for a wide range of aerospace sensing systems.

The end of this chapter links together many of the concepts in this and earlier chapters through a case study of a system design.

6.1 SINGLE-CHANNEL AND MULTISPECTRAL SENSORS

This section describes several of the critical components and design features of airborne and satellite imaging systems. With new sensors evolving at a rapid rate, this section emphasizes concepts rather than the details of specific sensors (though examples of current operational sensors are included). Sensors are somewhat arbitrarily divided into those with one to ten or so spectral channels (*multispectral*) and those with tens to hundreds of spectral channels (*hyperspectral*). In general,

airborne and satellite designs are intermixed, because in many ways they are more similar than dissimilar. For convenience, we will tend to treat all the sensors as though they are digital systems forming discrete pixels. This merely simplifies the terminology used in describing the systems. The reader should recognize that the digital concepts are merely the result of sampling a continuous signal that can be, and sometimes is, recorded instead (cf. Sec. 8.1).

6.1.1 Line Scanners

Remote sensing systems are most often categorized in terms of how the image is formed (e.g., one pixel at a time, one row at a time). In many ways, the simplest imaging sensor is the line scanner (cf. Fig. 6.1). These sensors employ a spinning (scan) mirror to project the image of the detector along a line on the ground perpendicular to the aircraft or satellite ground track. By sampling the signal from the detector, the across-track image lines can be formed. During the rotation of the scan mirror, the sensor platform advances slightly, and consecutive rotations of the mirror sweep out consecutive lines on the ground, which are sampled to form the across-track columns that make up the image (cf. Fig. 6.1). The angular extent of the image across-track is referred to as the *field of view* (*FOV*) of the imager, and the angular extent of the individual detector element is called the *instantaneous field of view* (*IFOV*). The projection of the detector onto the ground is referred to as the *ground instantaneous field of view* (*GIFOV*) or the *ground spot* of the sensor, i.e., at nadir

$$\text{GIFOV} = H \cdot \text{IFOV} \tag{6.1}$$

where H is the flying height above ground level. The conventional line scanner design uses square detectors that are sampled along track on pixel centers and the ideal ground track advances one GIFOV per rotation to sample contiguous lines with each rotation. Thus, every point on the ground is imaged (sampled) without gaps and without overlaps. Some systems employ oversampling to improve the spatial resolution of the reconstructed image (cf. Sec. 13.1).

Line scanners, like most nonframing imagers we will discuss, have a unique set of geometric distortions caused by the way the image is sampled and by the motion of the sensing platform during imaging. Space-based sensor platforms are usually geometrically stabilized such that the only significant motion of the platform during imaging is the along-track motion of the spacecraft. Aircraft platforms are often not stabilized so that the orientation of the aircraft can change from one line to the next or, in extreme cases, even from pixel to pixel within a line. The resulting distortions are illustrated in Figure 6.2. Aircraft aerodynamics are such that *pitch* and *yaw* (crab) are generally relatively constant errors typically removed in postprocessing. *Roll* effects vary considerably on a line-to-line basis. Roll can be corrected either by roll stabilizing the imager or by recording the amount of roll using a gyroscope signal and advancing or delaying each line of data by the number of IFOVs of roll. Figure 6.3 shows an image before and after roll compensation was performed using signals recorded from a gyroscope. Pitch and yaw can also

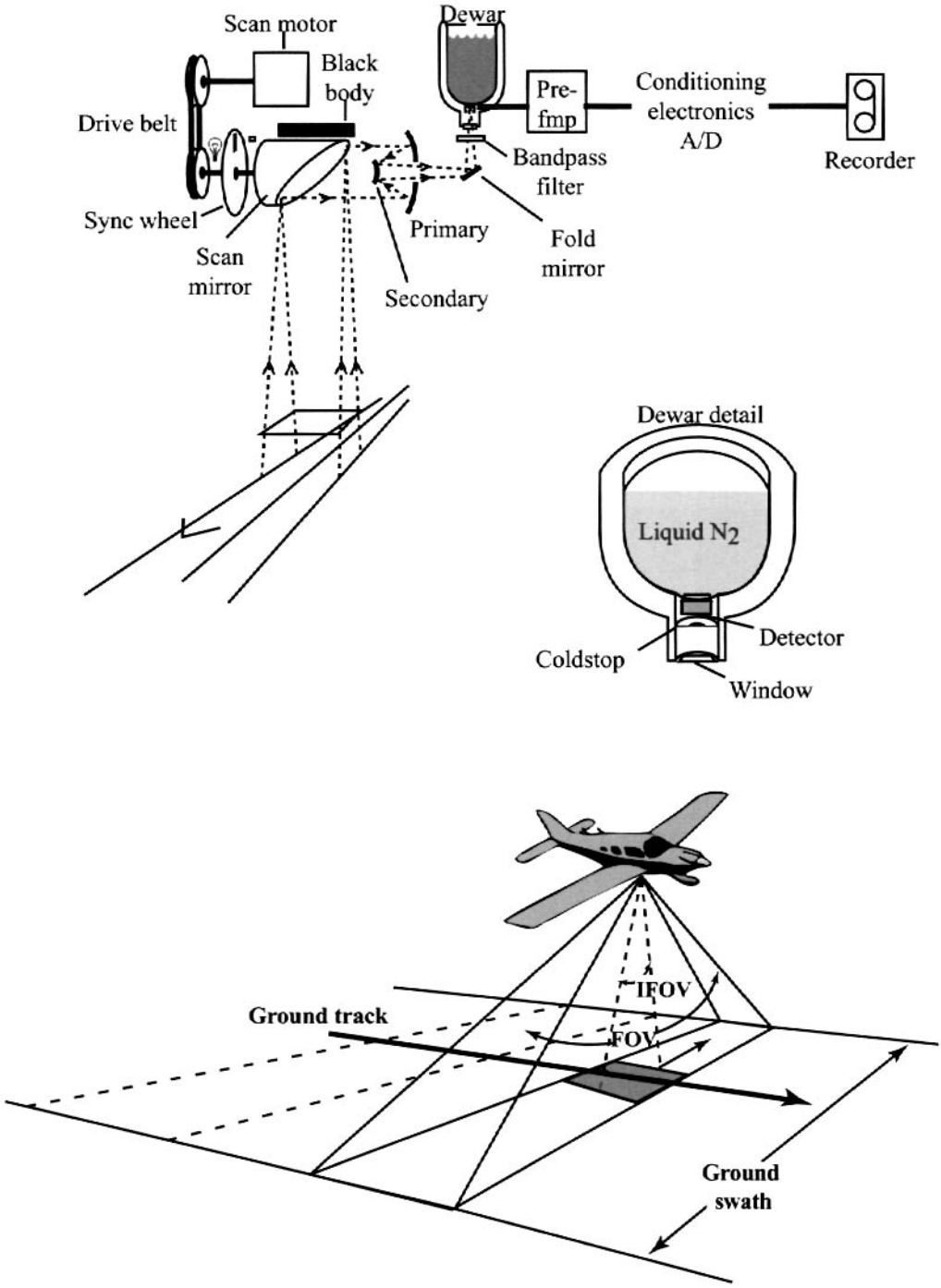


Figure 6.1 Line scanner design and collection scheme. A thermal infrared line scanner is shown with liquid nitrogen cooling of the detector and coldstop.

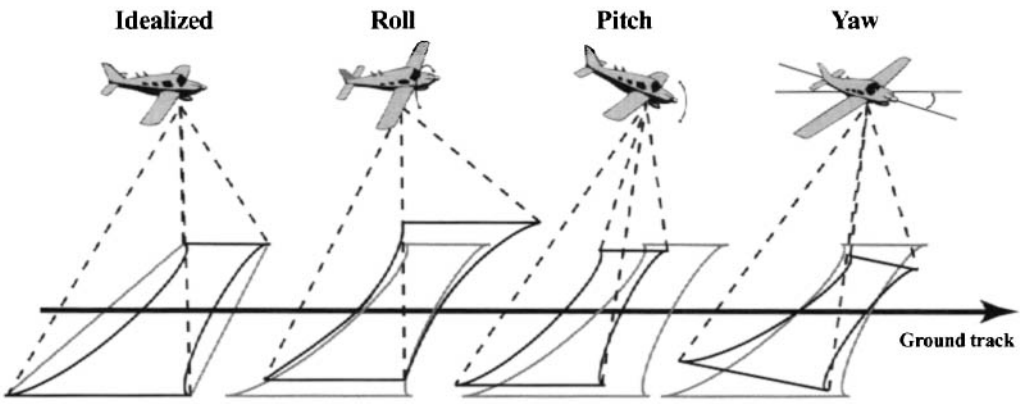


Figure 6.2 Geometric distortions due to aircraft orientation. Gray boundaries represent nominal coverage; black boundaries represent actual coverage.

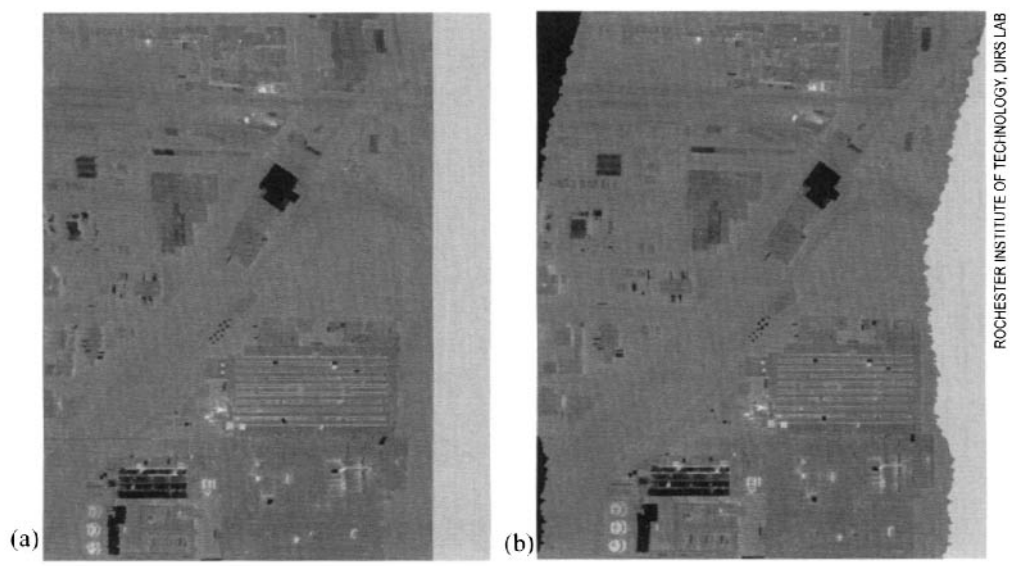


Figure 6.3 Portion of a thermal infrared (8–14 μm) line scanner image: (a) before and (b) after lines were shifted to correct for roll distortion.

be corrected using data from a three-axis gyro; and approximately constant errors in pitch and yaw can be removed using standard resampling techniques for image or map projection (cf. Sec. 12.3.1.2).

Another type of image distortion is introduced when the scan motor speed is not matched to the aircraft speed such that the aircraft does not advance exactly one GIFOV between scan lines. The error is called V/H (velocity over altitude) error and results in longitudinal compression or stretching of the image, as shown in Figure 6.4. The final type of distortion characteristic of line scanner imagery is called *tangent distortion* and results from the fact that the data are typically sampled on equal angular steps (e.g., every IFOV), and this results in each sample representing

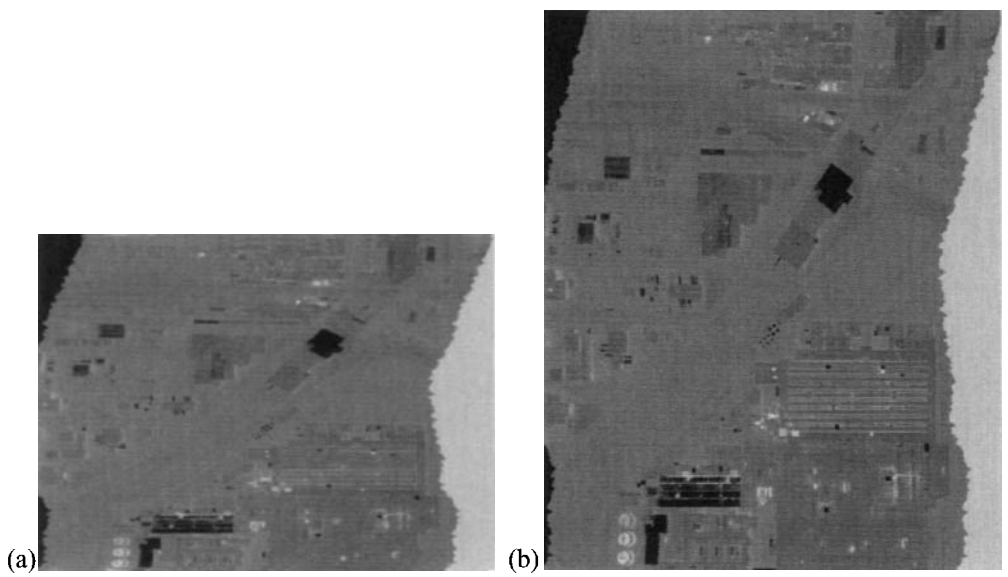
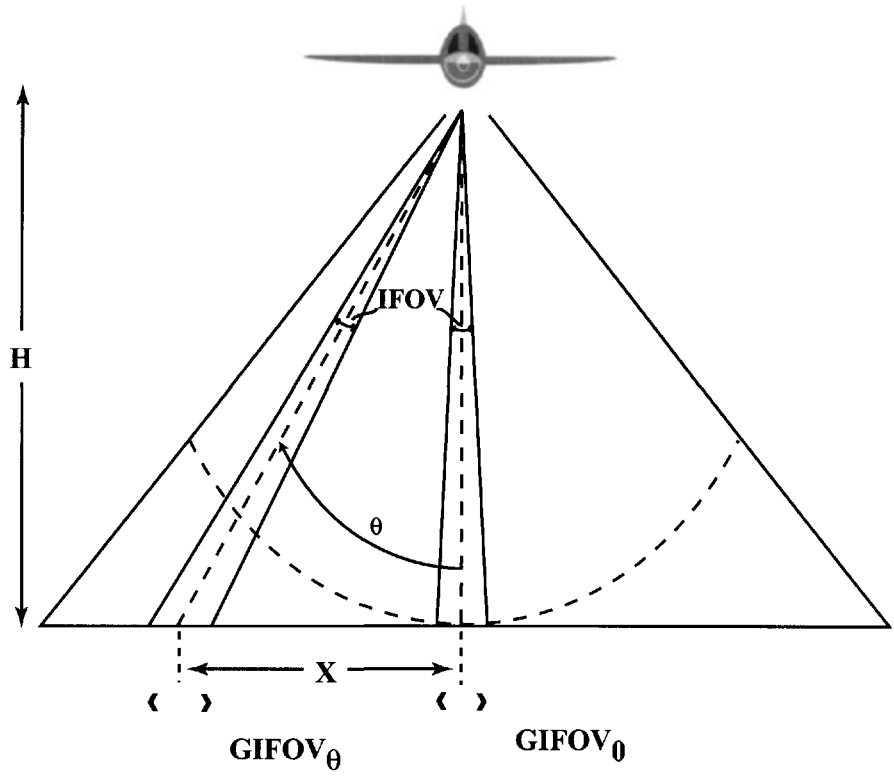


Figure 6.4 Thermal infrared image showing V/H distortion: (a) original image, (b) and after resampling to proper V/H ratio.

a larger projected area on the ground as we progress off axis (cf. Fig. 6.5). The image can be thought of as the projection of the ground onto a cylinder that has the effect of foreshortening the edges of the image. This can be seen in the image in Figure 6.5, where the buildings become compressed and straight diagonal roads become curved toward the edges. This type of systematic error can be removed through geometric resampling of the image as described in Section 12.3. These corrections, however, cannot restore the resolution lost due to the larger GIFOV off axis.

Many modern systems take advantage of the *global positioning system* (*GPS*) and *inertial navigation systems* (*INS*) to record the x,y,z location and roll, pitch, and yaw orientation of the instrument on small time centers so that the sensor's six degrees of freedom can be calculated for each rotation of the scan mirror (cf. Fig.6.6). These data can then be used to project the sampled data onto a geographic coordinate system where they can be resampled to form a geometrically rectified image (cf. Chap. 12 for a discussion of resampling). When terrain elevation data exist in the form of *digital elevation models* (*DEMs*), the GPS-INS data can be combined with DEMs to project the data onto an ortho-rectified space where they can be resampled into ortho-rectified images.

Line scanners often have large fields of view (90° to 120°) providing large ground swaths. This is possible with fairly simple optics, because only the very central portion of the lenses is used for imaging with the scan mirror pointing the telescope's optical axis off nadir. The line scanner design suffers from the disadvantage that a single detector does all the sampling, so the dwell time (the time the detector can spend gathering photons from a spot on the ground) is very short. This problem is exacerbated by the large "dead time" when the scan mirror is looking up inside the sensor. Most systems use some of this dead time to look at sources of



$$GIFOV_0 = IFOV \cdot H$$

$$GIFOV_\theta = H\left[\tan\left(\theta + \frac{IFOV}{2}\right) - \tan\left(\theta - \frac{IFOV}{2}\right)\right]$$

(a)

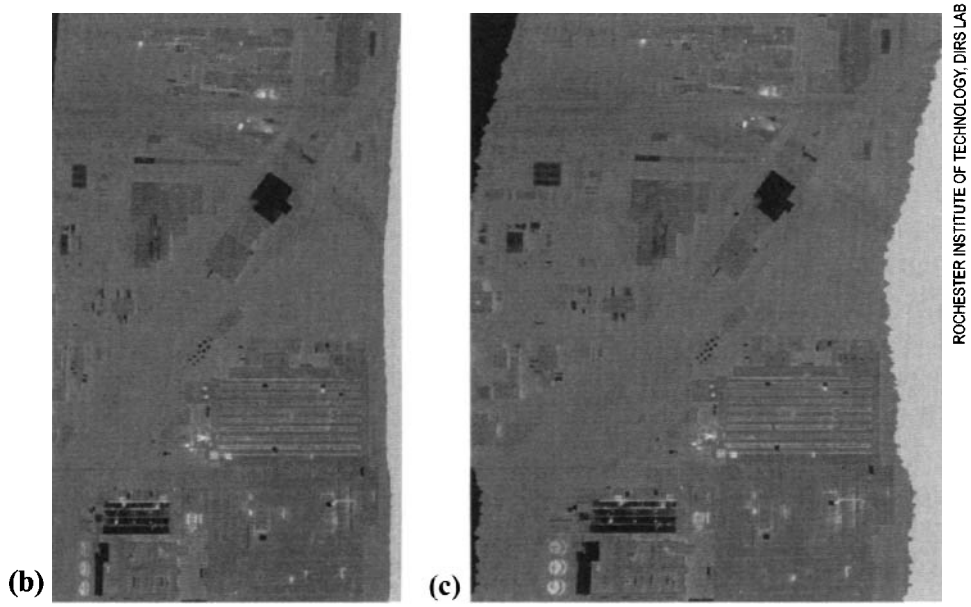
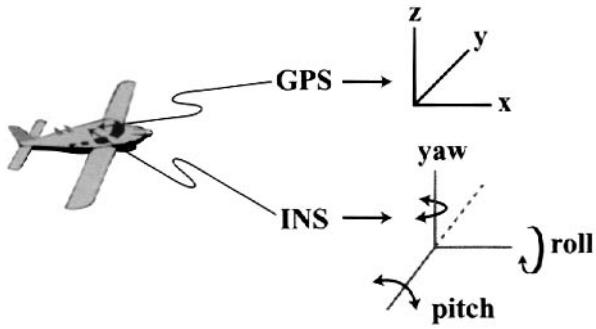
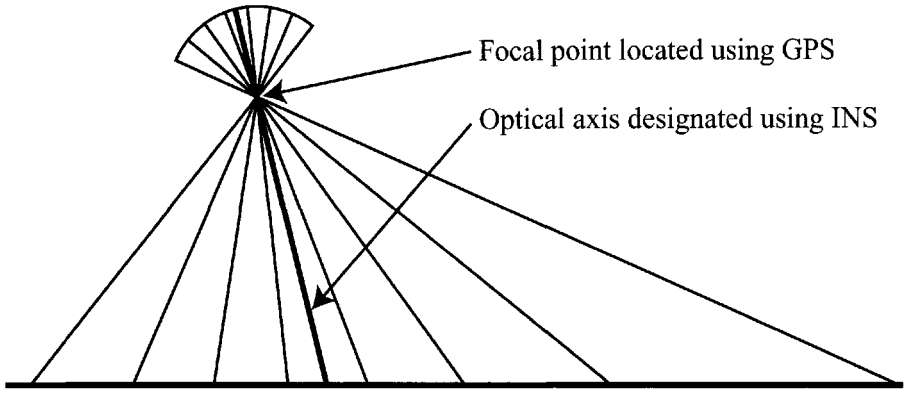


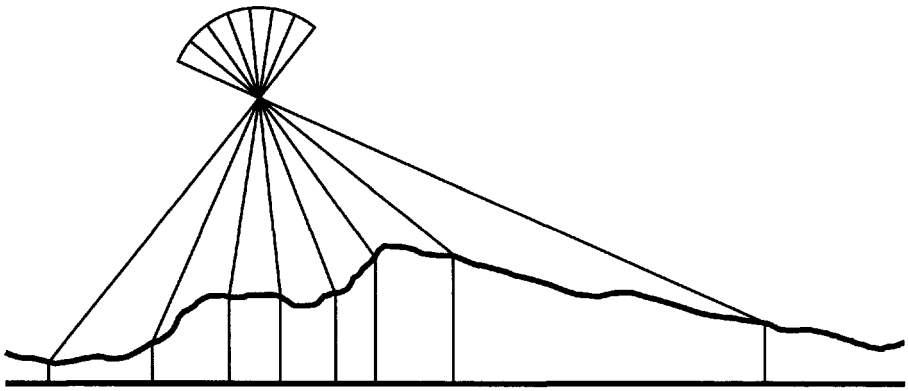
Figure 6.5 Tangent error effects. The diagonal roads in the thermal infrared line scanner image should be straight, and the tanks in the lower left roughly circular. (b) Original image and (c) image after tangent correction.



(a)



(b)



(c)

Figure 6.6 Illustration of use of GPS-INS data to reconstruct scene geometry: (a) GPS-INS is used to generate sensor location and orientation in world coordinate system; (b) sensor internal geometry, defined relative to the optical axis, is projected onto the ground coordinate system using GPS-INS to locate and orient the sensor's optical axis; (c) use of DEM to compute terrain intersection points and project data into an ortho-rectified space.

known radiance inside the scanner for calibration purposes (e.g., the blackbody in Fig. 6.1). A major advantage of the line scanner besides the simplicity of its optics is the inherent registration of multispectral data in many line scanner designs (cf. Fig. 6.7). Many of these systems place the entrance aperture to a monochromator at the focal plane of the optical system. This aperture is the limiting stop in the system defining the sensor's IFOV and sample size. The monochromator disperses the data spectrally with the detector width defining the spectral bandwidth of each channel. The spectral sampling can be performed with discrete detectors or with a linear array of detectors. The signal from each detector is amplified and processed to become a data channel for recording or transmitting to the ground. Because the data are sampled simultaneously through a common aperture, the resulting images are inherently registered [i.e., pixel (i,j) of the green image will fall directly on pixel (i,j) of the red image]. In many multispectral scanners, the spectral range cannot be covered with a single monochromator. In the scanner design illustrated in Figure 6.7, the thermal infrared energy is folded out of the primary beam with a beam splitter and focused onto a single cryogenically cooled detector. Line scanners commonly use a synchronization signal that lets the system know when to sample the calibration sources and when and how frequently to sample the Earth (cf. Fig. 6.8).

Line scanners have been extensively used in space for routine observation of the Earth at relatively low resolution. One of the most widely used data sets comes from the Advanced Very High Resolution Radiometer (AVHRR), which acquires multispectral data of the globe twice a day with approximately 1 km spatial samples (cf. Morain and Budge 2001). The Geosynchronous Operational Environmental Satellite (GOES) uses an interesting variation of the traditional line scanner design called *spin scan*. Rather than use a spinning scan mirror, the entire satellite, located in geosynchronous orbit, rotates to provide the line scan effect, and the mirror oscillates to provide the line advance as shown in Figure 6.9. The GOES images provide twice-hourly coverage of the hemisphere, permitting time-lapse motion sequences showing cloud dynamics.

6.1.2 Whisk-Broom and Bow-Tie Imagers

One simple way of overcoming the short dwell time of the line scanner design is to take several lines of data simultaneously. Then the time between repeat cycles of the scan mirror is reduced by the number of lines collected per sweep. This approach has been very successfully used in the *multispectral scanner* (*MSS*) instruments flown on Landsat satellites 1 through 5 starting in 1972. The scan mirror oscillates, as shown in Figure 6.10, acquiring several lines of data, but only in one direction. This method of data collection is analogous to using a *whisk broom* to sweep dust (data) into a dust pan (data stream). In the case of the Landsat *MSS*, six lines of data are collected with six different detectors with each mirror sweep. The next mirror sweep is timed so that the next six lines are immediately adjacent to the last, providing continuous ground coverage. Multiple spectral bands are acquired by locating additional detectors on the focal plane in the along-scan direction. The detectors are spectrally filtered to control the wavelength sampled. As the image

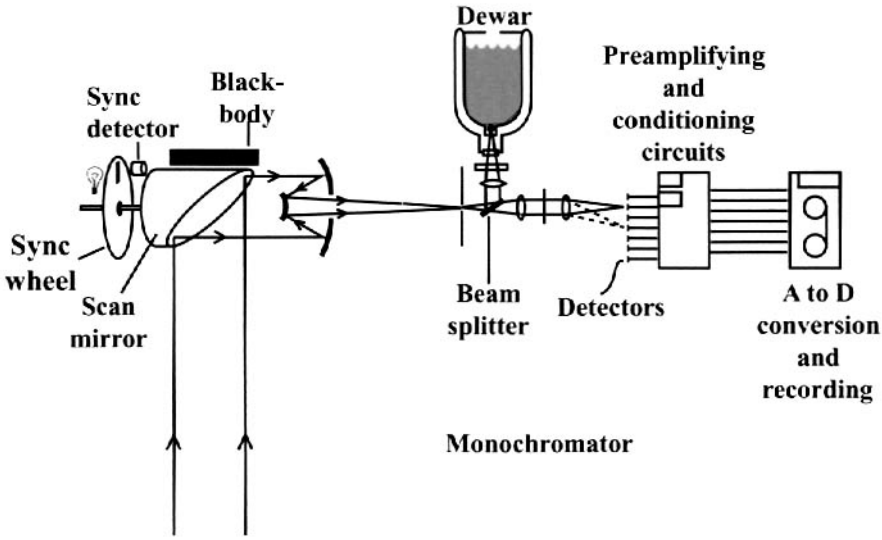


Figure 6.7 A multispectral line scanner design.

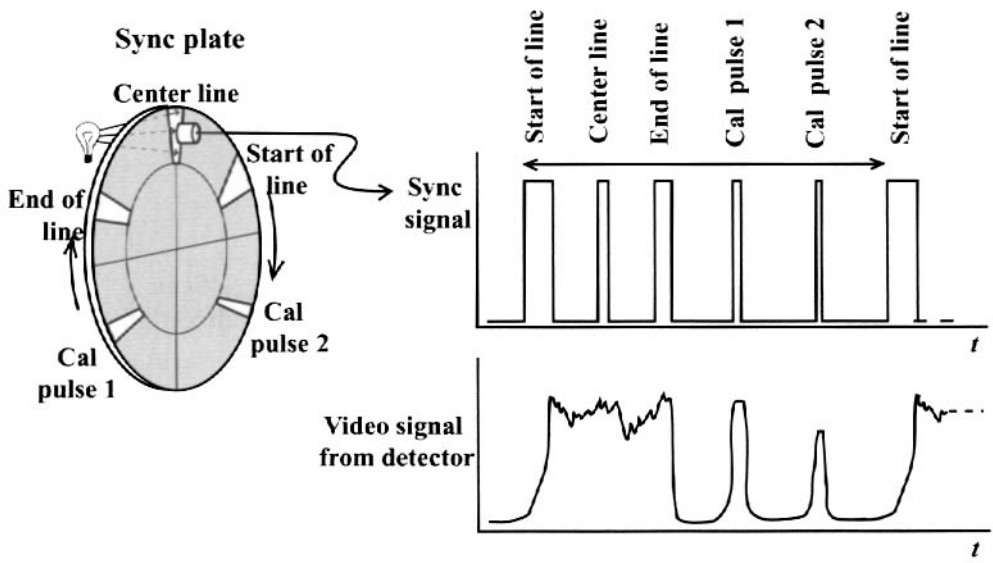


Figure 6.8 Synchronization signals. An opaque sync plate attached to the scan mirror has transparent windows to trigger the sync detector.

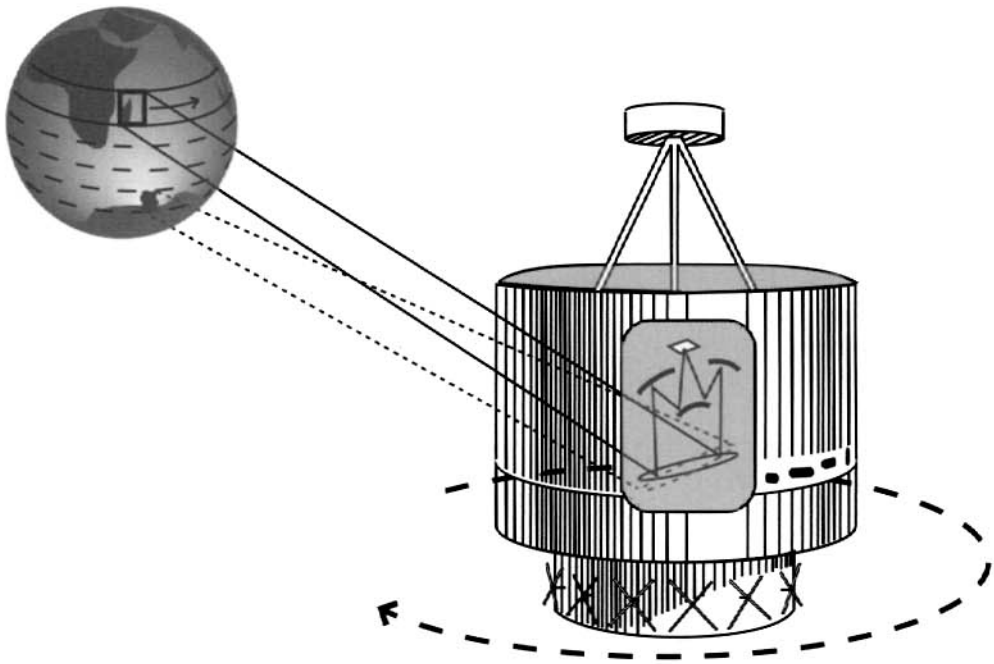


Figure 6.9 Spin scan coverage used from geosynchronous orbit (e.g., GOES).

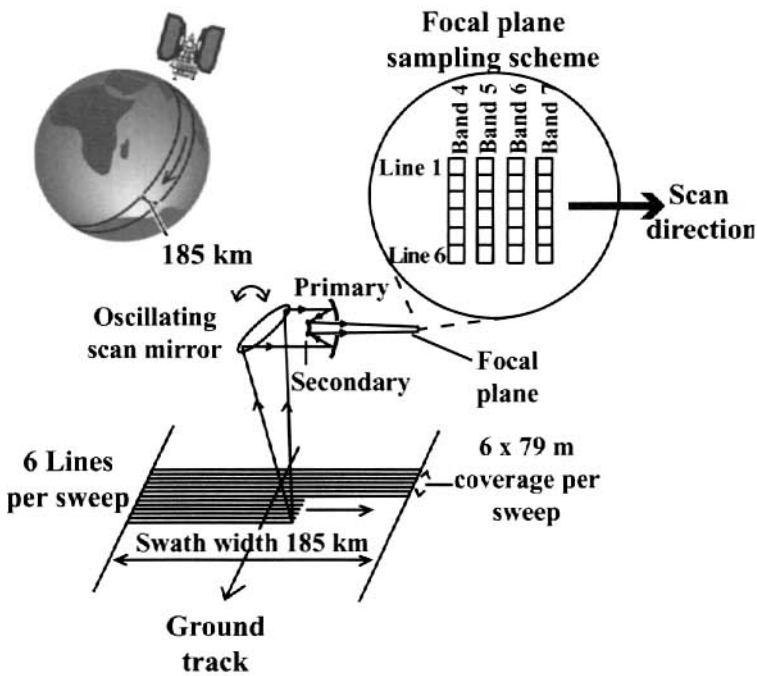


Figure 6.10 Whisk-broom design used on the Landsat MSS instrument.

is swept across the focal plane by the scan mirror, each point on the ground is sampled: first in one spectral band, then moments later by a second, etc. In reconstructing the images for spectral registration, the pixels in each spectral band must be shifted by a few pixels to properly align with the previous band. For most of the Landsat MSS instruments, four channels were collected, resulting in a total of 24 detectors. The actual detectors on the MSS are photomultiplier tubes for the green, red, and first IR channel and photo diodes for the second IR channel (cf. Table 6.1). The focal plane is sampled by light pipes (fiber optics), which carry the signal to the detectors. This allows physically large devices (PMTs) to be effectively located in close proximity on the focal plane.

The increased dwell time of the MSS whisk-broom design allowed it to achieve moderate spatial (79 m GIFOV) and spectral ($\Delta\lambda \approx 0.1$ μm) resolution. The whisk-broom approach still wastes approximately half the useful scan time by only taking data in one direction. The *Thematic Mapper* (*TM*), which was flown along with the MSS on Landsat 4 and 5, and the *enhanced Thematic Mapper* (*ETM+*) flown on Landsat 7 (note Landsat 6 did not achieve orbit) use an oscillating mirror that scans in both directions (cf. Fig. 6.11). As illustrated in Figure 6.12,

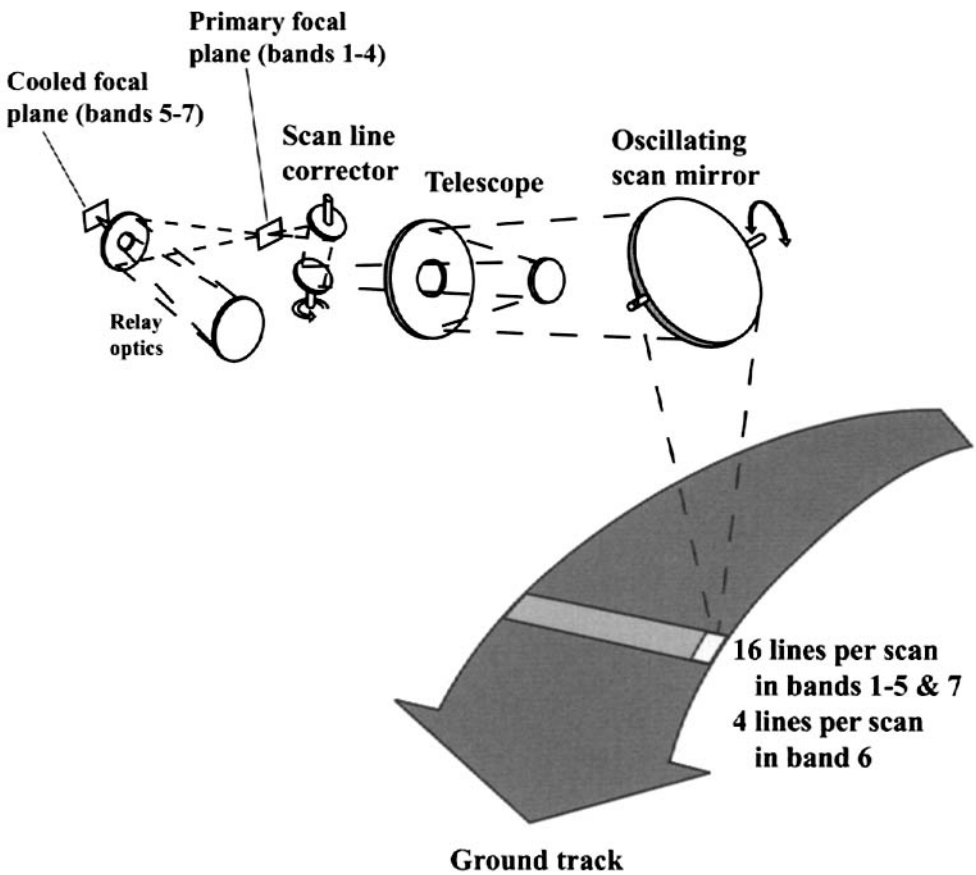


Figure 6.11 Landsat TM optical layout.

Table 6.1 Characteristics of Landsat MSS and TM Sensors

Sensor	Orbit Altitude	Repeat Period	FOV (degrees km)	GIFOV		Nominal Spectral Bands (μm)
MSS 1,2,3	Sun sychronous descending equatorial crossing = 9:30 a.m. 913 km	18 days	11.6 185 km	79 m	# 4	0.5 - 0.6
				79 m	# 5	0.6 - 0.7
				79 m	# 6	0.7 - 0.8
				79 m	# 7	0.8 - 1.1
MSS 3 only				237 m	# 8	10.4 - 12.6
MSS 4&5	Sun sychronous descending equatorial crossing = 9:45 a.m. 705 km	16 days	14.9 185 km	82 m	# 1	0.5 - 0.6
				82 m	# 2	0.6 - 0.7
				82 m	# 3	0.7 - 0.8
				82 m	# 4	0.8 - 0.11
TM 4&5	Sun sychronous descending equatorial crossing = 9:45 a.m. 705 km	16 days	14.9 185 km	30 m	# 1	0.45 - 0.52
				30 m	# 2	0.52 - 0.60
				30 m	# 3	0.63-0.69
				30 m	# 4	0.76 - 0.90
				30 m	# 5	1.55 - 1.75
				120 m	# 6	10.40 - 12.3
				30 m	# 7	2.08 - 2.35
ETM + 7	Sun sychronous descending equatorial crossing = 10:00 a.m. 705 km	16 days	14.9 185 km	30 m	# 1	0.45 - 0.52
				30 m	# 2	0.53 - 0.61
				30 m	# 3	0.63 - 0.69
				30 m	# 4	0.78 - 0.90
				30 m	# 5	1.55 - 1.75
				60 m	# 6	10.40 - 12.5
				30 m	# 7	2.09 - 2.35
				15 m	# 8	0.52 - 0.90

this approach will result in gaps and overlap regions in the ground coverage. To compensate for this, a pair of rotating parallel mirrors called the *scan line corrector* are included in the TM optical chain. These mirrors shift the image projected onto the detectors so that it is slightly ahead of the across-track location at the start of each scan and ends behind the start point. The corrector advances and retards in this fashion during both the forward and reverse scans of the primary mirror, such that the scan projections on the ground fall alongside each other (cf. Fig. 6.12). The TM takes advantage of the increased dwell time available from this bow-tie correction to increase the spatial and spectral resolution over the MSS. To achieve this, the ground coverage per mirror sweep is kept approximately the same, using

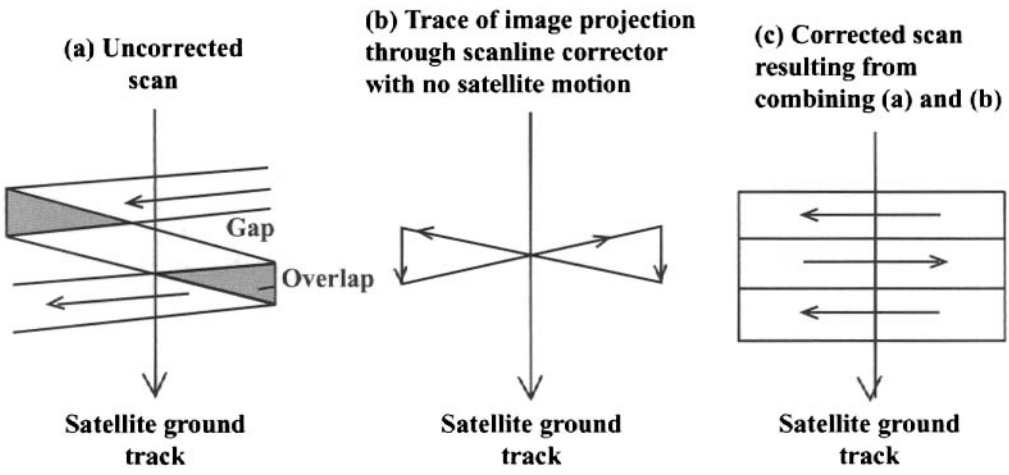


Figure 6.12 Bow-tie scan line correction used with Landsat TM and ETM+.

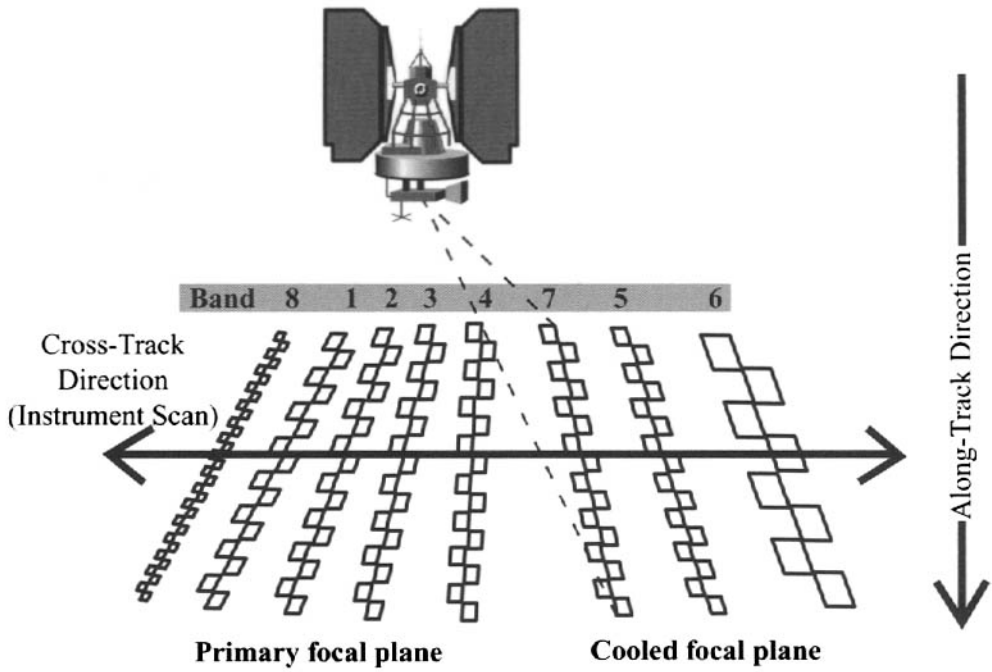


Figure 6.13 Projection of Landsat ETM+ detectors onto ground.

more (16) but smaller (30-m GIFOV) detectors. In addition, the number of bands is increased to seven, as listed in Table 6.1. The TM detectors are located on the focal plane using a staggered array (cf. Fig. 6.13) to make room for the individual detectors. Relay optics are used to focus a portion of the image onto a cooled focal plane where the SWIR and LWIR detectors are located. To achieve adequate signal to noise, the GIFOV of the LWIR band is four times (120 m) that of the other bands on TM and twice as large (60 m) on ETM+. In addition, the ETM+ has a 15

Figure 6.14 Landsat TM full-scene of eastern Lake Ontario and the Finger Lakes region of New York State.

m panchromatic band for spatial sharpening that has twice as many detectors (32) at half the size.

The signals transmitted to the ground are preprocessed to account for the staggered array effects, the spectral band offsets, and offsets between the data taken in the forward and reverse sweep. In addition, both MSS and TM data are nominally corrected for the effects due to the Earth's rotation during the time of data collection. As the satellite travels south in its descending node, the Earth's eastward rotation causes the effective ground track to drift westward. The resulting image is skewed and must be deskewed for proper ground projection. Figure 6.14 shows a full frame of Landsat data after ground processing, including the characteristic rhomboid shape indicating that it has been deskewed. A more complete treatment of the Landsat sensors can be found in Barker (1985), Markham and Barker (1985) and Goward et al. (2001).

The whisk-broom–style sensors have largely been used only in space because of the difficulties in correcting for geometric errors in an unstable platform. However, GPS-INS sensors are reducing these difficulties. Whisk-broom sensors

offer the advantage of increased dwell time allowing for either higher signal to noise or higher spatial and/or spectral resolution. They have the disadvantage that they require slightly larger fields of view (i.e., image quality must be maintained over the entire detector array) for the telescope, and the spectral data are not inherently registered. For Landsat TM, the detectors were contained within one-quarter of a degree of the optical axis within an area approximately 1 cm on a side, so the constraint on the optics was not very severe. Wrigley et al. (1985) report that the initial band-to-band registration of the Landsat TM focal planes exhibited several tenths of a pixel misregistration. However, they go on to report that, after post-launch corrections in the ground processing, band-to-band registration met the 0.2 pixel misregistration specification and that registration is typically good to better than 0.1 pixel. Similarly, for the ETM+ [Storey and Choate (2002)] report band-to-band registration to better than 0.1 pixel for ETM+.

We have emphasized the discussion of the Landsat series of instruments because of their long history of operation, the ready accessibility of the data, and the fact that the instruments (particularly ETM+) have been operated in a fashion designed to generate a global seasonal/annual archive of relatively high spatial resolution data [cf. Arvidson et al. (2001)]. Thus, the Landsat data represent the backbone for a large number of Earth remote sensing studies, particularly where temporal issues or high spatial detail over large areas are required.

6.1.3 Push-Broom Sensors

The *push-broom sensors* represent a further step toward increasing the dwell time to allow system designers to make signal-to-noise or resolution tradeoffs. These sensors use linear array detectors, as illustrated in Figure 6.15, to collect entire lines of data simultaneously. Multiple spectral bands are collected, with multiple linear arrays filtered for the bands of interest. Since 1986, the *French Système Probatoire d'Observation de la Terre* (*SPOT*) instruments have used a push-broom approach. With this approach, an individual detector element only needs to sample at one across-track location. This provides increased integration time for the sensor. The SPOT system uses this to achieve 5 m GIFOVs for a panchromatic band and 10 m GIFOVs for multispectral bands, as listed in Table 6.2. One of the disadvantages of the push-broom approach is that very long arrays are necessary to achieve a large ground swath. For example, the SPOT ground swath is only 60 km, compared to Landsat's 185 km and AVHRR's 2400 km. The SPOT system overcomes this limitation by using an across-track pointing mirror to allow it to look up to 31° to the left or right of the nominal ground track. In addition to increasing the probability of acquiring targets of interest, this allows for stereoscopic coverage using data from multiple orbits. The push-broom approach has the distinct advantage in satellite operation of having no movable parts. This increases the expected lifetime and may reduce power requirements. The detector technology, however, is far more sophisticated than the line scanner or whisk-broom approaches. Long arrays, even using the staggered-array approach illustrated in Figure 6.15, can be difficult to manufacture for some materials (e.g., HgCdTe). In addition, for focal planes that require cooling, it can be very difficult to design and provide power for

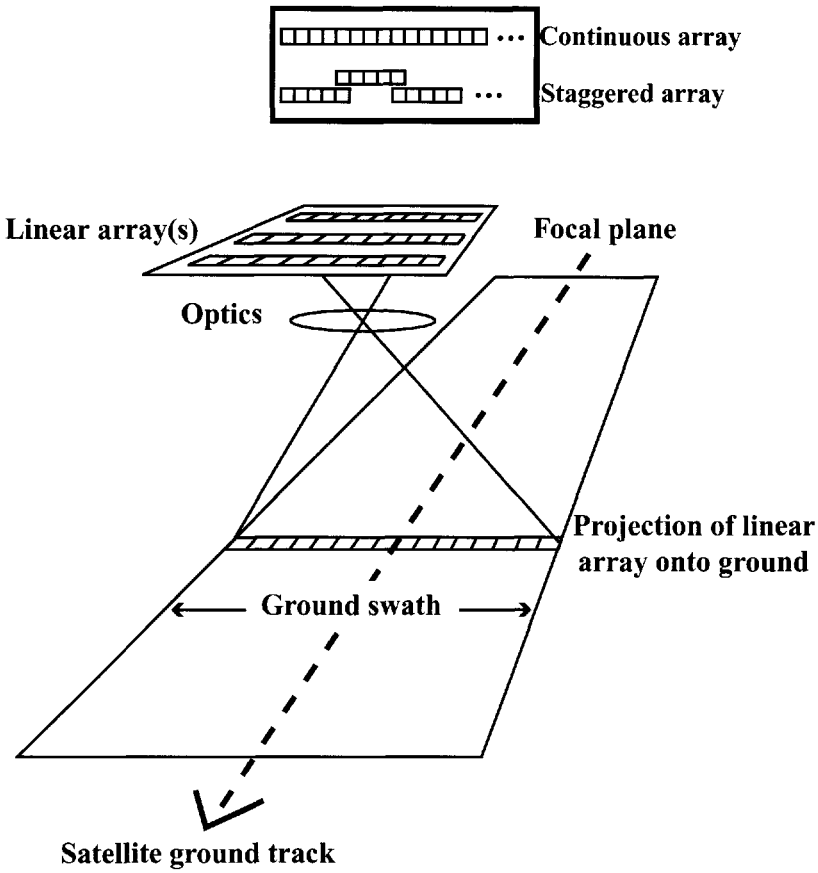


Figure 6.15 Push-broom sensor operation.

cooling large focal planes. Despite these limitations, advances in detector technology are pushing more and more space-based systems toward push-broom solutions because of the advantages of no moving parts and the long dwell times that allow for higher spatial or spectral resolution. The spectral bands still need to be registered by shifting to correct for the offset between the arrays; however, the lack of scan motion generally improves the inherent geometric fidelity of push-broom systems. This is most readily seen in the new generation of body pointing push-broom satellite instruments that are making high-resolution (less that 1 m) multispectral imagery available [cf. Cook et al. (2001) and Krause (2004)]. Two examples are listed in Table 6.3. These designs take advantage of large arrays and time delay and integration (TDI) techniques to allow very high-resolution image acquisition (cf. Fig. 6.16). However, they are typically limited to relatively narrow fields of view compared to the instruments introduced earlier. The ability to point the satellite allows for acquisition of imagery off the nominal ground track, which improves the repeat rate and also facilitates stereoscopic coverage. To date, these high-resolution sensors have been limited to using silicon-based detector technology restricting them to a few spectral bands in the VNIR region. However, this is not a fundamental limitation and advances in spectral range, spatial resolution, and ground swath width should all be forthcoming based on incremental improvements

Table 6.2 Characteristics of SPOT Instrument

Instrument	Orbits Altitude	Repeat Cycle	FOV	GIFOV	Spectral Bands (μm)
SPOT 1-3 1986, 1990, 1993	Sun synchronous descending equatorial crossing ≈ 10:30 a.m. 822 km	26 days	4.13° 60 km	10 m	0.50 - 0.73
				20 m	0.50 - 0.59
				20 m	0.61 - 0.68
				20 m	0.72 - 0.89
SPOT 4 1998	Sun synchronous descending equatorial crossing ≈ 10:30 a.m. 822 km	26 days	4.13° 60 km	10 m	0.61 - 0.68
				20 m	0.50 - 0.59
				20 m	0.61 - 0.68
				20 m	0.72 - 0.89
				20 m	1.58 - 1.75
SPOT 5 2002	Sun synchronous descending equatorial crossing ≈ 10:30 a.m. 822 km	26 days	4.13° 60 km	5 m	0.48 - 0.71
				10 m	0.50 - 0.59
				10 m	0.61 - 0.68
				10 m	0.78 - 0.89
				20 m	1.58 - 1.75

Table 6.3 Characteristics of Example High-Resolution Push-Broom Imagers

Instrument Launch Date	Orbit Altitude	Ground Swath at Nadir	GIFOV (nadir)	Spectral Bands (μm)
Ikonos 1999	Sun synchronous descending equatorial crossing ≈ 10:30 a.m. 681 km	11.3 km	0.8 m	0.45 - 0.90
			3.2 m	0.45 - 0.52
			3.2	0.51 - 0.60
			3.2	0.63 - 0.70
			3.2	0.76 - 0.815
QuickBird 2001	Sun synchronous descending equatorial crossing ≈ 10:30 a.m 681 km	16.5 km	0.6	0.45 - 0.90
			2.4	0.45 - 0.52
			2.4	0.52 - 0.60
			2.4	0.63 - 0.69
			2.4	0.76 - 0.90

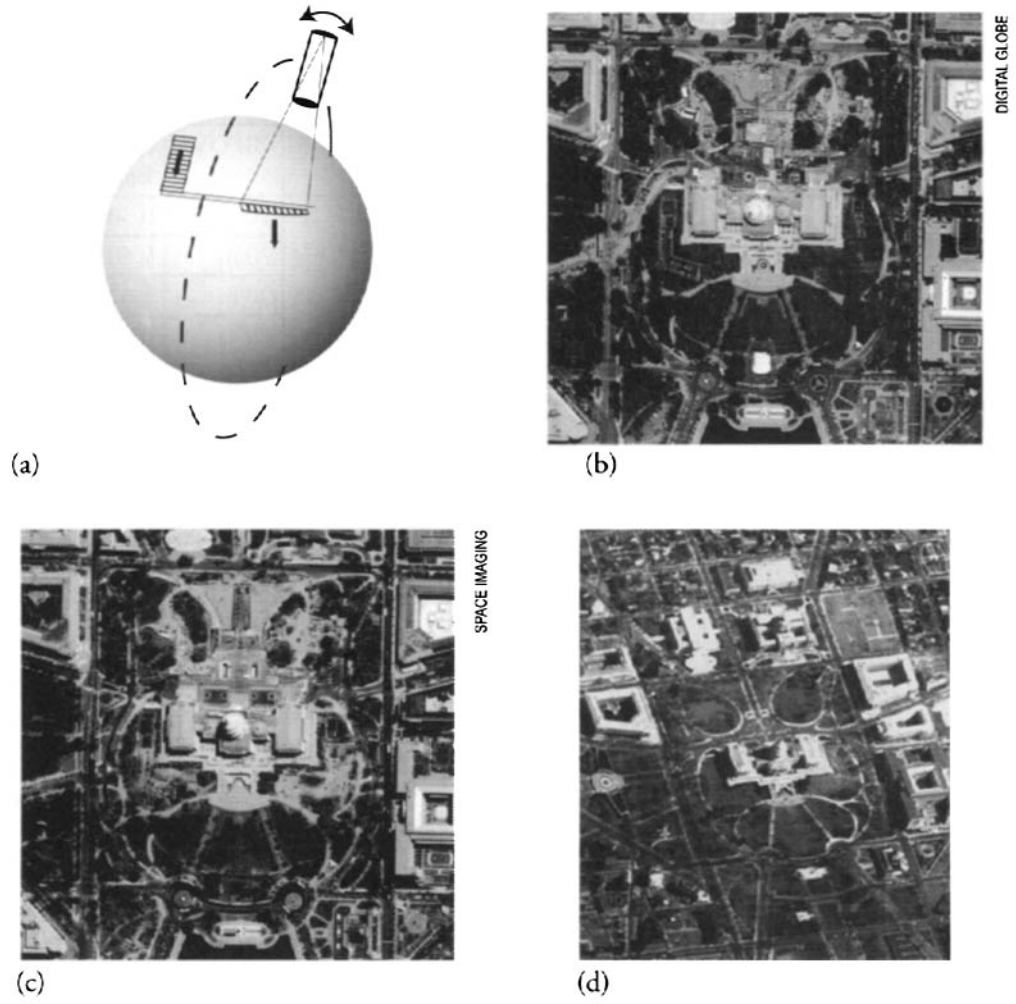


Figure 6.16 Illustration of body pointing push-broom concept (a) and example images:(b) Quickbird panchromatic image of the U.S. Capitol; (c) pan sharpened (see Chap. 12 for discussion of pan sharpening) Ikonos color image of the Capitol; (d) for comparison, a Corona "spy" satellite photographic image from February 19, 1966 is also included. See color plate 6.16.

in this basic design. The wider field of view normally required for airborne sensors and the high cost of linear arrays at longer wavelengths have limited the use of push-broom systems primarily to space systems. However, this is not a fundamental limitation, and with decreases in detector costs, more airborne push-broom systems can be expected (cf. Fig. 6.17).

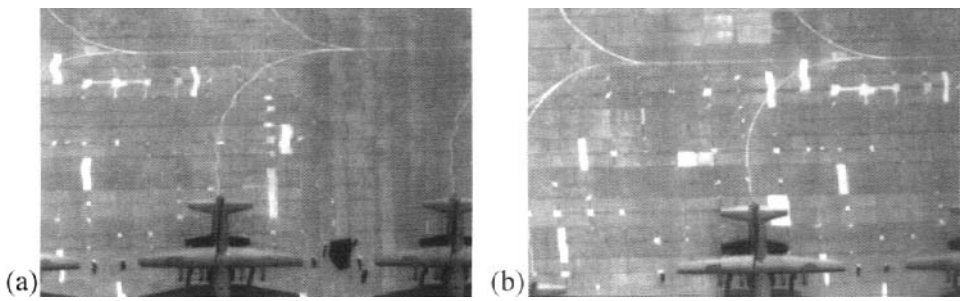


Figure 6.17 Visible images acquired with an airborne push-broom scanner. Note some residual roll effects in the lines painted on the runway in image (a).

6.1.4 Framing (2-D) Arrays

Another EO imaging approach is to use a two-dimensional (2-D) array of sensors and either a mechanical or an electronic shutter to control the integration (dwell) time. This is essentially the way a conventional film camera works with the 2-D array replacing the film. Since the forward motion of the sensor will blur the image, exposure times with this type of sensor must be restricted to less than the time it takes to move the sensor one GIFOV. As a result, there is no inherent gain from a signal-to-noise standpoint over a push-broom system. The major advantage of this approach is in terms of the geometric fidelity. Since the entire image is acquired simultaneously, distortions due to within-frame sensor motion are essentially eliminated. For this reason, 2-D array sensors are very attractive for aircraft use where platform stability is a serious concern. Two-dimensional detector arrays are still difficult to manufacture at many wavelengths, and large arrays (greater than 4K × 4K), even in the silicon region, become quite expensive. Another limitation of 2-D arrays is that it is difficult to obtain multispectral data. Figure 6.18 shows two approaches that are used to overcome this limitation. The sampling approach shown in Figure 6.18(a) is essentially the method used in solid-state color video cameras and amateur digital cameras. A filter mask is layered on top of the array such that different detector element (pixels) are sensitive to different spectral bands. Each spectral image is then produced by interpolating the values for the missing pixels (e.g., replacing those sampled in a different band), often taking advantage of the spectral correlation between bands in the interpolation process. This method reduces the spatial resolution by dividing the array among the spectral bands. An alternative approach shown in Figure 6.18(b) is to use multiple arrays. The incident image is collimated and divided by dichroic beam splitters into three components (e.g., green, red, and NIR) that are each imaged onto a separate 2-D array. The arrays must be very carefully aligned relative to each other to ensure proper registration when the multispectral images are combined [misregistration can be taken out in digital preprocessing (cf. Sec. 12.3) but this can be costly, and the resampling results in image degradation]. Clearly, the beam splitting loses a considerable amount of energy and the alignment must be very precise. In general, this approach is only used when a few (usually three) bands are required for low- to modest- resolution systems (usually video quality with a few hundred by a

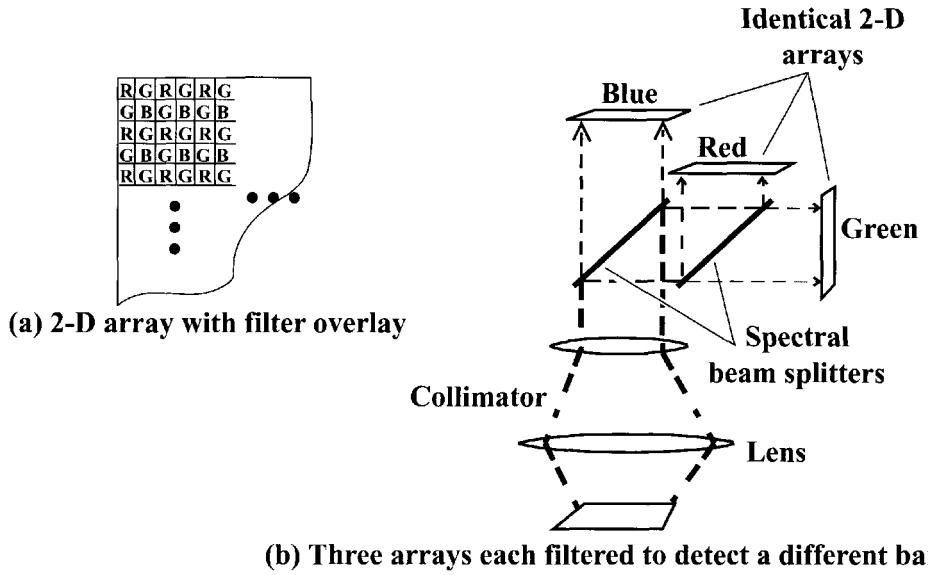


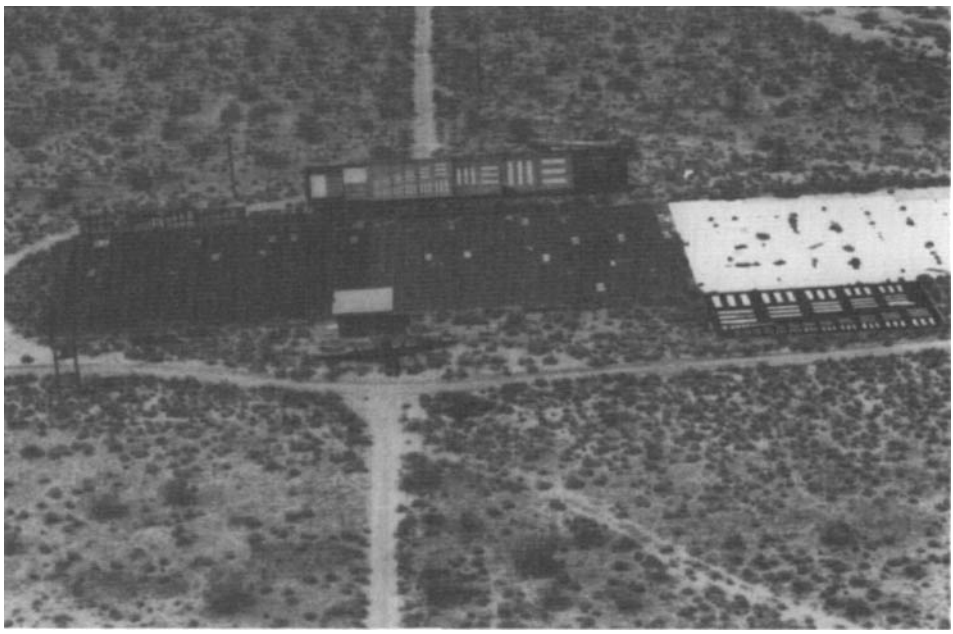
Figure 6.18 Multispectral framing sensor designs. (a) A 2-D array with filter overlay that is interpolated to from a multispectral image. (b) Multiple arrays used with beam splitters to simultaneously image in three spectral bands.

few hundred element detector arrays). In many cases, these 2-D arrays are used to generate video compatible images at 30 frames per second. The output signals are processed to look like conventional NTSC video for ease of storage on commercial videocassette recorders (VCRs). Figure 6.19 shows examples of aerial images obtained with 2-D–array sensors.

While framing systems are extensively used for aircraft applications to avoid geometric distortion problems, they have also been used in space. The *return beam vidicons* (*RBV*s) flown on Landsat 2 and 3 used a 2-D photosensitive array that was exposed and then raster-scanned with an electron beam in a fashion similar to conventional video. The RBVs on Landsat 3 had approximately 40-m resolution and used a reseau grid (i.e., a pattern of fiducial marks) to help ensure geometric fidelity. In the Landsat 3 era, this represented the highest-quality civilian data available for large-area cartographic mapping from space. Figure 6.20 shows an RBV image and the corresponding portion of a Landsat MSS scene showing the improved image quality that, coupled with the geometric fidelity, increased the value of the RBV products for mapping purposes.

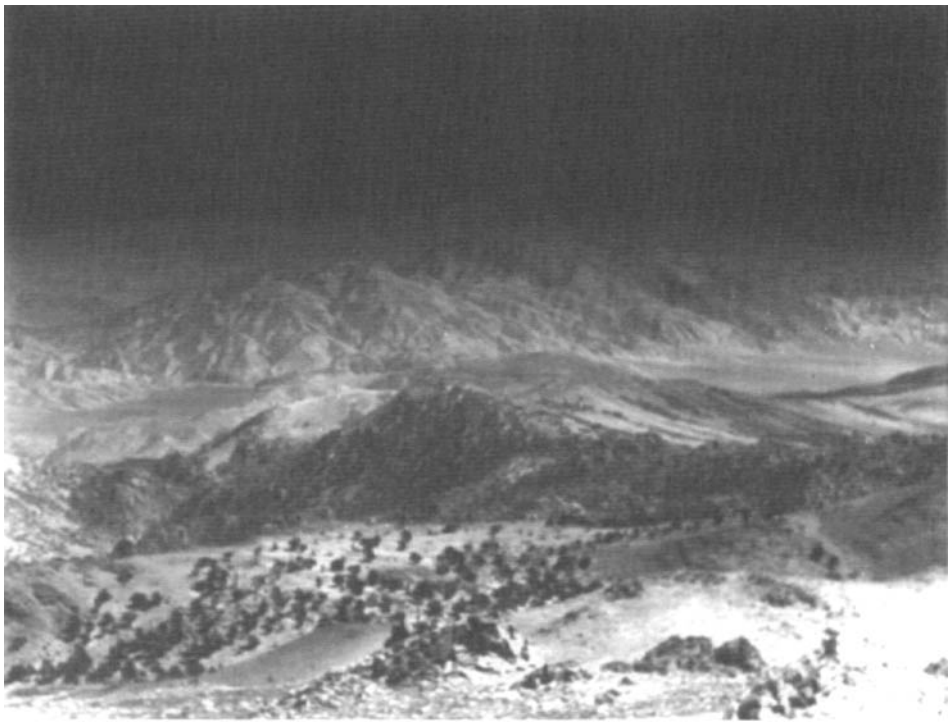
6.2 IMAGING SPECTROMETERS

Advances in one- and two-dimensional array detector technology, coupled with advances in digital data processing and storage, have opened the door for a new generation of sensors called *imaging spectrometers*. These sensors are designed to collect data on the spectral as well as the spatial characteristics of the imaged scene. Their use is motivated by the expectation that finer spectral sampling will convey more information about the makeup or condition of certain scene elements. Im-



RECON/OPTICAL

(a)



EASTMAN KODAK COMPANY

(b)

Figure 6.19 Examples of EO images acquired using 2-D arrays. (a) Visible image of test target acquired with a 2048 × 2048 array. (b) MWIR image acquired with a platinum silicide array (note the cold sky in the background).

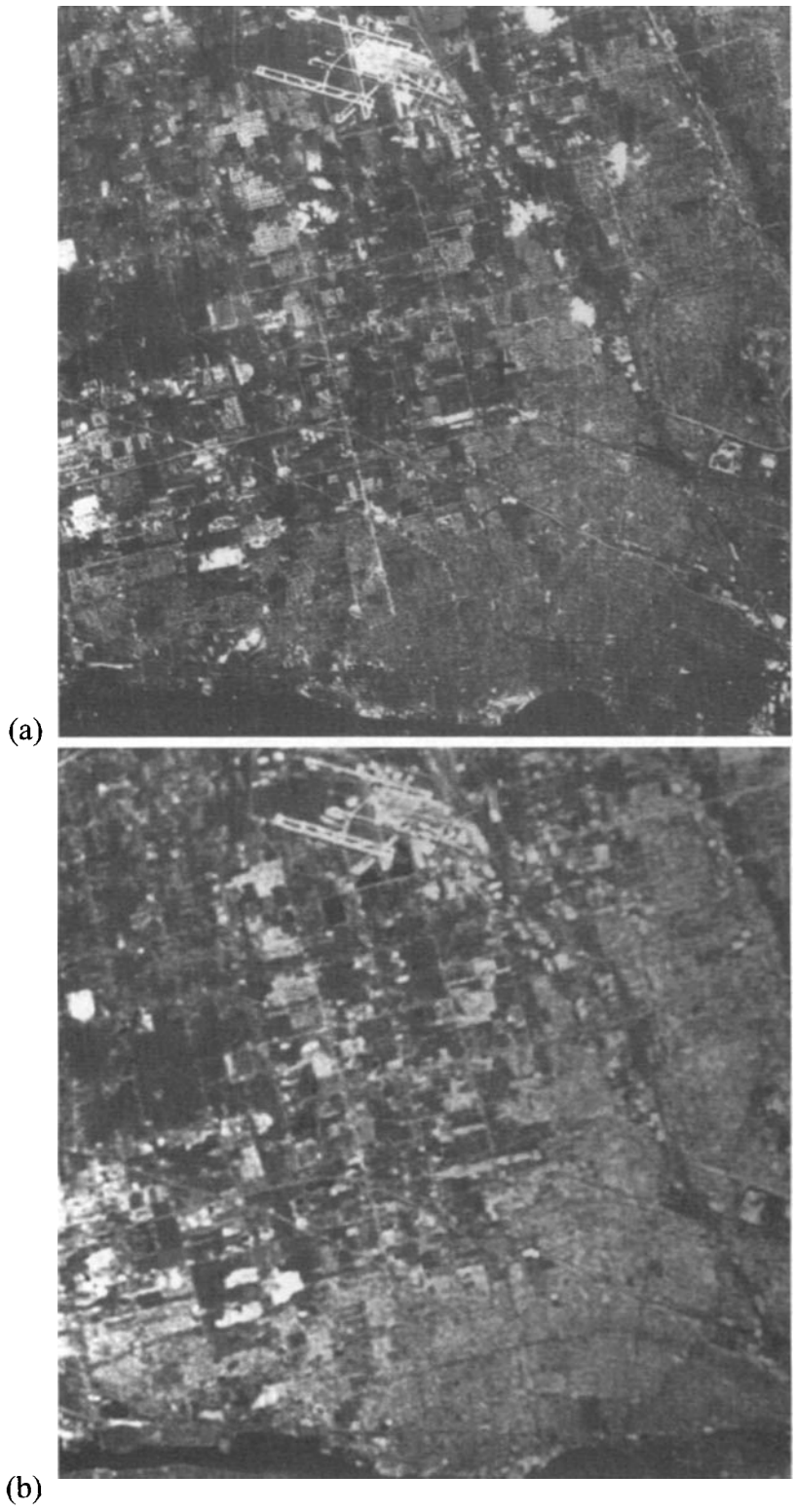


Figure 6.20 (a) Landsat 3 RBV image of Detroit and (b) a portion of Landsat MSS image of the same area showing the increased potential of the RBV for cartographic applications.

aging spectrometry brings to the forefront the fundamental tradeoffs in all sensor designs between spatial, spectral, and radiometric resolution. The limited numbers of photons from a scene can only be divided into a restricted number of information bins. We can think of these bins as spatial samples (i.e., number of pixels), spectral samples (i.e., number of spectral bands), and radiometric samples (i.e., number of gray levels). In most cases, for a fixed technological approach, it takes more photons to increase the resolution (i.e., number of bins) in any category, and the only way to obtain more photons is to steal them from one of the other bins. For example, to increase the spectral resolution, we must sacrifice signal-to-noise or spatial resolution. At present, these tradeoffs are largely made in an ad hoc fashion and depend on the scenario under study and the analytical algorithms being used. Hopefully, as more imaging spectrometer data become available, a practical approach to these tradeoffs can be developed that can help system designers and image analysts decide what spatial, spectral, and radiometric specifications are required for a specific analysis. In the meantime, system designs are attempting to push the limits in all three areas with a major emphasis on spectral resolution while we explore the potential of this relatively new dimension to remote sensing.

Before addressing "true" spectrometers, we should discuss one very important instrument that begins to bridge the gap between traditional space-based multispectral data and the many contiguous spectral bands of a full spectrometer. This is the moderate-resolution imaging spectrometer (MODIS) instrument [cf. Guenther et al. (2002)], which has flown on both the Terra (1999 launch) and Aqua satellites (2002 launch). MODIS is actually a 36-band multiple-line scanner acquiring data with a continuously rotating two-sided mirror to reduce dead time. Like the Landsat ETM+, it uses different numbers and sizes of detectors in the along-track direction to acquire data with a range of GIFOVs at nadir from 250 m (2 bands), through 500 m (5 bands), to 1000 m (29) bands. With a 10 km in track swath width, this requires 40, 20, and 10 detector elements per band at each of the three respective resolutions. By scanning ±55° from nadir, MODIS generates a 2330 km ground swath that enables full global coverage every 1–2 days based on its sun synchronous orbit. The 36 spectral bands for MODIS are spread over four different focal planes and use spectral filters to obtain the desired spectral sampling with the 36 bands covering the visible through the thermal infrared. MODIS data represent the first readily available global database of many spectral band data. Its regular repeat cycle means that cloud-free data of essentially anywhere in the globe can be acquired within a short time scale to allow temporal monitoring of the planet. With most of the spectral bands at 1 km (at nadir), MODIS has been most widely used for global or regional studies that take advantage of its temporal and spatial coverage (cf. Fig. 6.21). We describe MODIS as a transition instrument to a full spectrometer because it has discrete spectral bands rather than sampling of a spectral region with contiguous bands, which we will associate with a spectrometer. Nevertheless, because of the ubiquitous nature of the MODIS data, the fact that the *National Polar-orbiting Environmental Satellite System* (*NPOESS*) system will secure a long-term continuity of "MODIS-like" data, and the applicability of many of the spectral algorithms described in Chapter 10 to MODIS data, we have included

Figure 6.21 True color MODIS image of the East Coast of North America acquired March 6, 2000, showing the potential of MODIS to support global and regional Earth resources and Earth process studies. See color plate 6.21.

it in the spectrometer category. Indeed, in the near term, for many users it may be the only source of nearly spectroscopic data available for a study site.

The design approaches for "true" spectrometers are largely just extensions of traditional designs that take advantage of array detector technology. The most

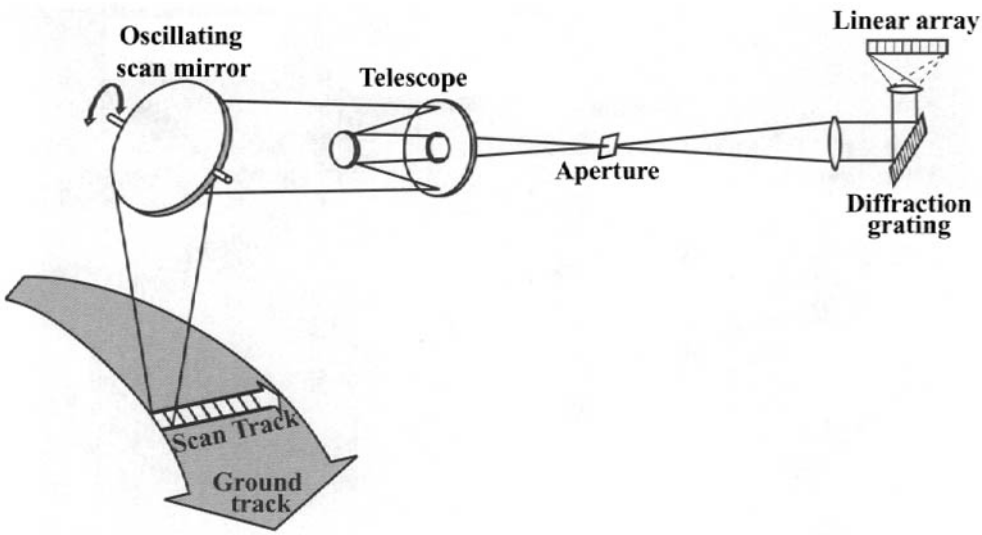


Figure 6.22 Conceptual diagram of an imaging spectrometer line scanner.

straightforward approach conceptually employs a line scanner with one or more monochromators whose entrance aperture(s) is located at the telescope's principle focal plane. The dispersed spectra are then sampled with linear array detectors, as illustrated in Figure 6.22. This is essentially the same design as the multispectral scanner shown in Figure 6.7 using a linear array spectrometer. NASA's advanced visible and infrared imaging spectrometer (AVIRIS) is a good example of this approach. In order to cover the spectral region from 0.4 to 2.5 μm in channels with approximately 0.01 μm spectral bandwidth on 0.01 μm spectral centers, it uses four spectrometers, one with a silicon array and three with InSb arrays [cf. Vane et al. (1993)]. This is accomplished by using fiber optics to carry the flux sampled at four points in the along scan direction on the focal plane to the entrance aperture to four monochromaters. The four regions of the spectrum recorded must then be spatially shifted and registered to form the spectrum of a given pixel. The AVIRIS instrument has a 1 milliradian IFOV resulting in 20 m pixels when flown from 20,000 m in an ER-2. (It can also be operated down to several km resulting in 2-4 m pixels.) Figure 6.23 is a sample image from the airborne AVIRIS sensor. The 2-D surface image was made by assigning the red, green, and blue brightnesses to three of the 224 spectral channels. The edge pixels have their spectra from 0.4 to 2.5 μm, color-coded to provide the depth axis of the "*image cube.*" Note that the black bands in the spectra are the atmospheric absorption bands around 1.4 μm and 1.9 μm. On a soft copy computer display, the location of the edges of the color cube can be interactively adjusted to let an analyst search spatially for spectral patterns. This is a useful quick-look tool. However, more detailed analysis is typically required using advanced computer analysis methods, since the human being is quickly overwhelmed by the volume of data associated with spectrometric images (cf. Chap. 11).

The AVIRIS sensor's primary function is to serve as an airborne test bed for satellite-based imaging spectrometers. It led the way, for example, for the MODIS

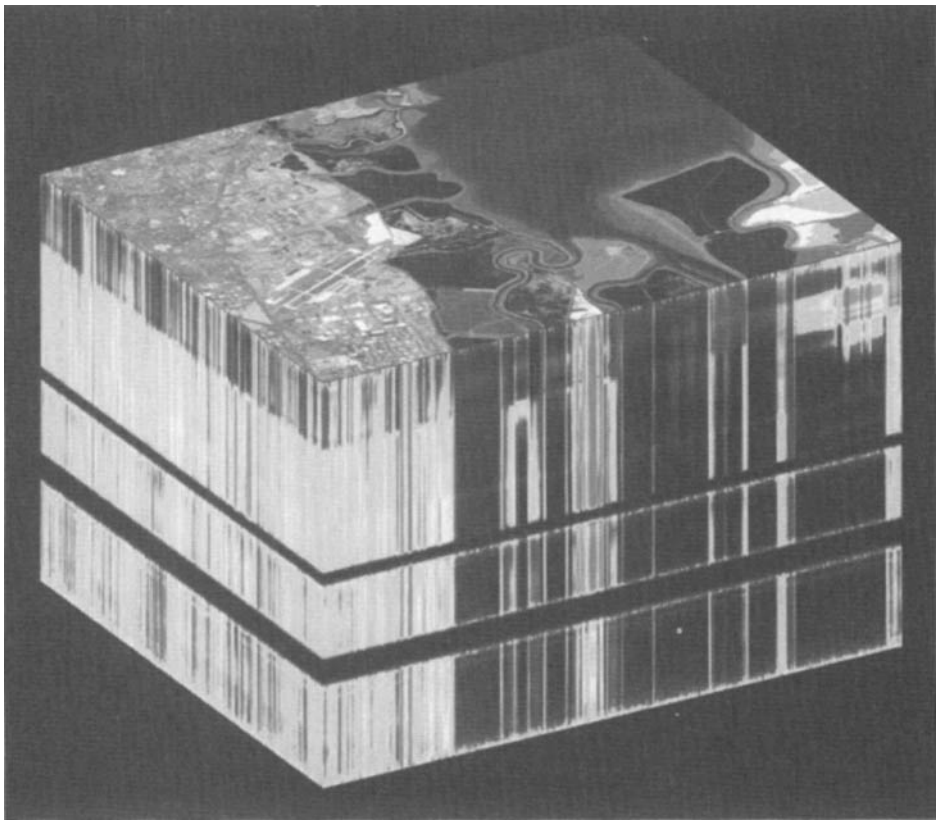


Figure 6.23 AVIRIS image cube of Moffet Field in California. The sensor has 224 channels, from 0.4 μm to 2.5 μm, each with spectral bandwidth of approximately 10 nm. See color plate 6.23.

instrument discussed above. In its role as test bed, the AVIRIS instrument has been modified many times, and AVIRIS imagery has served as the data for test and development of many new instrument calibration and data analysis approaches (cf. Proceedings of the AVIRIS workshops 1988-2005). As a result, we will use AVIRIS as the example in many of the discussions of instrument calibration, atmospheric correction, and spectral image analysis in the following sections and chapters The line scanner design for imaging spectrometry has all the strengths and weaknesses of conventional multispectral scanners discussed in Section 6.1. In particular, the short dwell time, coupled with the narrow spectral bands, means that the signal-to-noise or the spatial resolution (or both) must be sacrificed.

An alternative imaging spectrometer approach uses a push-broom design, as shown in Figure 6.24. In this design, an entire line of data is imaged onto the focal plane and sampled simultaneously by a slit that is the entrance aperture to a monochromator. The beam is spectrally dispersed perpendicular to the long axis of the slit and imaged onto a 2-D array. One dimension provides the spatial sampling in a traditional push-broom fashion; the other dimension provides the spectral sampling. NASA's Airborne Imaging Spectrometer (AIS) used this push-broom 2-D–array approach employing multiple monochromators to cover a spectral range from 0.4

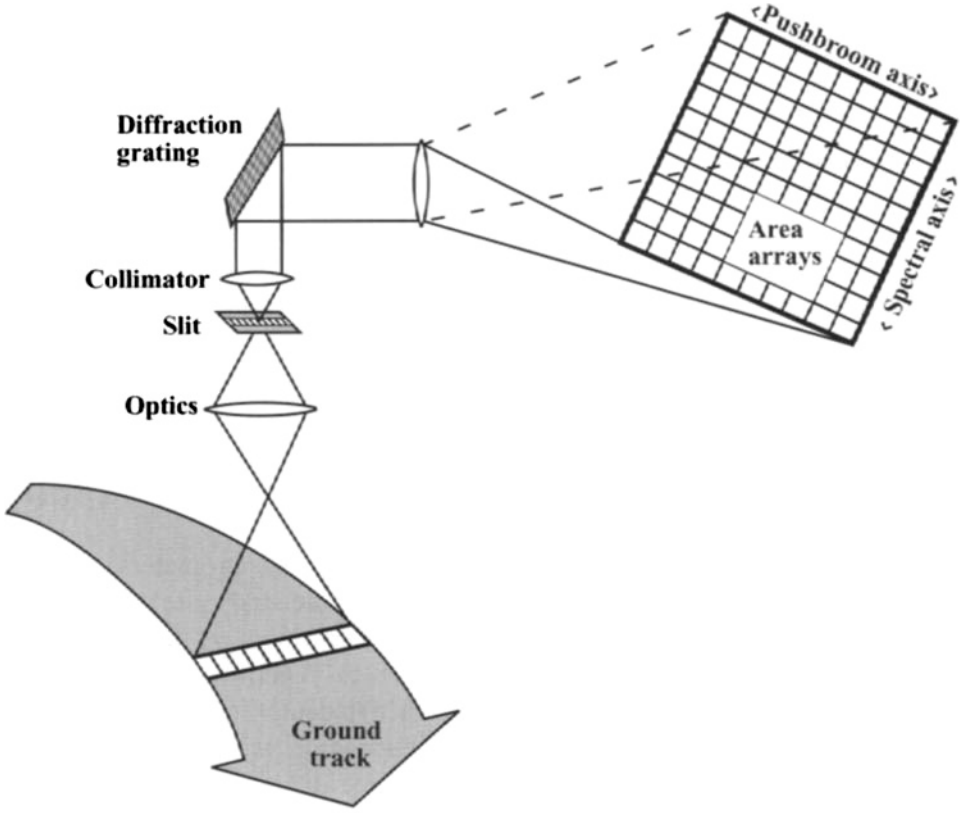


Figure 6.24 Conceptual illustration of an imaging spectrometer using a push-broom design.

to 2.5 μm [cf. Goetz et al. (1985)]. Two-dimensional-array detector technology has been a limiting factor with this approach. The push-broom axis needs to be quite large for many applications, and large 2-D arrays are still difficult to manufacture (especially in the IR bands). For example, AIS produced images that were only 64 pixels wide. As 2-D IR arrays became more available, the push-broom approach has been more widely used to take advantage of increased dwell time.

The hyperspectral digital collection experiment (HYDICE) was an early demonstration of an advanced push-broom imaging spectrometer that used a 2-D focal plane whose sensitivity extended from the visible through the short-wave infrared (0.4–2.5 μm) [cf. Rickard (1993)]. The 320-column, 210-row cryogenically cooled indium antimonide (InSb) focal plane was manufactured with three separate regions in the spectral (row) direction to achieve high sensitivity across the full spectral range. The spectral dispersion was achieved with a prism monochromator, resulting in an instrument with a spectral resolution of nominally 10 nm sampled on 10 nm centers. (Note the actual bandwidth is a function of wavelength and varies from approximately 5-to-15 nm). The HYDICE instrument achieved quite high spatial resolution (0.5 milliradian flown as low as 2 km); however, its coverage at these resolutions is very limited due to the small number (312) of usable detectors in the cross-track direction.

The Spatial Enhanced Broadband Array Spectrograph System (SEBASS) was another early 2-D array sensor that demonstrated the potential to perform push-broom imaging spectroscopy in the MWIR and LWIR [cf. Hackwell et al. (1996)]. The SEBASS instrument uses two 128 × 128 arrays cooled (along with the entire prism-based spectrometer) to 5K using liquid helium. The incident beam from the sampling slit at the telescope focal plane is split with a dichroic beam splitter to send the MWIR and LWIR wavelengths to two monochromaters. The result is 128 spatial pixels by 128 MWIR spectral pixels covering the range from 2.6-to-5.7 μm on 0.03 μm centers and 128 LWIR pixels from 7.5-to-13.5 μm on 0.05 μm centers. The SEBASS instrument has an IFOV of 0.5 milliradians and can operate as low as 1.5 km yielding very high spatial resolution, but with a very narrow swath width. Along with AVIRIS, the HYDICE and SEBASS sensors have been extensively used to study the potential for imaging spectroscopy over various spectral ranges, spatial resolutions, and signal-to-noise constraints (cf. Chaps. 10 and 11 for a discussion of imaging spectrometry algorithms).

A major limitation of the one of a kind airborne spectrometers described above is the difficulty in obtaining imagery of a specific study site or for a wide range of global phenomena. The *Hyperion* instrument flown on the EO-1 spacecraft represents an early effort to address this need for repetitive global coverage. The EO-1 mission (launched in 2000) was designed to demonstrate technology that might be used in future missions in the Landsat series. It was flown in formation with Landsat 7 (1 minute behind) and carried three demonstration instruments. The Advanced Land Imager (ALI) is a push-broom instrument designed to demonstrate a push-broom alternative to the ETM+ reflective bands with higher signal to noise achieved with the push-broom approach. The Hyperion instrument, in contrast, was designed to demonstrate that spectrometric image data could be obtained from space at Landsat spatial resolutions (30 m ground spot) [cf. Folkman et al. (2001)]. The Hyperion instrument employs two push-broom grating spectrometers. After passing through a slit on the primary focal plane, the beam is split with a dichroic beam splitter and passes through two grating spectrometers [one VNIR (70 bands) and one SWIR (172 bands)]. The resulting spectra are merged to form an image cube 250 pixels wide (7.5 km in the ground plane) by 220 spectral bands deep that can be push-broomed for hundreds of kilometers. By placing multiple spectrometers along side each other, wider swath widths should be achievable with this design in the future.

The third instrument flown on the EO-1 mission was the Linear Etalon Imaging Spectrometer Array (LEISA) Atmospheric Corrector (LAC). It represents a significantly different approach to imaging spectrometry (cf. Fig. 6.25). In this design, a linearly variable etalon (LVE) or wedge filter is placed in front of a 2-D array at the primary focal plane (i.e., without the filter in place, a full 2-D image of the scene would be formed). The LVE is essentially a narrow-band interference filter that only passes a narrow spectral band at any thickness location in the wedge. By changing the spacing between the interference layers (the wedge concept), the center wavelength passed is advanced across the filter and, therefore, the focal plane. Thus, each row of the array sees a different row of pixels on the ground in a different spectral band. With the instrument advance and detector readout properly

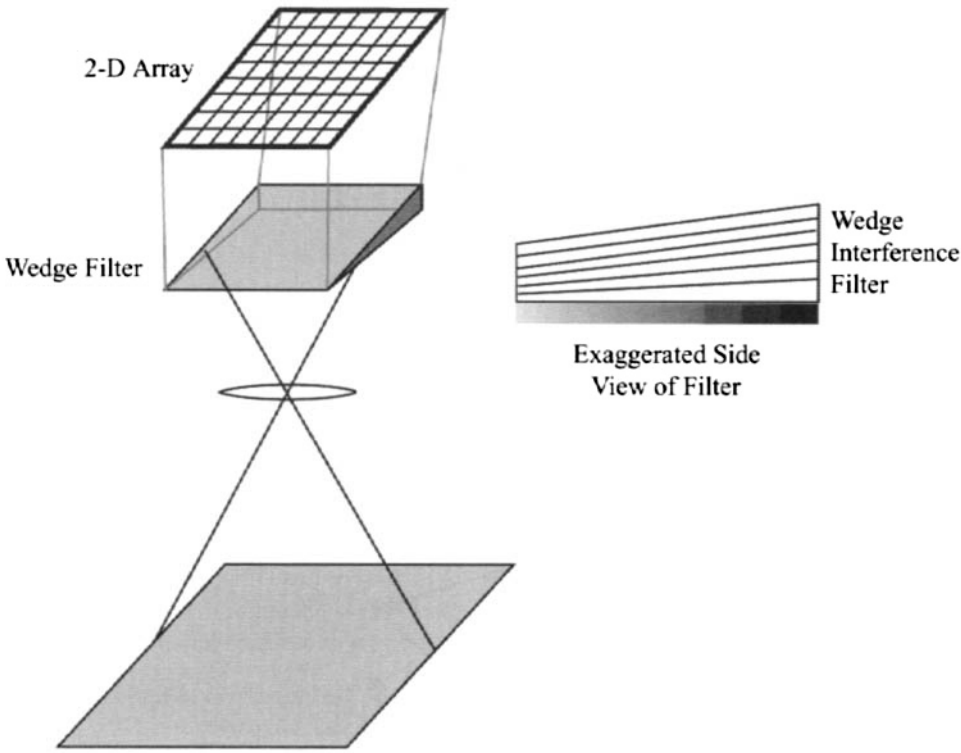
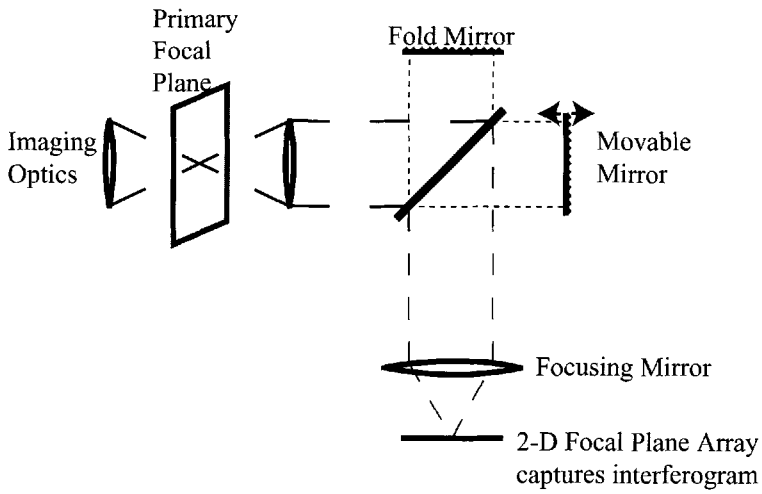


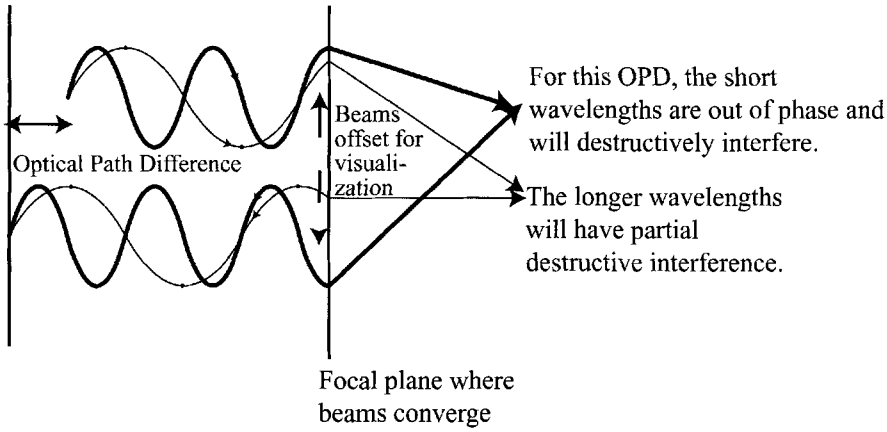
Figure 6.25 Illustration of the wedge filter concept for imaging spectrometry.

synchronized, the next sample from the array will advance the projection of a row of ground samples to the next row of detectors and, therefore, the next spectral band on the focal plane. Thus, to form a spectrum corresponding to a single pixel on the ground, we must separately shift and register down the entire length of the array in the spectral dimension. Clearly, this approach places considerable burden on the stability of the platform. In the case of the LAC, which was being demonstrated for atmospheric correction where spatial resolution was not an issue, this did not represent a critical limit given its 250 m ground samples and 256 spectral bands (0.89–1.58 μm).

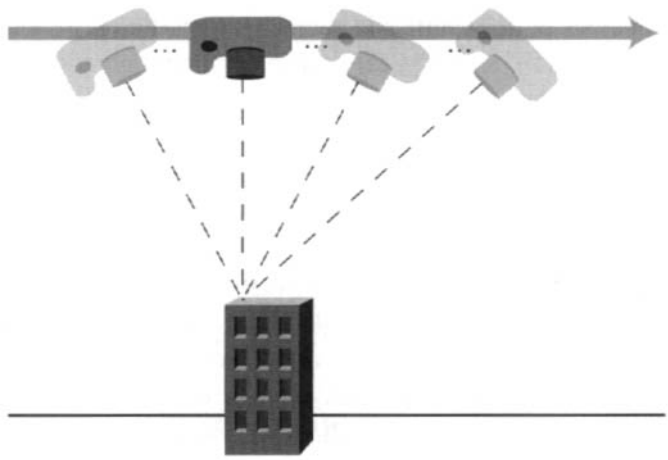
Another group of imaging spectrometers often used for very high spectral resolution use Fourier transform methods to invert interferograms to high-resolution spectra. One approach involves collimating the image beam as it enters a Michelson interferometer (Tuzman-Green configuration) and bringing the interferogram to focus on a two-dimensional array where each detector samples the interferogram corresponding to a discrete spatial location on the ground [cf. Wolfe (1997)]. Figure 6.26 shows a simplified drawing of the interferometric component of an imaging *Fourier transform spectrometer* (FTS). If we follow the energy from one point on the primary focal plane through the system, we see that the collimated beam is split. The upward path goes to a fixed mirror and returns to the beam splitter, and the right-hand path goes to a movable mirror and returns to the beam splitter, which at this point becomes a beam combiner. The combined beams are brought to focus on a point on the detector array, where they will interfere due



(a)



(b)



(c)

Figure 6.26 Conceptual illustration of a Michelson Fourier transform imaging spectrometer:(a) Simplified Michelson spectrometer shows ray paths. (b) An example waveform for a signal composed of only two wavelengths showing the phase offset due to optical path difference and the resulting differential interference at the focal plane. (c) Illustration of how the entire imaging device must ideally remain focused on the target as sequential temporal samples are acquired as the mirror moves.

to the wavelengths of EM energy present and the path difference between the two paths. Each motion of the mirror changes the path difference and forms a different signal level due to the changing interference pattern. By sampling the signal as a function of time, an interferogram is produced. Each pixel in the 2-D array will form over time an interferogram of a signal from a different location on the primary focal plane (i.e., a different location on the ground). If we stack these time samples to form a interferogram cube and then take a one-dimensional Fourier transform in the z direction (i.e., for the signal acquired by a single detector element), the result can be converted into a spectrum with the resolution controlled by the number of path differences sampled (i.e., the number of mirror locations). Clearly, the temporal nature of this approach presents problems and requires the 2-D sensors to attempt to stare at a fixed point on the ground during the full acquisition of the interferogram. The Sagnac FTS, which generates a spatial interferogram and operates in a push-broom mode, offers an alternative design [cf. Yarbrough et al. (2002) or Meigs et al. (1996)]. Both of these general Fourier transform designs have a number of subtle limitations that regrettably are beyond our scope in this treatment, and the reader should consider the relevant literature carefully.

6.2.1 Imaging Spectrometer Issues

The discussion, thus far, of imaging spectrometers has tried to provide an overview through example sensors of some of the common instrument approaches and provide a general background on the sensors that have generated some of the most accessible data. In this section, we want to discuss some of the subtler issues that relate to the integrity of the data from imaging spectrometers, particularly as it may eventually impact the performance of image analysis algorithms as discussed in Chapters 10 and 11.

One of the first potential areas of concern with imaging spectrometers is the potential to have significantly poorer radiometric resolution as expressed in terms of signal to noise. If we simply compare a few band multispectral sensor to a spectrometric sensor, we might well reduce the spectral bandpass by an order of magnitude (e.g., from 100 nm to 10 nm). If everything else between the two systems were equal (which it never is), we could expect one-tenth the signal and, if the noise were additive, a factor of 10 decrease in the signal to noise. In fact, the throughput of a monochromator is often less than the broad-band filters used in multispectral systems, which would exacerbate this problem. Most imaging spectrometers compensate for the reduced photon flux by employing faster optics (lower F#), larger detectors (i.e., lower spatial resolution), longer dwell time, or some combination of these to boost signal to noise to acceptable levels. As we will see in Chapter 10, spectral analysis algorithms are often searching for very subtle differences between spectra, which may, in some cases, demand even higher signal-to-noise performance than we may accept from panchromatic or multispectral systems.

Another concern with spectrometers is the location of the band center, band edges, and characterization of out of band response. For sensors with band widths of many tens to hundreds of nanometers, uncertainties of a few nanometers in the location of a response function typically result in relatively minor radiometric er-

rors. On the other hand, when the width of the response function is only a few nanometers and the spectral features of interest are of comparable size, similar uncertainties of a few nanometers can result in significant radiometric errors and fundamental changes in the shape of features in the observed spectrum. Figure 6.27 shows an example of how this error in knowledge of spectral response location could significantly influence the observed signal from the potassium signature from forest fires described by Vodacek et al. (2002). Also of concern with some spectrometer designs is the out-of-band response. Most designs have relatively well-behaved spectral response functions that fall quickly to near zero outside the nominal bandpass. However, some designs may have significant response from immediately adjacent regions or may even allow leakage from a spectral region somewhat removed from the band of interest.

A very common source of variation in spectral response in imaging spectrometers results from the slight curvature of the image of the sampling slit when projected onto the 2-D array used with both prism and grating push-broom imagers. This effect, illustrated in Figure 6.28, is referred to as *spectral smile* (perhaps more appropriately as spectral frown) because of the curved shape of a monochromatic feature on the focal plane. Its impact is clearest if we think about forming a single-band image from any given row of a push-broom spectrometer array. Ideally, that image could be treated as characterizing the spectral content of the scene in a single spectral band with a common spectral response for every pixel. In fact, the spectral smile causes the spectral response to change from column to column across the image. The band center and band edges drift slightly across the scene (usually the bandwidth is not significantly affected). This means that identical spectral signatures may manifest themselves differently at different locations in the across track direction (cf. Fig. 6.27). Clearly, this type of error can be a problem if it is not characterized (i.e., if we don't know about it); however, it can also be problematic even if fully characterized, in that with severe cases we may have to adapt an algorithm to adjust for the spectral response changes across a scene. The spectral smile effects may be combined with geometric distortions caused by misalignment of the detector array on the focal plane of the monochromator and/or by spectral projection (keystoning) effects, as illustrated in Figure 6.28.

The discussion of keystoning and misalignment raises another common concern with image spectrometers often referred to as spectral purity (although it is actually a spatial issue). Ideally, a spectrum formed by drilling down through a hypercube at any pixel location would only contain spectral information about that location (i.e., the spectrum would be pure). In fact, as we see in the keystoning and misalignment examples in Figure 6.28, as we sample a spectrum down a column in a push-broom array, we may begin to have spectral information from adjacent pixels contaminating the spectrum at certain wavelengths. This problem becomes a driving force in the wedge filter design shown in Figure 6.25, where we must register each line of data to the previous one over all spectral samples to form a single spectrum. Even in the line scanner approach, where the spectra are largely forced to be spectrally pure by having a single common sampling aperture, problems can arise when, for practical purposes, multiple spectrometers and sampling apertures are employed. For example, in the AVIRIS case, each spectrometer has its own

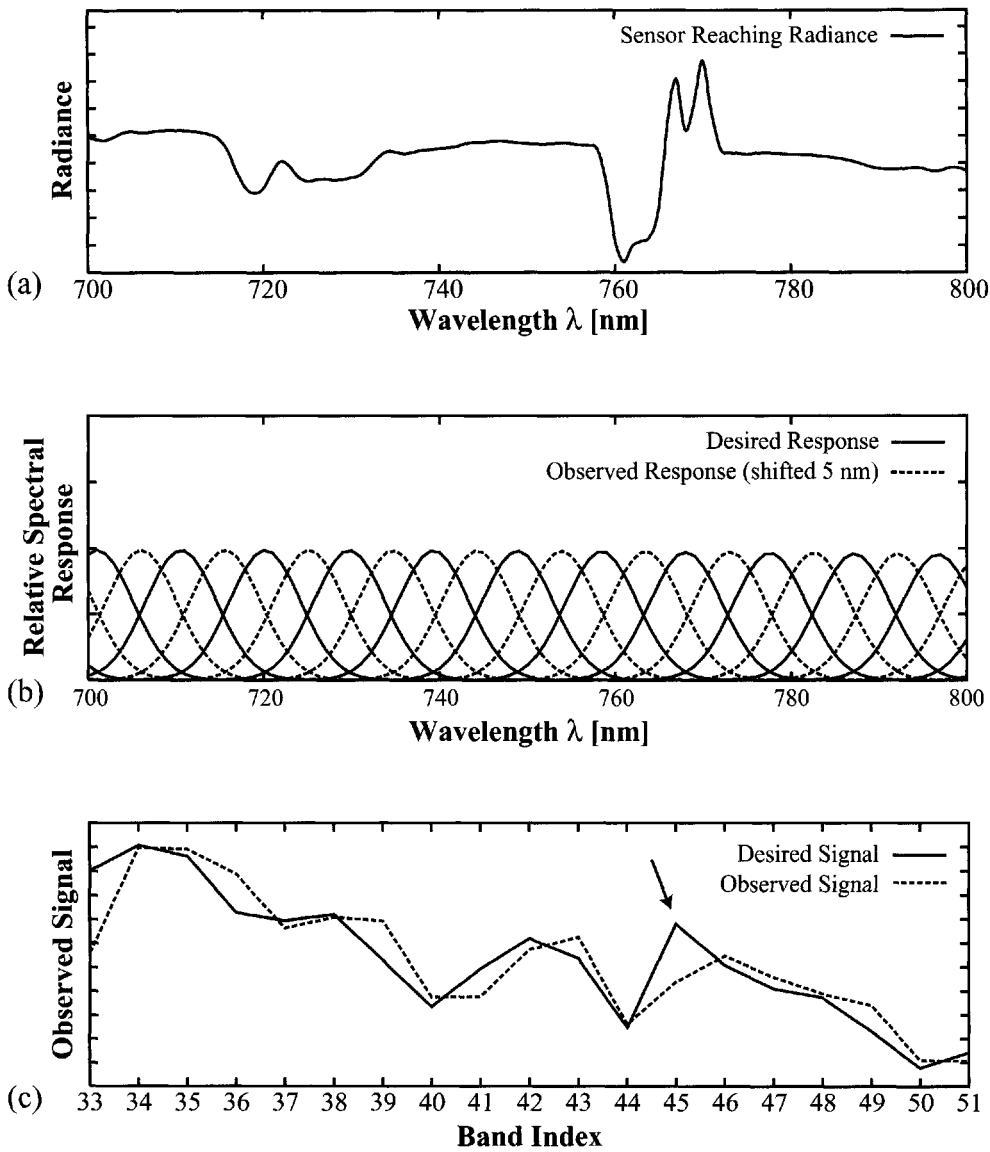


Figure 6.27 Illustration of the effects of error in spectral response on observed spectral shape. In this case, the target spectrum includes the narrow potassium emission line near 770 nm associated with burning wood adjacent to the 760 nm oxygen absorption feature as shown in (a). The desired (designed) spectral response of spectral channels of a spectrometer is shown in (b) along with the actual spectral response that might unknowingly exist without careful characterization. Finally, the output signal obtained by convolution of each band in (b) with the target spectrum in (a) yields the desired and observed spectra in (c), illustrating how our signature could be significantly suppressed without our knowledge.

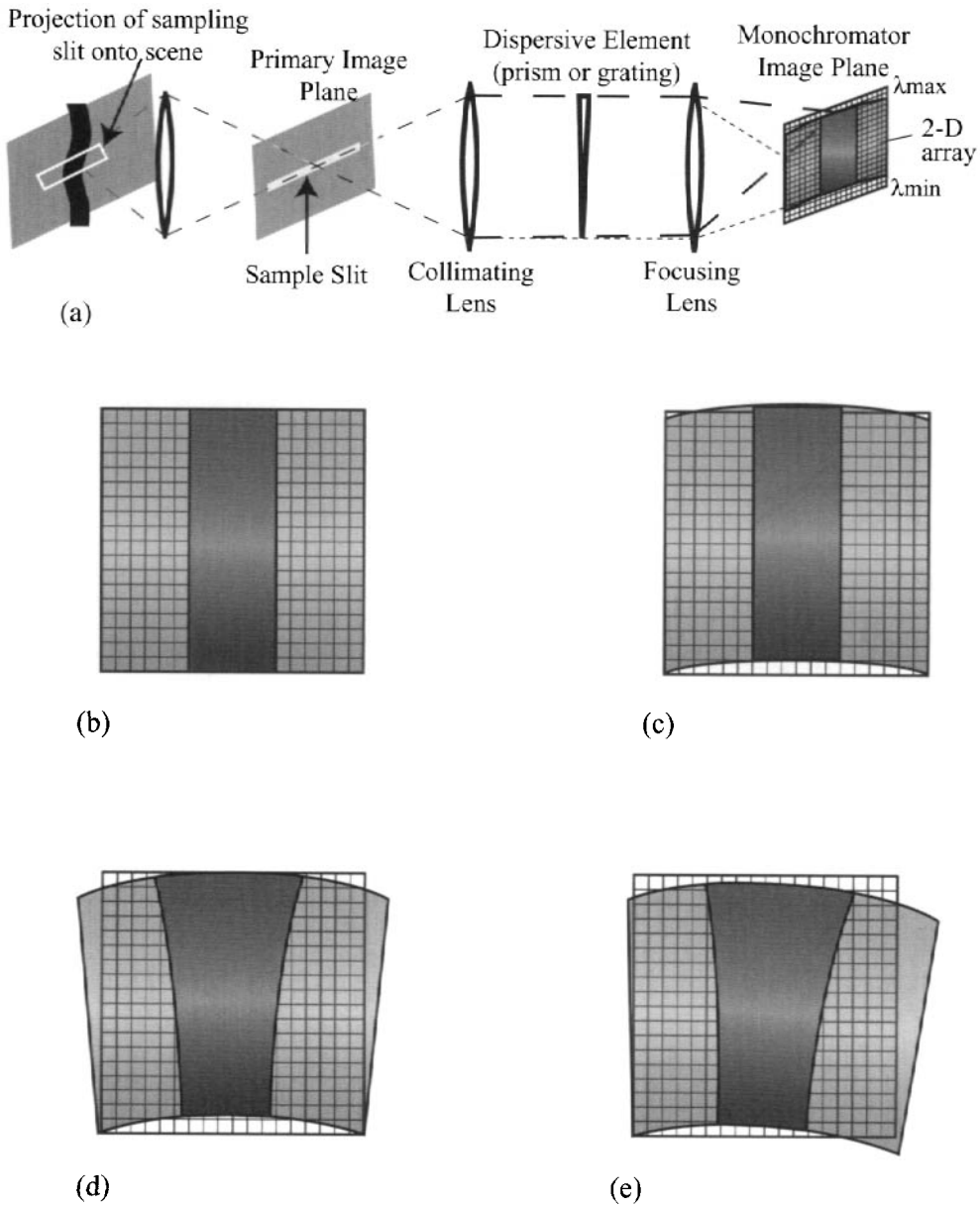


Figure 6.28 Illustration of spectral smile concept and the confounding effect of array misalignment (tilt). (Note that in many designs the collimating and focusing functions use a common lens and the dispersive element is a reflective component.) (a) Source of smile error (note curvature is exaggerated for illustration purposes). (b) Idealized spectrum projected on detector array, (c) Spectrum showing smile. (d) Spectrum showing smile and key stoning effects due to optical projection and misalignment. (e) Spectrum showing smile, key stoning, and tilt due to array misalignment. See color plate 6.28.

sampling aperture and the four spectral segments must be shifted across track in image reconstruction to form the spectrum.

A final common issue with imaging spectrometer data is its shear volume. These systems generate data files that are 12–14 bits (cf. Chap. 7 for quantization concepts) per sample by several hundred spectral samples by the image pixel dimensions in size. For satellite systems, this represents an enormous bandwidth issue as we move to higher resolution systems. Even for airborne systems, the issues of data storage and data processing can be daunting for operational systems where many images of large areas must be processed regularly.

6.2.2 Agile Spectrometers

One method that has been devised to address the data volume and bandwidth issues associated with imaging spectrometers is to only retain the spectral bands of interest. For example, if we were only interested in one or two spectral features, we might only retain a few spectral bands on or near that feature and a few more for contrast and not store or downlink the remaining bands. This, of course, requires us to know in advance what we are interested in for each image acquisition and which bands will be of interest. In the research domain, this is seldom possible, but for operational programs it may be. However, the approach described above still requires the full capability to collect all data and then the logic to discard the unneeded data. An alternative approach uses what is referred to as *agile spectral sensing* to change how a spectrometer spectrally samples the scene. One such approach uses a *liquid crystal tunable filter* (LCTF) in the imaging path of a framing array to select the narrow spectral band to be sampled [cf. Stevenson et al. (2003)]. The LCTF can rapidly be tuned from one narrow band to another across a fairly broad spectral range. The spectral bands must then be shifted and registered in the same fashion as we described for the wedge filter approach except that now each frame read out of the array is a single spectral band. The LCTF is programmed to sample different numbers of bands and different spectral bands depending on the spectral properties of the targets of interest. As we learn more about the spectral character of targets of interest, this approach may be increasingly popular for operational programs. However, as we will see in Chapter 7, much of the information being extracted from spectral data comes from analysis of subtle spectral variations that may require sampling of substantial amounts of the spectrum.

6.3 LUMINESCENCE SENSORS

As we continue to look at sensors with narrower and narrower spectral lines, a new source of photons can become significant in certain spectral bands. The source of this flux is stimulated *luminescence*. This occurs when a high-energy photon causes an electron in an absorbing molecule to transition from a stable level to a higher energy state. The electron can then return to the lower energy level by giving up the transition energy through collision (thermal energy), by emission of a photon equal to the transition energy (i.e., a photon at the same wavelength as

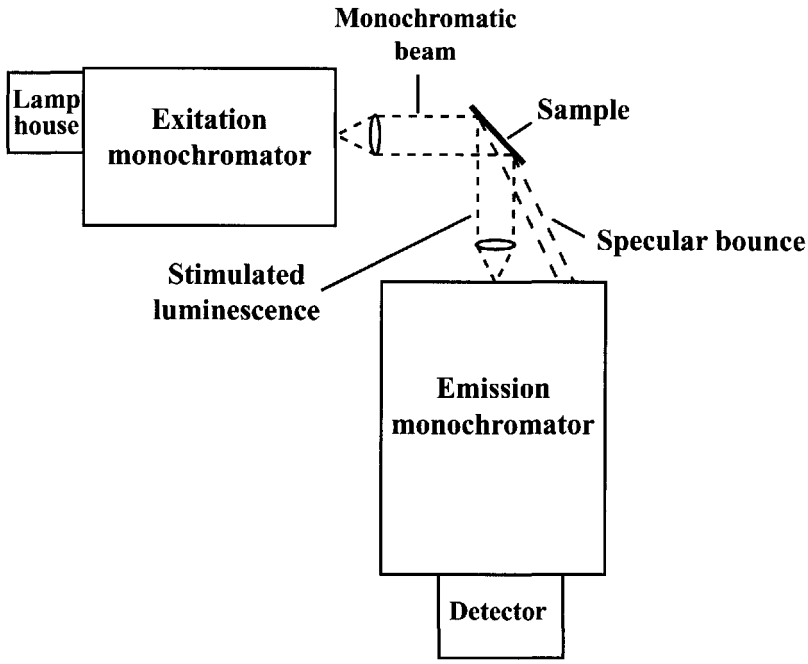


Figure 6.29 Luminescence spectrometer.

the absorbed energy), or by a step process where the decay takes place through two or more energy levels emitting lower energy longer wavelength photons at each step. This last process, known as luminescence, is quite common. However, the flux levels involved are usually quite small and are masked by the reflected flux levels. Luminescence can be measured in the laboratory with a luminescence spectrometer, as shown in Figure 6.29. The sample chamber can contain either a cuvette for liquid samples (most standards are dyes in varying concentrations) or solid samples. The emission spectrometer is generally located at right angles to the beam from the excitation spectrometer. Solid samples are oriented such that any energy from a specular bounce does not enter the emission spectrometer. This reduces stray-light problems. A sample can then be characterized in terms of its emission and/or excitation spectra. The emission spectra are obtained by setting the excitation monochromator at a fixed wavelength (e.g., a laser illumination line that would be used to study laser-induced luminescence) and then scanning the emission monochromator. An excitation spectra is produced by fixing the wavelength of the emission spectrometer (e.g., at a particular line to be sensed by an imaging system) and then scanning the excitation monochromator. Sample spectra are shown in Figure 6.30 for rhodamine WT dye, which is commonly used as a luminescence standard. Watson (1981) suggests that luminescence levels can be expressed in terms of the concentration of a standard material required to produce the observed level (e.g., equivalent ppb of rhodamine WT). Luminescence can be studied with an active source such as a laser used to illuminate the target. The induced luminescence can then be observed using any of the imaging sensors we have described. To reduce the laser power levels, scanning systems are often used so that only a small region around the GIFOV of the sensor needs to be illuminated.

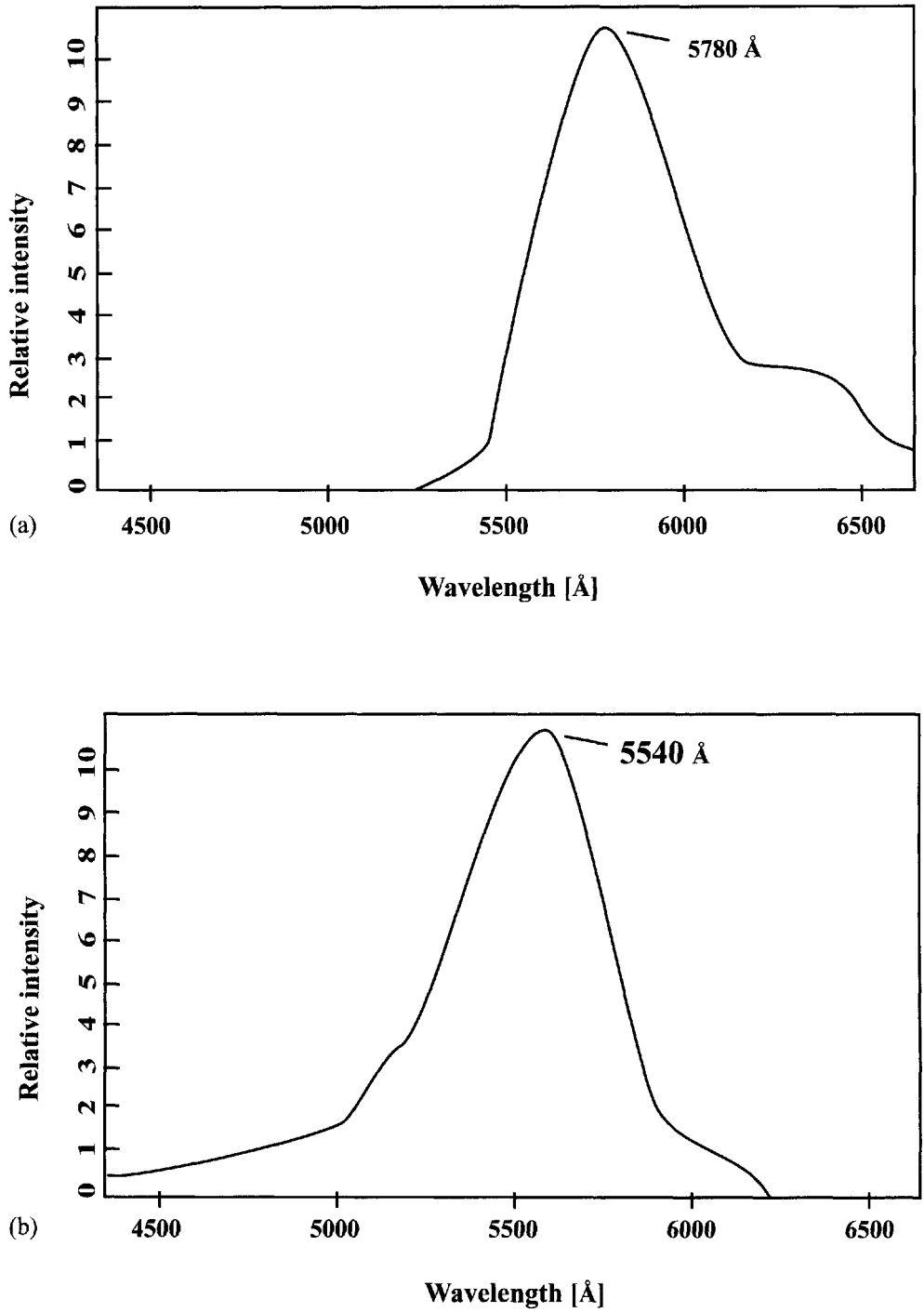


Figure 6.30 Luminescence spectra of rhodaminne WT. [Adopted from Hemphill et al. (1969.)] (a) Emission spectrum for an excitation wavelength of 554 nm. (b) Excitation spectrum for an emission wavelength of 578 nm.

The sensor can collect energy either in a single band at wavelengths longer than the laser or at several wavelengths to take advantage of variations in the emission spectra from different materials. Depending on the emission characteristics of the material being studied, sensing may be more effective under low-ambient illumination conditions.

An alternative approach described by Hemphill et al. (1969) uses passive sensing systems by taking advantage of the Fraunhofer absorption lines in the solar spectrum described in Section 3.5. By sensing in very narrow spectral bands centered on the Fraunhofer lines, the solar-stimulated luminescence can be detected above the reflected background flux. This can be accomplished by measuring the ratio of the irradiance in an absorption line to that in the adjacent continuum and comparing this to the ratio of the exitance in a Fraunhofer line to that from the adjacent continuum, as illustrated in Figure 6.31. We then have the solar irradiance ratio (R_E) given by

$$R_E = \frac{E_b}{E_a} \tag{6.2}$$

and the exitance ratio given by

$$R_M = \frac{E_b r + M_\eta}{E_a r + M_\eta} = \frac{M_b}{M_a} \tag{6.3}$$

where M_η is the exitance due to stimulated luminescence, and the reflectance r is assumed constant over the narrow spectral interval *a-b*. Clearly, in the absence of luminescence, R_E equals R_M, and the difference between R_E and R_M is a measure of luminescence. We can formally define a luminescence coefficient that is a material property as

$$\eta = \frac{M_\eta}{E_a} \tag{6.4}$$

In terms of the measurable values of irradiance and exitance, the luminescence coefficient can be found from

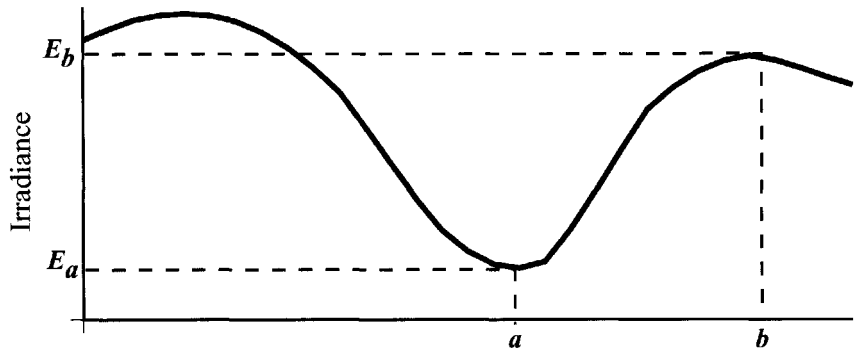
$$E_b r = M_b - \eta E_a \tag{6.5}$$

$$E_a r = M_a - \eta E_a \tag{6.6}$$

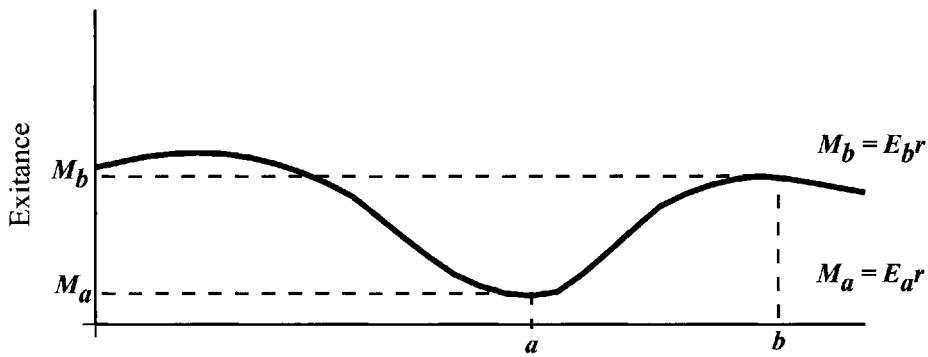
and eliminating r and rearranging yields

$$\eta = \frac{E_b M_a - E_a M_b}{E_a E_b - E_a^2} \tag{6.7}$$

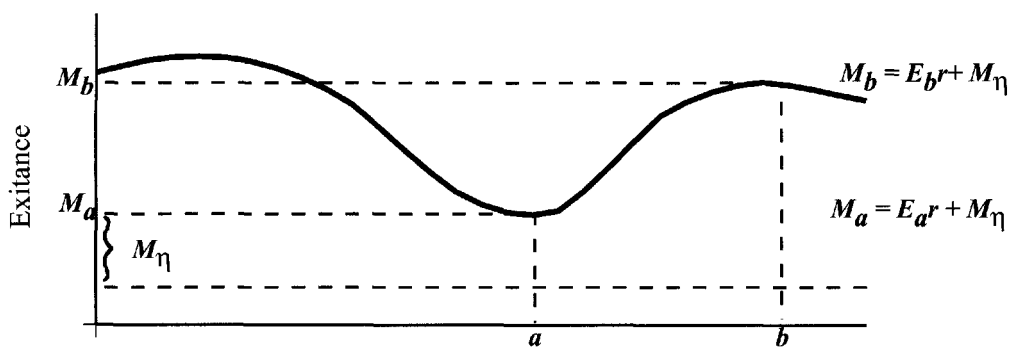
Watson and Theisen (1981) describe an imaging instrument called a *Fraunhofer line discriminator* (FLD) designed to produce images of solar-induced fluorescence levels in the Fraunhofer lines. It consists of a 1° field of view radiometer coupled to a line scanner configuration, as shown in Figure 6.32. The radiometer has a pair of coupled chopper blades that let it alternately look at the solar irradiance and then at the exitance from the ground. The flux is then sent through a beam splitter to a pair of Fabry-Perot etalon/filters and then to a detector to sample either the Fraunhofer line (e.g., the sodium D_2 line at 589 nm) or the adjacent continuum.



(a) Irradiance spectra in a Fraunhofer line and sampling points.



(b) Exitance spectra from the ground in the absence of luminescence.



(c) Exitance spectra from the ground when luminescence occurs.

Figure 6.31 Examples of sampling points at a and b near a Fraunhofer line (a) for irradiance spectra, (b) for exitance spectra in the absence of luminescence, and (c) with luminescence.

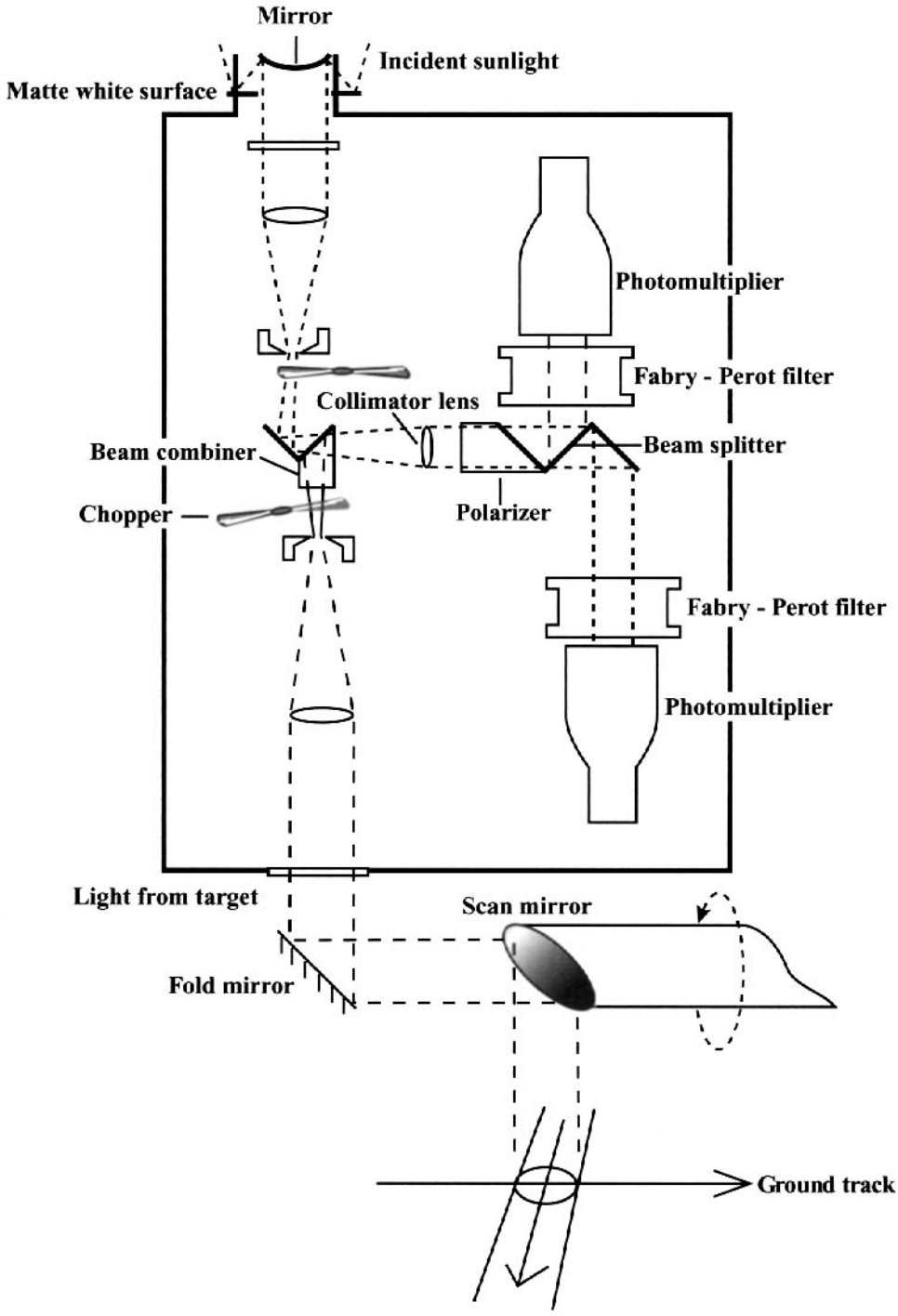


Figure 6.32 Schematic of an FLD instrument. [Adapted from Hemphill et al. (1981).]

Thermal control of the etalon/filter combination allows sampling in a very narrow (≈0.1 nm) band centered on the absorption line. The second etalon/filter combination is adjusted to sample just far enough away from the line center to avoid the absorption wings. Between the two fields of view (up and down) and the two spectral channels, the FLD instrument is able to acquire estimates of E_a, E_b, M_a, and M_b for use in Eqs. (6.2), (6.3), and (6.7). For simplicity, we are assuming the FLD instrument is flown low enough to minimize any atmospheric effects. These instruments have typically been fairly low resolution due to their narrow spectral band width and are flown at low altitude and low speed for improved GIFOV and dwell time, respectively.

A major advantage of luminescence imagery is that the luminescence signatures are uncorrelated with reflectance or thermal signatures. As a result, they represent a new dimension in the material identification space that offers exciting opportunities for improved image analysis when used synergistically with traditional methods. Most early work has focused on single-line instruments, but it is possible to build multispectral luminescence instruments for improved discrimination of luminescence signatures.

6.4 CALIBRATION ISSUES

In this section, we will be primarily concerned with ways to ensure the internal radiometric calibration of EO sensors. However, we will briefly touch on two other calibration issues. The first is geometric calibration, which can be thought of in two reference frames: How well do we know the relative orientation of one pixel to another (internal orientation)? How well can we project the optical axis onto the ground (external orientation)? As mentioned in section 6.1, the inherent internal geometric fidelity is best with a framing system, worse with a push broom, and poorest with a scanner. Recognize that any individual scanner may have very high geometric integrity. The designer just had to work harder to provide it. The overall geometric calibration (external orientation) is the main focus of photogrammetry (cf. Sec. 2.2) and will not be addressed here.

The second calibration topic we will briefly return to is spectral registration (cf. Sec. 6.2). In many multispectral and hyperspectral systems, the integrity of the spectral signature can be compromised due to misregistration of the various spectral bands. When this occurs, the spectral data we associate with one ground spot is actually some mixture of that spot and the surrounding pixels (i.e., we have a blurred spectral signature). This is essentially a geometry and sampling problem very closely related to the internal geometry problem discussed above. This problem is minimized when the spectral data are all sampled simultaneously through a common limiting aperture as illustrated in Figure 6.7. It is worse when pixel shifting occurs such as in whisk-broom and push-broom systems, and worse still when entire images must be resampled for registration (e.g., framing systems). As with geometric integrity, the reader is cautioned that these rules of thumb apply to the generic designs and that specific sensor performance may deviate from them if substantial effort has gone into ensuring spectral integrity. In most cases, spectral

registration is controlled by the internal geometry of the sensor and internal timing sequences. Calibration involves imaging controlled geometric structures (e.g., edges) in the laboratory and defining relative locations of sampling locations relative to sample rates so that appropriate shifts can occur during image reconstruction to geometrically align all spectral samples. In flight (particularly immediately after launch), it may be necessary to verify with controlled targets that the internal geometry and timing circuits are performing as calibrated. Regrettably, if shifts occur, they are likely not in integer pixel values, and some level of spectral mixing may occur even for the best characterized system.

The primary calibration issue of concern to us in this section is how to provide internal calibration to relate the observed signal (typically in digital counts) to the effective radiance reaching the sensor. This is normally done using on-board calibration standards when high radiometric integrity is required. As an alternative, laboratory preflight calibration and periodic postlaunch calibration assessment programs can be used if no calibration drift is expected or if accuracy specifications are reduced. Both methods require filling the collection aperture with known radiance levels in the laboratory and observing the output from the sensor. So we will begin by considering laboratory radiance sources.

The first step in radiometric calibration is often spectral calibration to define the relative spectral response of a spectral band. This can be accomplished using a broad-band source (e.g., tungsten halogen lamp or hot blackbody) whose spectral flux is sequentially passed through a scanning monochromator that in turn irradiates the entrance aperture of the sensing system. Two full-aperture approaches are shown in Figure 6.33. It is also possible to compute the relative spectral response at the component level and cascade the effects together (e.g., spectral transmission of optics, spectral transmission of filters, and spectral response of detectors); however, because the spectral response can be impacted by internal geometry, it is usually best to perform full-aperture tests where possible with the illumination field mimicking actual collection geometry. For the full-aperture approach, the relative spectral response per detector can be computed by running the sensor in its normal mode (or as close as possible given the test conditions) and scanning through the spectrum. The peak normalized spectral response can then be expressed as

$$R'(\lambda) = \frac{\hat{S}(\lambda)}{\hat{S}_{\max}(\lambda)} \tag{6.8}$$

where

$$\hat{S} = S(\lambda)\left[\frac{L(\lambda)}{L_{\max}(\lambda)}\right]^{-1} \tag{6.9}$$

and $S(\lambda)$ is the signal out of the sensor for flux at wavelength λ (e.g., for an end-to-end test, this might be recorded digital counts), $\hat{S}_{\max}(\lambda)$ is the maximum signal observed, and $L(\lambda)/L_{\max}(\lambda)$ is the relative radiance introduced from the calibration system at each wavelength. Note that for many narrow bandpass systems $L(\lambda)/L_{\max}(\lambda)$ is assumed to be unity over the spectral band.

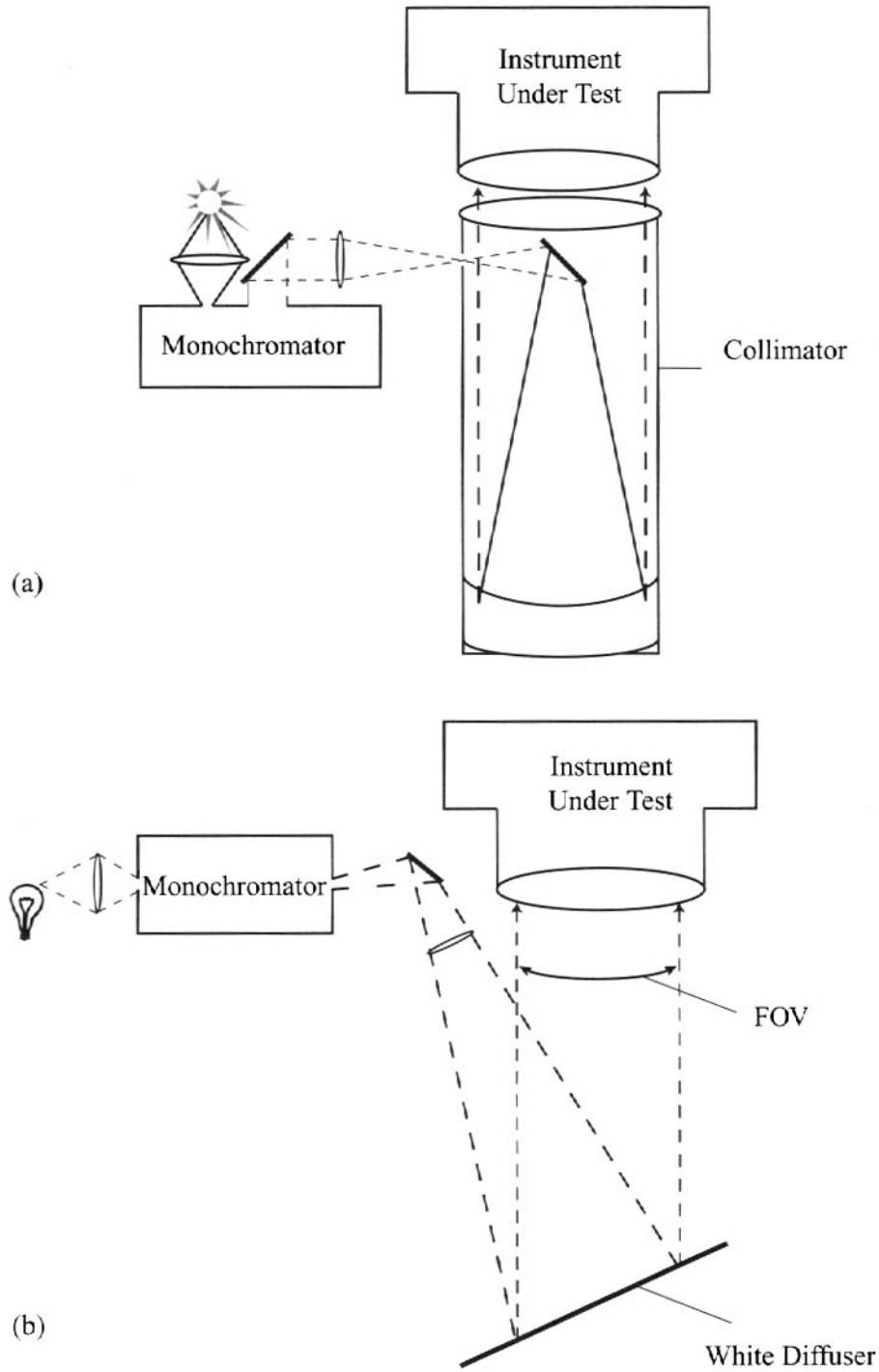


Figure 6.33 Illustration of two methods of spectral calibration: (a) using a monochromator and a collimator and (b) using diffuse reflectance from a white Lambertian reference panel.

In addition to the full spectral shape of the response function ($R'(\lambda)$), the spectral shape is also often characterized by the bandpass expressed as the *full width at half the maximum* (*FWHM*) as illustrated in Figure 6.34. The nominal band edges are also commonly defined using the half maximum of the relative spectral response, although other criteria can be defined. For spectrometer systems, the relative spectral response can often be approximated by fitting a Gaussian curve to the spectral samples from the spectral response calibration process. As shown in Figure 6.34, use of the Gaussian fit can help in defining both the bandwidth and the band center. The band shape in this case may be characterized either by the band center and the standard deviation of the best fit Gaussian or by the band center and the FWHM derived from the Gaussian fit (i.e., bandwidth = $2.35\sigma_\lambda$) as shown in Figure 6.34.

The relative response function can also be used to compute the out-of-band response on a channel-by-channel basis. One method to specify out-of-band response is to compute the fraction of signal from outside the nominal band limits to the total signal according to

$$F_{ob} = \frac{\int_0^\infty L_\lambda R'(\lambda)d\lambda - \int_{\lambda_1}^{\lambda_2} L_\lambda R'(\lambda)d\lambda}{\int_0^\infty L_\lambda R'(\lambda)d\lambda} \tag{6.10}$$

where F_{ob} is the fractional out of band response, L_λ is the spectral radiance of a nominal target (e.g., a Planckian blackbody at 5000K or 300K in the thermal), and λ_1 and λ_2 are the lower and upper band limits. Note λ_1 and λ_2 may be defined as the FWHM values, or some other specification may be used (e.g., when response first falls to 10% of the maximum when moving away from the peak response). The out-of-band response is designed to capture potential problems from spectral leakage such as shown in Figure 6.34(a), which shows a band where the out-of-band response might be unacceptably high for some applications.

After the spectral response of each channel is characterized, the overall channel response to absolute radiance levels can be calibrated. A common procedure for laboratory calibration in the reflective portion of the spectrum is to fill the entire entrance aperture of the sensor with known radiance levels from an *integrating sphere*. Figure 6.35(a) shows the integrating sphere concept, and Figure 6.35(b) shows a large sphere that NASA has used to calibrate sensors. To produce the uniform radiance field, the integrating sphere must be large compared to the aperture. For large spheres to achieve radiance levels similar to what the sensor will see operationally, a number of sources are distributed inside the sphere outside of the sensor's direct line of sight. Multiple bounces from the approximately Lambertian highly reflective walls produce a nearly uniform radiance field over an area larger than the sensor's entrance aperture. To change the radiance level, sources are turned off or on. This preserves the spectral distribution of the flux. The spectral radiance from the sphere is carefully measured and uniformity verified before use [cf. Barker et al. (1985a)].

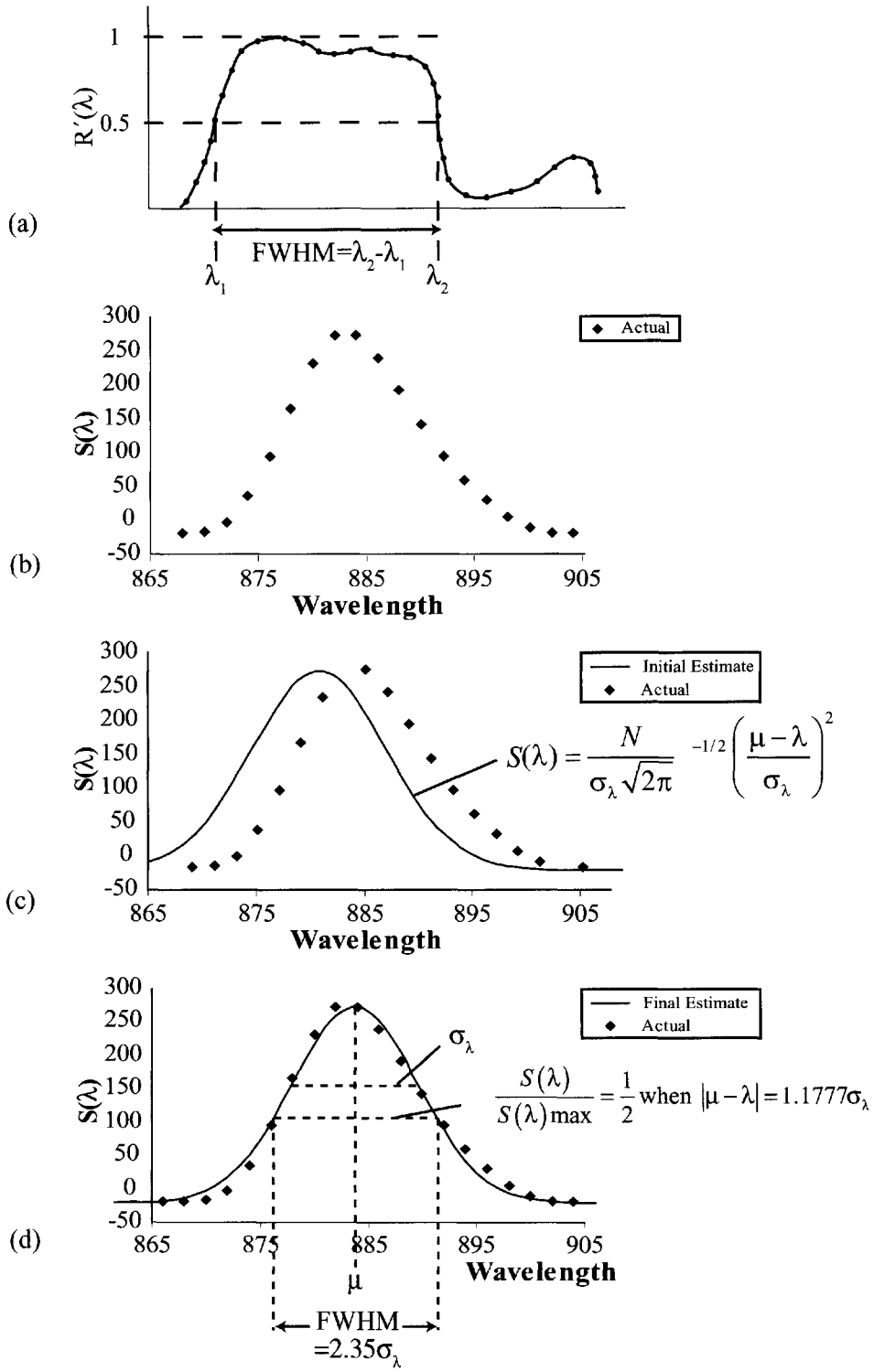
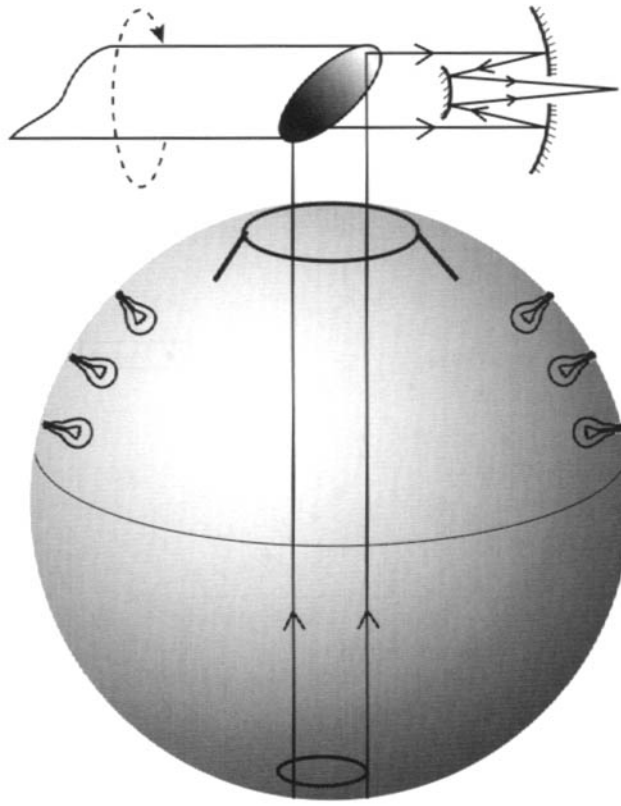


Figure 6.34 Illustration of relative spectral response calibration concepts: (a) the relative response samples from a broadband sensor and the full-width half-maximum (FWHM) calculation of bandwidth; (b) samples from a single band of an imaging spectrometer; (c) and (d) iterative Gaussian fits to the samples in (b) where $S(\lambda)$ is raw signal, N is scaled to fit the magnitude of the raw data, μ is the band center, and σ is the standard deviation of the Gaussian fit.



(a)

NASA JPL

(b)

Figure 6.35 Integrating spheres used for sensor calibration: (a) sphere design; (b) sphere used in calibration of the AVIRIS Sensor.

The sensor output, e.g., digital count (DC), as a function of effective incident radiance (L) or effective spectral radiance can then be evaluated over the expected dynamic range of the sensor. Linearity can be verified and the functional relationship between output signal and radiance defined. For a sensor with linear response, this would take on the form

$$DC = gL + b \tag{6.11}$$

where g is the sensor *gain* [counts $W^{-1}m^2sr$] (or counts $W^{-1}m^2sr\ \mu m$ if effective spectral radiance is used) and b is the sensor *bias* [counts]. The gain (g) and bias (b) are found by linear regression of pairs of DC and L values for various lamp settings in the integrating sphere. This process is repeated for each spectral band and each sensor setting (e.g., each gain setting for a multiple gain sensor). Note that for this approach to be effective, the spectral distribution of the calibration sphere must simulate the spectral shape of the expected radiance from the Earth over the spectral channel of interest (cf. Fig. 3.8).

The discussion above uses a source-based standard where we assume that the absolute spectral radiance of the integrating sphere can be treated as known and stable based on the stability of the tungsten halogen lamps used as sources. An alternative approach that is becoming popular as more and more stable detectors are being developed over a broader spectral range is the use of detector standards. In this case, we would use a detector whose signal out as a function of absolute radiance is known very accurately. If we then match the standard detector's spectral response to the sensor bandpass under test (or used a spectrally sampled standard detector), we can directly measure the radiance from any source. Thus, in our discussion above, we would still use the integrating sphere to provide a source of uniform radiance with a known spectral shape; however, rather than use the sphere calibration to define the absolute radiance, we would measure it with the detector-based standard. This is done each time the instrument under test looks at the sphere (i.e., at each radiance level). As we drive to higher and higher radiometric calibration standards, we can expect to see increasing use of detector standards, including development of instrument-specific transfer radiometers such as the Landsat transfer radiometer (LXR) used in detector based calibration of the Landsat ETM+ [cf. Markham et al. (1998)].

The calibration procedures in the self-emissive region of the EM spectrum is essentially the same as for the reflective region, except that the source is changed. The most straightforward source to use is a blackbody with a uniformly radiating surface larger than the entrance aperture of the sensor. The radiance from the blackbody is changed by changing its temperature. With the use of thermistors embedded in the blackbody for monitoring the temperature and as feedback controls in the temperature control circuits, very precise temperatures (and therefore radiance values) can be produced. If the entrance aperture is so large that maintaining thermal uniformity becomes difficult, the radiance field can be produced using a collimator with an exit aperture slightly larger than the entrance aperture of the sensor (cf. Fig. 6.36). The radiance from the collimator can be measured using laboratory instruments or computed if the throughput (G#) of the collimator is well known. Collimators are also used in this fashion in the reflective region of the

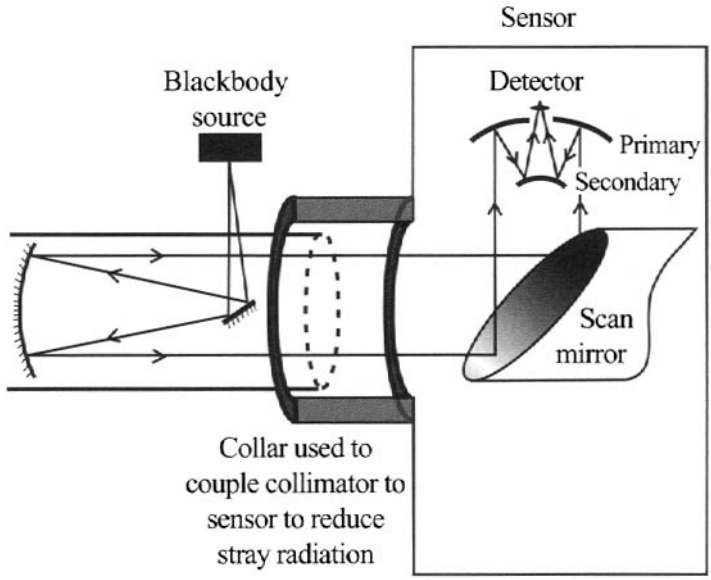


Figure 6.36 Radiometric calibration of a sensor using a calibration source and a collimator.

spectrum. In general collimators are avoided, where possible, in absolute calibration, since the calibration of the collimator can introduce an additional element of uncertainty (error) in the sensor calibration. The calibration of the thermal channels proceeds in the same fashion as the reflective channels using effective radiance (L) and digital count (DC) values associated with different blackbody temperatures to define the response function, i.e.,

$$\mathrm{DC}=f(L) \tag{6.12}$$

where f is the function relating DC to L, which can often be approximated as a linear function of the form shown in Eq. (6.11). The stability of this type of calibration is very much a function of the temperature of the sensor elements (particularly the optics) and is subject to considerable drift as the sensor temperature varies.

For systems whose response is assumed stable over time, the laboratory calibration may complete the calibration process, and the resultant calibration values (g and b) are used for analysis of image data. Generally, some form of validation test is performed to ensure the stability of the calibration over time. For aircraft sensors, the system is often recalibrated after each flight to ensure that pre- and postflight calibrations agree. For satellite systems, the test process can be quite a bit more difficult. Rao and Chen (1993) describe how the stability or instability of the NOAA AVHRR sensors have been monitored using uniform desert regions as stable targets after correction for viewing geometry effects. They point out a major limitation of using only preflight calibration values. Any change in sensor performance will result in a change in the calibration values, of which the user may be unaware. For example, the visible and NIR AVHRR sensors were shown to have a steady decay in gain of several percent per year. As a result, the user of sensors

calibrated in this fashion must be very careful of trusting cited calibration values, unless the current validity of those values has been carefully evaluated.

To reduce the uncertainties in radiometric calibration values, many sensors employ some type of on-board calibrator. The most straightforward of these are full-aperture sources located ahead of all of the optical elements in the sensor. This approach is very commonly used in line scanners, as shown in Figure 6.1. By filling the entire aperture and being ahead of all of the sensor's optical elements, this provides essentially a full calibration of the entire sensor that reproduces in flight the laboratory calibration. Most thermal infrared systems, because of their inherent temporal instability, use two on-board blackbodies set to different temperatures (radiance levels). Each time the calibration sources are observed (e.g., every line in a line scanner), a two-point linear calibration can be performed and the sensor gain (g) and bias (b) terms updated. Spectral channels with greater inherent stability may be calibrated less frequently. For example, the radiance from a calibration source in a visible channel may be changed every 100 lines. If five radiance levels are used in the calibrator, it will take 500 lines of data to develop a calibration update. In the case of satellite systems, full-aperture calibration may only be performed occasionally (e.g., monthly or quarterly) with the assumption that the instrument is stable over shorter times or that short-term variability is monitored and corrected by other calibration methods (see the internal calibration discussions below). Several satellite instruments have deployed full aperture calibrators in the reflective regime [cf. Xiong et al. (2003a) and Clodius et al. (1999)] employing diffusing panels in front of the telescope as shown in Figure 6.37. To first order, the spectral radiance from such a diffuser can be expressed as

$$L_\lambda = E_s' \cos\sigma_d \frac{r_p(\lambda)}{\pi} \tag{6.13}$$

where σ_d is the angle from the normal to the diffuser to the sun and $r_p(\lambda)$ is the bi-directional reflectance factor for the diffuser for the relevant sun-panel-telescope orientation. In principle, this approach only requires a well-calibrated reflectance panel. In practice, small corrections are required for the fact that some of the light

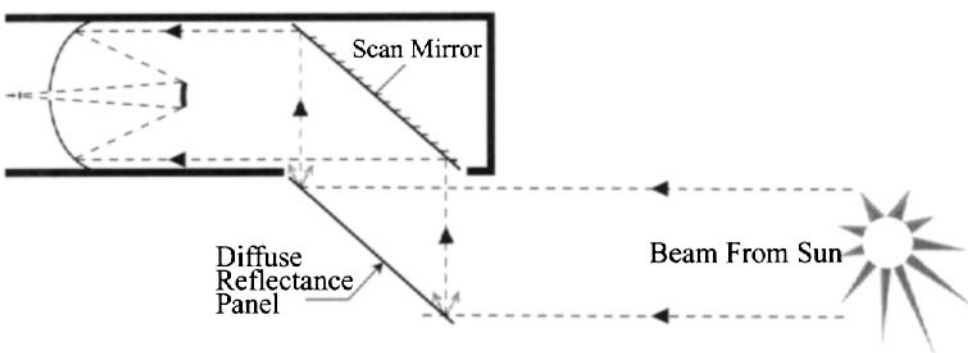


Figure 6.37 Illustration of the use of a full-aperture calibration panel. The panel is deployed in front of the telescope such that diffused, reflected radiance fills the entrance aperture.

flooding into the telescope reilluminates the panel and is therefore diffused back into the telescope. In addition, the panel's reflectance typically changes slowly over time, and some method to monitor and correct for these long-term changes is required. In the thermal infrared, a blackbody panel (door) can be placed over the telescope. If the temperature and emissivity are well known, the panel/door becomes a source of known radiance. This approach was used for thermal calibration of MODIS using a panel and the multispectral thermal imager (MTI) using a door [cf. Xiong et al. (2003b) and Clodius et al. (2000)].

Having full-aperture calibrators can be very costly from a weight and volume standpoint as the sensor aperture increases. As a result, many alternative methods are used. For example, in the TM, a calibration "wand" is "waved" in the optical path while the scan mirror is reversing direction [cf. Barker et al. (1985b)]. The calibrator is introduced after the telescope when the beam diameter has been considerably reduced. Using fiber optics from three lamps, it irradiates the detectors with controllable irradiance levels as it is swept across the beam. The temperature of a surface on the calibrator provides one thermal IR calibration point, and a mirror on the calibrator reflects the energy from a small blackbody onto the IR detectors to provide a second calibration level. Thus, the thermal band is fully updated each line.

In all cases where on-board calibrators are used, the calibration sources must themselves be calibrated. This is particularly true when something other than a full-aperture source of known radiance is used (e.g., the TM calibrator). The on-board calibrators are usually calibrated by first carefully calibrating the sensor using laboratory sources as described above. Then, for those sensor conditions, the digital-count-to-radiance relationship is well known, and any digital count can be converted to a known radiance level. If no changes are made to the sensor configuration except to let it "see" the on-board calibration source(s), then the digital count observed can be converted to the calibration radiance level that will be associated with that source condition. This process is repeated until all source settings (e.g., all calibration lamp levels) are calibrated to known radiance levels. This transfers the calibration of the laboratory radiance standards to the on-board calibration sources. It is important to recognize that the radiance from the standards is not necessarily derived by this process. Instead, we derive the value for what the radiance reaching the front of the optical system would have to be to produce the same flux on the detector as is being produced by the calibration source.

On-board calibration sources would appear to be the solution to sensor drift, instability, and sensor and electronic-aging problems. Because the sensor's calibration is continually updated, we should always be able to convert the observed signal to the effective radiance reaching the sensor. Regrettably, there are several obstacles that may continue to limit this process for certain sensors. The first limitation was introduced in describing the TM's calibration approach. If the calibrator is not ahead of all of the sensor optical elements (including windows), then it cannot account for changes that may occur to the temperature or surface characteristics of the forward surfaces. For space-based systems like the TM, the fouling of the mirror surfaces is expected to be a very slow process having only very minor long-term impact on the sensor. Furthermore, the temperature of the fore optics can be

monitored or controlled and slight corrections introduced into the calibration equation using a fore optics model to attempt to adjust for changes. The net result is that designs that do not have the sources ahead of all optical elements can be calibrated. However, the calibration can be in error since changes in the system can occur that are not corrected for in the calibration process. As a result, it is important to perform periodic calibration checks on this type of system. For systems that do not use scan mirrors (e.g., push-broom and framing systems), it can be very difficult to view calibration sources on a regular basis. The calibration of these sensors is further complicated by the increased physical size of the focal plane and the increased angular field that must be exposed to a uniform radiance field. If the calibrator flux is not uniform, detector-to-detector calibration error will be introduced, which affects not only the radiometric calibration but also the appearance of the images by introducing striping and shading artifacts. One way to calibrate array systems is to periodically turn the pointing mirror, if there is one (e.g., SPOT), so that the sensor looks internally at calibration sources. Alternately, the entire sensor can be turned to look at calibration sources (i.e., deep space or the moon). Both these maneuvers are risky with satellite systems since they place the sensor in a nonimage collection mode where it could become "stuck." Consequently, this type of calibration is usually performed much less frequently than is commonly the case with scan mirror systems. In summary, as Norwood and Lansing (1983) point out, radiometric calibration difficulties increase as we go from scanning, to push- broom, to framing sensors.

An additional difficulty plagues all satellite calibration systems (and even some aircraft sensors if flight schedules preclude periodic calibration maintenance). This difficulty arises from decay in the calibrator itself. Calibration relies on the sources producing the same flux levels over time. On satellites, this may be over periods of many years. Over time, the sources, the calibration control electronics, or the calibration monitoring sensors can change. Thus, it is necessary to periodically calibrate the calibrator. One way to do this is with a redundant calibrator that is not used operationally. If it is activated on a periodic basis, its prelaunch radiance can be compared to the radiance predicted by the operational calibrators. Any discrepancies can then be used to update the calibration of the calibrator. This redundant approach can be very expensive, and with full-aperture calibration it may be nearly impossible to accomplish. In addition, in some cases it just raises the question of whether the backup calibrator or the operational units are in error.

Teillet et al. (1990) and (2001) suggest that one sensor system can be used to check the calibration on a second. For sensors that have nearly coincident acquisitions, the spectral radiance at the top of the atmospheres along a common line of sight should be the same. By correcting for differences in spectral response functions, some level of cross-calibration can be achieved.

Nearly all the problems described above can be overcome by periodic laboratory calibration updates. For satellite systems, this can be a bit difficult unless we use the world as our laboratory. Slater et al. (1987) describe how this has been done in the reflective region using large, nearly uniform reflectance surfaces (White Sands, New Mexico) observed through a very clear, dry, well-characterized atmosphere. Under these conditions, Thome (2001) estimate that by measuring the

reflectance and modeling the atmosphere, the radiance at the top of the atmosphere can be predicted to 3–5%. When these measurements are made coincident with a satellite overpass, they provide a single well-known calibration point. The white sands represent a high radiance level where DC and radiance are known. If the sensor sees essentially zero flux (e.g., in the back scan) and is biased such that the digital count associated with this flux is observed, then a second point is available to provide a linear calibration update for the system.

Schott et al. (2001) describe how a similar approach can be used to calibrate the thermal channels on satellite systems. A calibrated thermal sensor on an aircraft is used to underfly a satellite sensor with a similar spectral response. The airborne sensor is used to sense the radiance near the top of the atmosphere (e.g., 7 km), and then atmospheric propagation models are used to predict the slight changes in the radiance that the satellite would see at the very top of the atmosphere. By observing several targets over a range of radiance values, a digital-count-to-radiance calibration update for the satellite can be developed. Schott et al. (2004) indicate that, when care is taken in target selection and the use of radiometric correction techniques, the error in top-of-the-atmosphere radiance values using this technique can be reduced to better than 0.3 K. This result is for the LWIR spectral region, with the error expressed in apparent temperature. In practice, this vicarious calibration approach has been used to update the thermal calibration of the Landsat ETM+, which was observed to be in error by approximately 2 K immediately after launch [cf. Schott et al. (2001)].

Our discussion thus far has addressed ways to characterize the mean-level radiometric performance of sensor systems, but it has not addressed how to characterize the system noise. In Chapter 5, we discussed the concept of noise largely at the component (i.e., detector) level. From the user's perspective, it is the system-level noise or often signal to noise relative to a reference level about which we will be most concerned. For systems dominated by additive (bias) noise, we can easily characterize the *noise* if a dark-field image or sample is acquired. This is done by closing a shutter somewhere in the optical train of a reflective system and sampling the detector as though it were in imaging mode. The resulting image is referred to as a dark-field image (or image segment). The standard deviation of the dark-field image for each detector represents a dark noise estimate in digital counts. The *noise-equivalent radiance* (dark) can then be calculated using the gain terms from Eq. (6.11) to yield

$$NER = N_d / g \tag{6.14}$$

where NER is the noise equivalent effective radiance or noise equivalent effective spectral radiance, depending on how g is calculated; N_d is the dark noise; and g is the gain in counts per unit radiance or counts per unit spectral radiance. For instruments that look at a bright target (e.g., with a solar diffuser, an internal lamp, or an integrating sphere in the lab), the bright noise level can also be computed (i.e., standard deviation about the mean signal level). If the dark noise and the bright noise are comparable during initial systems test, then the system is bias noise limited and only dark noise need be considered in the future. For systems where the noise is

signal dependent, we may choose to characterize the noise at several signal levels and develop a signal-dependent model of the noise.

The process of noise characterization in the thermal region is quite similar except that a dark noise level is of little interest since we seldom observe targets near 0 K. Instead, the noise is typically characterized by signals (images or image segments) obtained when the sensor "observes" blackbody calibration standards either during internal calibration or during laboratory calibrations. Once again, the standard deviation about the mean for a fixed target is computed to yield the noise in digital counts. The noise equivalent radiance is then computed as shown in Eq. (6.14) using gain values derived using the results of the analysis associated with Eq. (6.12).

As discussed in Chapter 5, noise without reference to a signal level is not particularly informative. Thus, we often characterize the performance of a system in terms of the signal to noise expected for a certain signal level. The signal level used may depend on the instrument or the application. For example, the signal to noise expected for a top-of-the-atmosphere reflectance of 0.18 with a solar zenith of 30° might be specified. For any given channel where we have computed the relative spectral response, we can compute the top of the atmosphere spectral radiance to be

$$L_\lambda(18) = \frac{\int E'_{s\lambda} \cos\sigma \frac{r}{\pi} R'(\lambda) d\lambda}{\int R'(\lambda) d\lambda} = \frac{\int E'_{s\lambda} \cos 30 \left(\frac{0.18}{\pi}\right) R'(\lambda) d\lambda}{\int R'(\lambda) d\lambda} \tag{6.15}$$

This leads to an expected signal to noise for an object with an apparent reflectance of 18% of

$$\frac{S(18)}{N} = \frac{L_\lambda(18)}{NER_\lambda} = \frac{L_\lambda(18)}{N / g_\lambda} \tag{6.16}$$

where NER_λ is the spectral noise equivalent radiance for an 18% reflector, N is the noise expected for an 18% signal (note that, for a system with bias limited noise, this would be the dark noise), and g_λ is system gain [counts $W^{-1}cm^2sr\ \mu m$]. A similar expression can be derived for the expected signal to noise for a thermal band, where the signal to noise is typically characterized for a target with some apparent temperature (e.g., 273 K). This would yield an expected signal of

$$L_\lambda(T) = \frac{\int L_{bb\lambda}(T) R'(\lambda) d\lambda}{\int R'(\lambda) d\lambda} = \frac{\int L_{bb\lambda}(273) R'(\lambda) d\lambda}{\int R'(\lambda) d\lambda} \tag{6.17}$$

and an expected signal to noise of

$$\frac{S(273)}{N} = \frac{L_\lambda(273)}{NER_\lambda} = \frac{L_\lambda(273)}{N / g_\lambda} \tag{6.18}$$

where NER, in this case, is the noise equivalent radiance associated with a target at 273 K or for a system dominated by additive noise any convenient noise estimate can be used.

The reader is cautioned that many instruments report signal to noise against different signal levels. For example, the AVIRIS instrument uses a 50% apparent reflectance and HYDICE a 5% apparent reflectance. Thus, assuming similar solar elevation specifications and additive noise, we need 10 times better signal to noise from AVIRIS to have comparable performance. It is also important to recognize that if we actually image a 50% reflector with AVIRIS, the observed signal to noise will not be the specified value. This is because the actual signal will be affected by atmospheric effects, solar elevation, and target orientation at the time of imaging, which are not included in the specified value (i.e., 50% apparent reflectance) as described above.

Before leaving this subject, we should note an overarching problem associated with sensor calibration. This is the fundamental difficulty associated with making accurate measurements of radiometric quantities. Grum and Becherer (1979) point out that, unlike other physical quantities where measurements to small fractions of a percent are common, measurements of radiant energy to the 1% level are extremely difficult, and in many cases errors of up to 10% are considered adequate. A great deal of care and experience are required to obtain measurements approaching the 1% level, even in the laboratory, and approaching similar levels on airborne or space-based systems is extremely difficult. We should note that in the thermal region where temperature measurements can be used to assist the process, absolute levels of better than 1% are readily achieved. However, in this region, application requirements push accuracy needs below the 1% level (cf. Chap. 7). In summary, sensor calibration is still difficult to achieve with high accuracy. As the remote sensing community increasingly tries to make its results repeatable and quantitative, absolute calibration becomes more and more important. Particularly as we move to an era of trying to make long-term environmental studies using remote sensing, it will be critical that we have highly calibrated sensor data. Chapter 7 addresses in depth the overall issues associated with the use of calibrated sensor data.

6.5 SENSOR CASE STUDY

In this section, we will perform some simple calculations of sensor performance to attempt to tie together the concepts introduced in this chapter. To make a tractable example, we will not consider a full sensor design but a sensor upgrade such that the degrees of freedom are reduced. Furthermore, to make the example more interesting, we will take an actual problem for our case study. The sensor in question is a thermal infrared line scanner using a design similar to that shown in Figure 6.1. The original system performance specs are listed in Table 6.4. Ground spot size (scene element) will be a critical issue in any upgrade considerations, so we first need to convert system specifications to a meaningful ground sample size. To compute the GIFOV of the system, we need to find the IFOV and know the flying height. The IFOV can be expressed as

Table 6.4 Specification of the Original IR Line Scanner Used in the Case Study	
Optics	Dall-Kirkham-Cassegrain
Aperture	$d = 3$ inches
F number	$F = 4$
Mirror scan rate	$S_o = 110$ Hz
8 - 14 μm filter	Filter factor = 0.7
IR tran window	Filter factor = 0.9
Detector specifications, - HgCdTe	
•FOV as defined by cold shield	20°
•Length of side	$\ell = 0.75$mm
•Specific detectivity	$D_\lambda = 3.31 \cdot 10^{10}$ cm $Hz^{1/2}W^{-1}$
•Responsivity	$R = 1306$ V/W
System noise expressed as apparent temperature	$N_o = 0.25$ -0.35 K (gain dependent)
Total FOV	120°
Aircraft speed range	$v = 100$–180 mph
Derived Values	
Focal length	$f = F\# \cdot d = 4 \cdot 76.2$ mm = 304.8 mm
Instantaneous FOV	IFOV = $\ell/f = 0.75$ mm/305.5 mm $= 2.5 \cdot 10^{-3} = 2.5$ milliradians
Bandwidth	$\Delta f = 0.14 \cdot 10^6$ Hz
Throughput	$G\# = 40.4$ sr^{-1}
NEΔT (optics and detector)	NE$\Delta T_0 = 0.09$ K
Noise equivalent power	$NEP_0 = 8.5 \cdot 10^{-10}$W

$$\text{IFOV} = \frac{l_o}{f} = \frac{l_o}{F\#d} = \frac{0.75[mm]}{4 \cdot 76.2[mm]} = 2.5 \cdot 10^{-3}[\text{radians}] \tag{6.19}$$

where l_o is the side dimension of the square detector, f is the focal length, d is the diameter of the entrance aperture (3 inches), and $F\#$ is defined to be f/d. This 2.5-milliradian system was usually flown at 1000 feet above ground level (AGL) (H) to yield a GIFOV of

$$\text{GIFOV} = H \cdot \text{IFOV} = 1000[\text{ft}] \cdot 2.5 \cdot 10^{-3} = 2.5\ [\text{ft}] \tag{6.20}$$

At the time in question, a primary use of the line scanner was to study heat loss from residential structures using data from nighttime flyovers during the winter. A combination of engineering studies and pressure from competitors had defined a requirement for a scanner with a GIFOV at nadir of 1 ft, a system noise

equivalent temperature difference of less than 0.3 K, and a ground swath of at least 2000 ft. From an engineering standpoint, the easiest solution was to buy a commercially available system that met the specifications. The management response was *that the solution exceeded the value the company placed on the upgraded system by more than an order of magnitude,* and a more affordable solution was required. So with very limited resources, our task was to identify and implement an upgrade to the system to meet the improved specifications. We already met the noise and FOV specifications, but we needed to improve on GIFOV. The easiest solution was to fly lower. However, to meet the GIFOV specification, we would have had to recommend flying the system at 400 ft. This is below safety limits for operational collections and would not provide an adequate ground swath. As part of this consideration, it was determined that the current operational flying height of 1000 ft was also a practical minimum. This led to a computation that the system IFOV must be

$$\text{IFOV}=\frac{\text{GIFOV}}{H}=\frac{1\ [\text{ft}]}{1000\ [\text{ft}]}=1\cdot 10^{-3}\ [\text{radians}] \tag{6.21}$$

or less. A 1 milliradian system would have a detector size of

$$l_n=\text{IFOV}\cdot f=1\cdot 10^{-3}\cdot 304.8\ [\text{mm}]=0.3\ [\text{mm}] \tag{6.22}$$

where l_n is the side dimension of the detector; the subscript n will be used as necessary to designate the new system and o the original system. Detectors of this size were readily available in HgCdTe, and this still appeared to be the most appropriate detector material. Improvements in fabrication technology allowed the vendors to specify an effective D^* value for a new detector of $3.3\cdot 10^{10}$ [$W^{-1}cm\ Hz^{1/2}$] and an effective responsivity of 6300 [VW^{-1}]. They also agreed that (for a small fee) they would hand select to try to exceed these values for a single detector. After considering a variety of alternatives, we decided to evaluate whether the system could be improved sufficiently by using a smaller detector. As we will see, this is not as simple a solution as it may at first sound.

Knowing that the flux on the detector would decrease in proportion to the area (6.25 times) and that we could only count on a four-fold increase in responsivity and possibly no increase in specific detectivity, a more detailed assessment to determine if the current system was detector noise limited seemed in order. This would help us determine if we needed to consider only detector specifications or if improvement in the preamplifier (preamp) noise specifications should also be considered. We also need to determine if the scan rate is sufficient to allow full coverage with the smaller GIFOV. The minimum aircraft speed (v) is 100 [mph]. The ground advance between lines (X_o) is:

$$X_o=v\,t_o=v\ /s_o=\frac{100\ [\text{mph}]}{110\ [\text{Hz}]}=1.33[\text{ft}] \tag{6.23}$$

where t_o is the time per revolution of the scan mirror and s_o is the scan rate expressed in revolutions per second. At the original scan rate and minimum air speed, the improved scanner will have a gap between scan lines of 0.33 ft when flown at 1000 ft. To have adjacent scan lines, the scan mirror would have to spin at

$$S_n = \frac{v}{X_n} = \frac{100\ [\text{mph}]}{1\ [\text{ft}]} = 147\ [\text{revolutions sec}^{-1}] \tag{6.24}$$

This increase in scan mirror rate would require a change in the drive motor and would reduce the dwell time, further increasing our already substantial concerns about meeting the noise specifications. For the present, we will plan on making this modification, but consider it a tradeoff option.

While we are considering scan speeds, we should also analyze the frequency response required by the original and the upgraded system since we will need this for the noise calculations. First, we compute the time per sample t_{sp}, or the dwell time, as the time it takes the mirror to sweep out one IFOV, i.e.,

$$t_{so} = t_o\left(\frac{\#\text{IFOV}}{\text{rev}}\right)^{-1} = \frac{1}{s_o}\left(\frac{\#\text{IFOV}}{\text{rev}}\right)^{-1}$$

$$= \frac{1}{110}[\text{sec}]\left(\frac{2\pi\ [\text{radians}]}{2.5\cdot 10^{-3}\ [\text{radians}]}\right)^{-1} = 3.6\cdot 10^{-6}[\text{sec}] \tag{6.25}$$

The frequency bandwidth (f_o) is computed as

$$f_o = \frac{1}{2t_{so}} = \frac{1}{2\cdot 3.6\cdot 10^{-6}[\text{sec}]} = 0.14\cdot 10^{6}[\text{Hz}] \tag{6.26}$$

For the proposed 1 milliradian system with the increased scan rate, the dwell time (t_{sp}) would be

$$t_{sp} = t_p\left(\frac{\#\text{IFOV}}{\text{rev}}\right)^{-1} = \frac{1}{s_n}\left(\frac{\#\text{IFOV}}{\text{rev}}\right)^{-1}$$

$$= \left(\frac{1[\text{sec}]}{147}\right)\left(\frac{2\pi[\text{radians}]}{1\cdot 10^{-3}[\text{radians}]}\right)^{-1} = 1.1\cdot 10^{-6}[\text{sec}] \tag{6.27}$$

where t_p is the time per revolution at the increased speed, and the subscript p is used to indicate proposed values.

The resulting frequency bandwidth would be

$$f_p = \frac{1}{2t_{sp}} = \frac{1}{2\cdot 1.1\cdot 10^{-6}[\text{sec}]} = 0.45\cdot 10^{6}[\text{Hz}] \tag{6.28}$$

If the scan rate is not increased, we would have a dwell time t_{sn} of

$$t_{sn} = t_o\left(\frac{\#\text{IFOV}}{\text{rev}}\right)^{-1} = \frac{1}{s_o}\left(\frac{\#\text{IFOV}}{\text{rev}}\right)^{-1}$$

$$= \frac{1}{110\ [\text{Hz}]}\left(\frac{2\pi\ [\text{radians}]}{1\cdot 10^{-3}\ [\text{radians}]}\right)^{-1} = 1.45\cdot 10^{-6}[\text{sec}]\infty \tag{6.29}$$

with a corresponding frequency bandwidth f_n of

$$f_n = \frac{1}{2t_{sn}} = 0.35 \cdot 10^6 [\text{Hz}] \tag{6.30}$$

As shown in Figure 5.15, the HgCdTe detectors can pass these signals without a problem. By trying to pass a 450-kHz signal [cf. Eq. (6.28)] through the scanner recorder system, it was found that the preamplifier was the only element that would not pass this signal. A new preamp was planned that would roll off at the required detector bandwidth, so this was not a problem.

We needed to know the NEΔT of the detector relative to the noise performance of the sensor. To begin, we solve for the NEP of the original system as

$$\text{NEP}_0 = \frac{1}{D} = \frac{1}{D*}\sqrt{A_o}\sqrt{f_o} = \frac{0.075[cm]\sqrt{0.14 \cdot 10^6 [Hz]}}{3.3 \cdot 10^{10} [cmHz^{\frac{1}{2}} \text{W}^{-1}]}$$

$$= 8.5 \cdot 10^{-10} [\text{W}] \tag{6.31}$$

where A_o is the area of the original detector. The NEΔT of the original system can then be expressed as

$$\text{NE}\Delta\text{T}_0 = \frac{\text{NEP}}{\Delta\Phi / \Delta T} \tag{6.32}$$

where $\Delta\Phi/\Delta T$ is the change in flux on the detector associated with a unit change in temperature. To solve for this value, we first define the change in radiance on the sensor per unit change in temperature to be

$$\frac{\Delta L}{\Delta T} = \int_8^{14} L_{\lambda 301} d\lambda - \int_8^{14} L_{\lambda 300} d\lambda = 8.41 \cdot 10^{-5} [\text{Wcm}^{-2}\text{sr}^{-1}\text{K}^{-1}] \tag{6.33}$$

where we are implicitly assuming all the specifications are for a 300K target.

Using the $G\#$ concept from Eq. (5.7), we can describe the change in irradiance on the detector per unit change in temperature as

$$\frac{\Delta E}{\Delta T} = \frac{\Delta L}{\Delta T}\frac{1}{G\#} \tag{6.34}$$

The change in flux per unit change in temperature for the original sensor can then be expressed as

$$\frac{\Delta\Phi}{\Delta T} = \frac{\Delta E}{\Delta T} A_o = \frac{\Delta L}{\Delta T}\frac{1}{G\#} A_o \tag{6.35}$$

where

$$G\#_o \cong \frac{4F\#^2}{\tau_f \tau_o \pi} = \frac{4 \cdot 4^2}{0.8(0.9 \cdot 0.7)\pi} = 40.4[\text{sr}^{-1}] \tag{6.36}$$

τ_o is the transmission loss due to obscuration by the secondary optic in the Cassegrain telescope (0.8 for this system), and τ_f is transmission loss due to the bandpass filter (0.7) and the window on the dewar (0.9) in the original system.

Inserting this result into Eq. (6.35) yields

$$\frac{\Delta\Phi_o}{\Delta T} = 8.41\cdot10^{-5}[\mathrm{Wcm^{-2}sr^{-1}K^{-1}}]\cdot\frac{1}{40.4[\mathrm{sr^{-1}}]}\cdot(0.075\ [\mathrm{cm}])^2$$
$$= 1.17\cdot10^{-8}[\mathrm{WK^{-1}}] \tag{6.37}$$

which combined with Eq. (6.32) yields

$$\mathrm{NE}\Delta\mathrm{T}_0 = \frac{\mathrm{NEP}_0}{\frac{\Delta\Phi_0}{\Delta T}} = \frac{8.5\cdot10^{-10}[\mathrm{W}]}{1.17\cdot10^{-8}[\mathrm{WK^{-1}}]} \cong 0.1[K] \tag{6.38}$$

Since the observed system noise N_o is 0.25 to 0.35 [K] and depends on the system gain, we conclude that the system is not quite detector noise limited. If we assume that the bulk of the remaining system noise is from the preamplifier, we can solve for the preamp noise as:

$$N_{ao} = (N_o^2 - \mathrm{NE}\Delta\mathrm{T}_o^2)^{\frac{1}{2}} = \left[(0.25\text{ to }0.35)^2[\mathrm{K}] - (0.1\mathrm{K})^2\right]^{\frac{1}{2}}$$
$$= 0.23\text{ to }0.33\ [\mathrm{K}] \tag{6.39}$$

In any new system, we must recognize that in using a similar preamp design with components with comparable specifications, we must anticipate similar noise levels for similar signal levels. Because the preamp noise increases with gain, we must expect that if we have lower signal levels, we will have even greater noise associated with the preamp.

We can perform a quick assessment of the relative signal levels expected from the proposed detector (S_p) compared to the original (S_o) as follows:

$$\frac{S_p}{S_o} = \frac{A_n}{A_o}\frac{R_p}{R_o}\frac{t_{sp}}{t_{so}}\frac{\tau_n}{\tau_o} \tag{6.40}$$

since we know that the signal out of a sensor will be proportional to area, responsivity, dwell time, and the transmission of the optics. In our case, knowing that we need more signal, we plan to improve the filter factor by having the bandpass filter incorporated into the dewar window for an overall transmission of 0.8. The expected signal ratio using the vendor's estimated responsivity would be

$$\frac{S_p}{S_o} = \frac{(0.03\ [\mathrm{cm}])^2}{(0.075\ [\mathrm{cm}])^2}\frac{6300\ [\mathrm{VW^{-1}}]}{1306\ [\mathrm{VW^{-1}}]}\frac{1.1\cdot10^{-6}\ [\mathrm{sec}]}{3.6\cdot10^{-6}\ [\mathrm{sec}]}\frac{(0.8)}{(0.9\cdot0.7)} = 0.3 \tag{6.41}$$

This means that we would require nearly three times as much gain from the preamp to achieve the same signal levels. This places us at the high end of the preamp noise range.

At this point, we determined to try to maintain as much signal level as possible so that the scan mirror motor speed would not be changed. This decision was augmented by concerns that scan mirror deformation at higher speed might degrade image quality [cf. Feng et al. (1994)]. Based on this decision, we can compute the expected NEΔT of the new detector using the vendor's predicted performance for D^*. The expected noise equivalent power would be

$$\mathrm{NEP}_p = \frac{\sqrt{A_n}\sqrt{f_n}}{D^*_p} = \frac{0.03[cm]\sqrt{0.35\cdot 10^6[Hz]}}{3.3\cdot 10^{10}[cmHz^{\frac{1}{2}}\mathrm{W}^{-1}]} = 5.38\cdot 10^{-10}[\mathrm{W}] \tag{6.42}$$

and the expected noise equivalent change in temperature due to the detector specs would be:

$$\begin{aligned}\mathrm{NE}\Delta\mathrm{T}_\mathrm{p} &= \frac{\mathrm{NEP}_\mathrm{p}}{\Delta\Phi/\Delta T} = \frac{\mathrm{NEP}_\mathrm{p} G\#_n}{\Delta L/\Delta T\cdot A_n} \\ &= \frac{5.38\cdot 10^{-10}[\mathrm{W}]32[\mathrm{sr}^{-1}]}{8.41\cdot 10^{-5}[\mathrm{Wcm}^{-2}\mathrm{sr}^{-1}\mathrm{K}^{-1}](0.03\ [\mathrm{cm}])^2} = 0.2[K]\end{aligned} \tag{6.43}$$

where the $G\#$ of the new system is

$$G\#_n = \frac{4F\#^2}{\tau_f\tau_o\pi} = \frac{4\cdot 4^2}{0.8\cdot 0.8\pi} = 32\ [\mathrm{sr}^{-1}] \tag{6.44}$$

If we predict a preamp noise (N_{ap}) of 0.3 [K] (due to low signal and high gain), we would have a system noise expressed in apparent temperature of

$$\begin{aligned}N_p &= \left[(\mathrm{NE}\Delta\mathrm{T}_\mathrm{p})^2 + (N_{ap})^2\right]^{\frac{1}{2}} \\ &= \left[(0.2\ [\mathrm{K}])^2 + (0.3\ [\mathrm{K}])^2\right]^{\frac{1}{2}} = 0.4\ [\mathrm{K}]\end{aligned} \tag{6.45}$$

This expected performance is slightly worse than our specification. In addition, we have been doing simplified calculations with mean, or in some cases peak, values for our figures of merit, rather than spectrally integrated expected values over the bandpass. This will tend to exaggerate our performance figures. Nevertheless, we decided to proceed based on the detector vendor's efforts to exceed specifications and our plans to improve the preamp to reduce noise. Finally, we held out, as a rather costly fallback option, the possibility of using three or four detectors and a *time delay and integration* (*TDI*) approach. This would involve three detectors arranged on the focal plane in the along-scan direction in much the same way as the Landsat MSS multispectral detectors are arranged (cf. Fig. 6.10). In this case, rather than collecting multispectral images of the scene, we acquire multiple copies of the same scene that could be pixel shifted (time delayed) to be in registration and then averaged (integration) to form an image with reduced noise. The noise in the resulting image will be reduced by the square root of the number of detector elements. In our case with three detectors, this should yield

$$N_p(\mathrm{TDI}) = \frac{\mathrm{N_p}}{\sqrt{3}} = \frac{0.4\ [\mathrm{K}]}{\sqrt{3}} = 0.23\ [\mathrm{K}] \tag{6.46}$$

which would be an acceptable noise figure for this system. The relatively large cost of this option was due to two incremental detectors, preamplifiers and recorder input electronics, as well as the cost of having the detectors mounted in the dewar. While not large in absolute dollars, it would nearly triple the cost of this supposedly low-budget upgrade.

So, a much younger version of your author held his breath and ordered the single detector, hoping to keep the costs and his fledgling career on track. Knowing we would not be detector noise limited, we asked the vendor to look in particular for a high-responsivity detector and to push the D^* value up if possible. The resultant detector specifications were a specific detectivity (D_n^*) of $9.6 \cdot 10^{10}$ [$cmHz^{1/2}W^{-1}$] and a responsivity (R_n) of $11.9 \cdot 10^3$[VW^{-1}].

These numbers made the single detector option continue to look feasible. The $NE\Delta T_n$ of the new detector should be

$$NE\Delta T_n = \frac{\frac{NEP_n}{\Delta\Phi}}{\Delta T} = \frac{NEP_n G\#}{\Delta L / \Delta T \cdot A_n}$$
$$= \frac{1.85 \cdot 10^{-10}[W](32[sr^{-1}])}{8.41 \cdot 10^{-5}[Wcm^{-2}\ sr^{-1}\ K^{-1}](0.03[cm])^2} = 0.1K \tag{6.47}$$

where

$$NEP_n = \frac{\sqrt{A_n}\sqrt{f_n}}{D^*_n} = \frac{0.03[cm]\sqrt{0.35 \cdot 10^6[Hz]}}{9.6 \cdot 10^{10}[cmHz^{\frac{1}{2}}W^{-1}]} = 1.85 \cdot 10^{-10}[W] \tag{6.48}$$

This indicates that we should see similar noise levels from the new detector as from the original.

The relative signals from the new detector compared to the original should be

$$\frac{S_n}{S_o} = \frac{A_n}{A_o}\frac{R_n}{R_o}\frac{t_{sn}}{t_{so}}\frac{\tau_n}{\tau_o}$$
$$= \frac{(0.03\ [cm])^2}{(0.075\ [cm])^2}\frac{11.9 \cdot 10^3[V/W]}{1306\ [V/W]}\frac{1.45 \cdot 10^{-6}[sec]}{3.6 \cdot 1^{-6}[sec]}\frac{0.8}{(0.9)(0.7)} = 0.75 \tag{6.49}$$

indicating that gain levels from the new preamp will only need to be slightly higher than the old. The actual observed noise level for the entire system after assembly was approximately 0.3 K, just meeting our design specifications.

Figure 6.38 shows a portion of an image acquired with the upgraded system, along with a simulated image of how the same scene would have appeared if imaged by the original system. Note that in addition to the aesthetic differences, small details would not have been observed in the original scene. For example, the shape/structure of the hose/cable running from the support vehicle to the airplane is difficult to see in the 2.5 milliradian image.

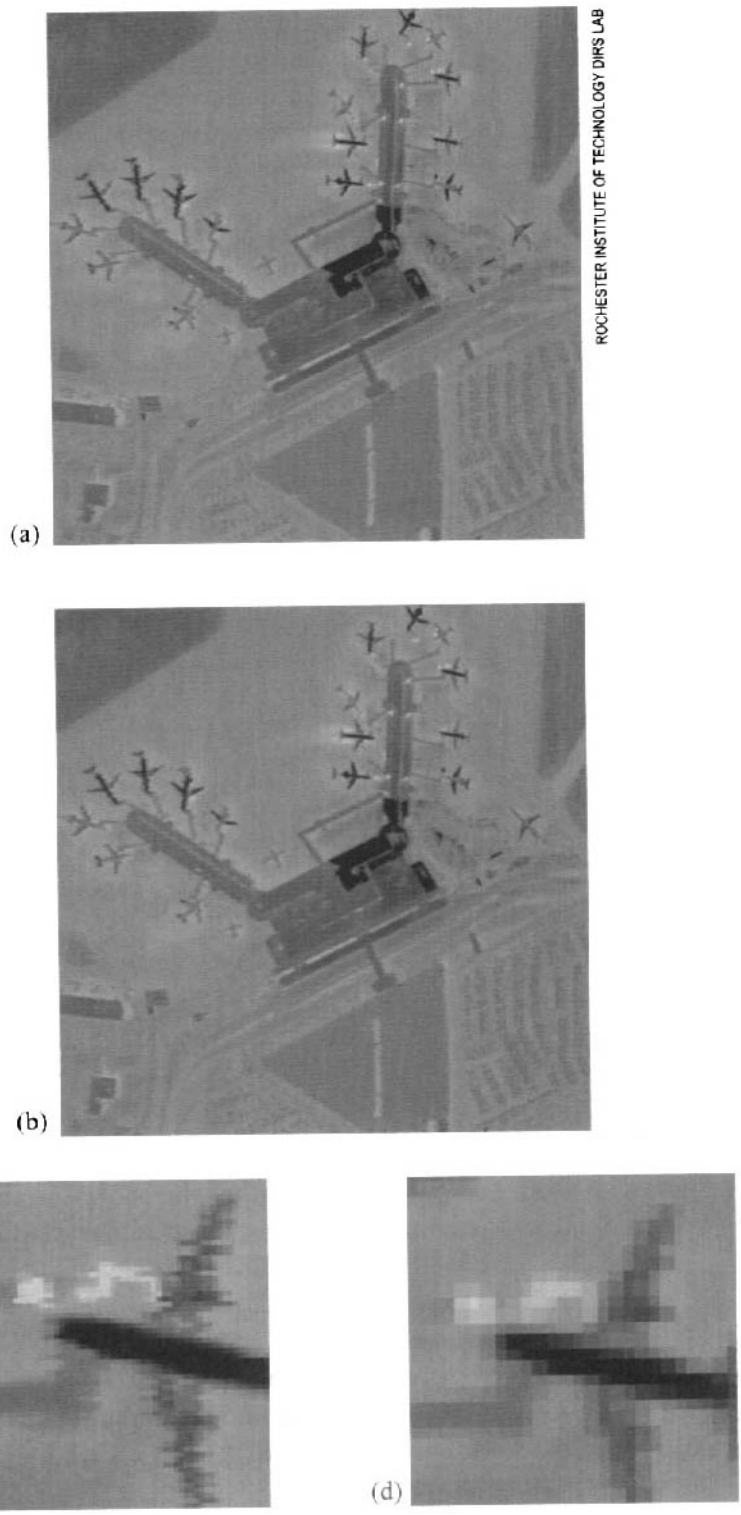


Figure 6.38 (a)Portion of LWIR image acquired with an infrared line scanner after sensor upgrade, and (b) simulation of how scene would have been imaged by the original system. Enlargements (c) and (d) show a portion of image (a) and (b), respectively. Note that the line-to-line variations have not been compensated for in this image.

6.6 REFERENCES

Arvidson, T., Gasch, J., & Goward, S.N. (2001). Landsat 7's long-term acquisition plan—an innovative approach to building a global imagery archive. *Remote Sensing of Environment*, Vol. 78, No. 1-2, pp. 13-26.

Barker, J.L., ed. (1985). "Proceedings of the Landsat-4 Science Characterization Early Results Symposium." Feb. 1983, NASA Goddard, NASA Conference Publication 2355.

Barker, J.L., Abrams, R.B., Ball, D.L., & Leung, K.C. (1985a). "Radiometric calibration and processing procedures for reflective bands on Landsat-4 Protoflight Thematic Mapper." Proceedings of the Landsat-4 Science Characterization Early Results Symposium, NASA, Goddard, 1983, NASA Conference Publication 2355, pp. 47-86.

Barker, J.L., Ball, D.L., Leung, K.C., & Walker, J.A. (1985b). "Prelaunch absolute radiometric calibration of the reflective bands on the Landsat-4 Protoflight Thematic Mapper." Proceedings of the Landsat-4 Science Characterization Early Results Symposium, NASA, Goddard, 1983, NASA Conference Publication 2355, pp. 277-372.

Chen, H.S. (1985). *Space Remote Sensing Systems. An Introduction.* Academic, Orlando, FL.

Clodius, W.B., Bender, S.C., Kay, R.R., Smith, B.W., Atkins, W.H., Christensen, R.W., Little, C.K., Salewski, E.F., & Rappoport, W.M. (1999). MTI on-orbit calibration. *Proceedings SPIE*, Vol. 3753, pp. 380-391.

Clodius, W.B., Bender, S.C., Atkins, W.H., Christensen, R.W., Little, C.K., Kay, R.R., & Bridenstine, D. (2000). Initial MTI on-orbit calibration performance. *Proceedings SPIE*, Vol. 4132, pp. 290-305s.

Cook, M.K., Peterson, B.A., Dial, G., Gerlach, F., Hutchins, K., Kudola, R., & Bowen, H.S. (2001). IKONOS technical performance assessment. *Proceedings SPIE*, Vol. 4381, pp. 94-108.

Feng, X., Schott, J.R., & Gallagher, T.W. (1994). Modeling the performance of a high-speed scan mirror for an airborne line scanner. *Optical Engineering*, Vol. 33, No. 4, pp. 1214-1222.

Folkman, M.A., Pearlman, J., Lushalan, B.L., & Jarecke, P.J. (2001). EO-1/Hyperion hyperspectral imager design, development, characterization and calibration. *Proceedings SPIE*, Vol. 4151, pp. 40-51.

Goetz, A.F.H., Vane, G., Solomon, J.E., & Rock, B.N. (1985). Imaging spectrometry for earth remote sensing. *Science*, Vol. 228, pp. 1147-1153.

Goward, S.N., Masek, J.G., Williams, D.L., Irons, J.R., & Thompson, R.J. (2001). The Landsat 7 mission: Terrestrial research and applications for the 21st century. *Remote Sensing of Environment,* Vol. 78, Nos. 1-2, pp. 3-12.

Grum, F., & Becherer, R.J. (1979). *Optical Radiation Measurements: Vol. 1 Radiometry*. Academic Press, NY.

Guenther, B., Xiong, X., Solomonson, V., Barnes, W.L., & Young, J. (2002). Performance of the earth observing system moderate resolution imaging spectroradiometer, first year of data. *Remote Sensing of Environment*, Vol. 83, pp. 16-30.

Hackwell, J.A., Warren, D.W., Bongiovi, R.P., Hansel, S.J., Hayhurst, T.L., Mabry, D.J., Sivjee, M.G., & Skinner, J.W. (1996). LWIR/MWIR imaging hyperspectral sensor for airborne and ground-based remote sensing. *Proceedings SPIE*, Vol. 2819, pp. 102-107.

Hemphill, W.R., Stoertz, G.E., & Markle, D.A. (1969). Remote sensing of luminescent materials. Proceedings of the 6th International Symposium of Remote Sensing of Environment, University of Michigan, Ann Arbor, MI.

Kramer, H.J. (2001). *Observation of the Earth and Its Environment: Survey of Missions and Sensors*. Springer-Verlag, Berlin.

Krause, K.S. (2004). Relative radiometric characterization and performance of the QuickBird high-resolution commercial imaging satellite. *Proceedings SPIE*, Vol. 5542, pp. 35-44.

Markham, B.L., & Barker, J.L., eds. (1985). Special Landsat image data quality assessment (LIDQA). *Photogrammetric Engineering and Remote Sensing*, Vol. 51, No. 9.

Markham, B.L., Schafer, J.S., Wood, F.M., Jr., Dabney, P.W., & Barker, J.L. (1998). Monitoring large aperture spherical integrating sphere sources with a portable radiometer during satellite instrument calibration. *Metrologia*, Vol. 35, pp. 643-648.

Meigs, A.D., Butler, E.W., Jones, B.A., Otteen J.L. III, Sellar, R.G., & Rafert, B., (1996). Airborne visible hyperspectral imaging spectrometer: Optical and system-level description. *Proceedings SPIE*, Vol. 2819, pp. 278-284.

Morain, S.A., & Budge, A.M., eds. (2001). *The Manual of Remote Sensing,* 3rd ed. Earth Observing Platforms and Sensors, ASPRS, Bethesda, MD.

Norwood, V.R., & Lansing, J.C., Jr. (1983). Electro-optical imaging sensors, Chapter 8, in *Manual of Remote Sensing*, 2nd ed., Colwell, R.N. ed. American Society of Photogrammetry, Falls Church, VA.

Rao, C.R.N., & Chen, J. (1993). Calibration of the visible and near infrared channels of the advanced very high resolution radiometer (AVHRR) after launch. *Proceedings SPIE*, Vol. 1938, pp. 56-66.

Rickard, L.J. (1993). HYDICE: An airborne system for hyperspectral imaging. *Proceedings of SPIE*, Vol. 1937, pp. 173-179.

Schott, J.R., Barsi, J.A., Raqueño, N.G., & deAlwis, D. (2001). Calibration and analysis of Landsat thermal data for water resource studies. *Remote Sensing of Environment*, Vol. 78, Nos. 1-2, pp. 108-117.

Schott, J.R., Brown, S.D., & Barsi, J.A. (2004). Calibration of Thermal Infrared (TIR) Sensors. In *Thermal Remote Sensing in Land Surface Processes*. J. Luvall & D. Quattrochi, eds. CRC Press, Boca Raton, FL.

Slater, P.N., Biggar, S.F., Holm, R.G., Jackson, R.D., Mao, Y., Moran, M.S., Palmer, J., & Yuan, B. (1987). Reflectance and radiance-based methods for in-flight absolute calibration of multispectral sensors. *Remote Sensing of Environment*, Vol. 22, pp. 11-37.

Stevenson, B.P., Kendall, W.B., Stellman, C.M., & Olchowski, F.M. (2003). PHIRST light: A liquid crystal tunable filter hyperspectral sensor. *Proceedings SPIE*, Vol. 5093, pp. 104-113.

Storey, J.C., and Choate, M. (2002) Landsat 7 initial on-orbit geometric calibration and performance. *Proceedings SPIE*, Vol. 4049, pp. 143-154.

Teillet, P.M., Slater, P.N., Ding, Y., Santer, R.P., Jackson, R.D., & Moran, M.S. (1990). Three methods for absolute calibration of the NOAA AVHRR sensors in-flight. *Remote Sensing of Environment*, Vol. 31, pp. 105-120.

Teillet, P.M., Barker, J.L., Markham, B.L., Irish, R.R., Fedosejevs, G., & Storey, J.C. (2001). Radiometric cross-calibration of the Landsat-7 ETM+ and Landsat 5 TM sensors based on tandem data sets. *Remote Sensing of Environment,* Vol. 78, Nos. 1-2, pp. 39-54.

Thome, K.J. (2001). Absolute radiometric calibration of Landsat 7 ETM+ using the reflectance-based method. *Remote Sensing of Environment*, Vol. 78, Nos. 1-2, pp. 27-38.

Vane, G., Green, R.O., Chrien, T.G., Enmark, H.T., Hansen, E.G., & Porter, W.M. (1993). The airborne visible/infrared imaging spectrometer (AVIRIS). *Remote Sensing of Environment*, Vol. 44, pp. 127-143.

Vodacek, A., Kremens, R., Fordham, A., VanGorden, S., Luisi, D., Schott, J.R., & Latham, D. (2002). Remote optical detection of biomass burning using a potassium emission signature. *International Journal of Remote Sensing*, Vol. 23, No. 13, pp. 2721-2726.

Watson, R.D. (1981). Quantification of luminescence intensity in terms of rhodamine WT standard. In Workshop on Applications of Luminescence Techniques to Earth Resources Studies, W.R. Hemphill & M. Settle, eds. LPI Tech Rpt. 81-03. Lunar and Planetary Institute, Houston, TX, pp. 19-21.

Watson, R.D., & Theisen, A.F. (1981). Electronic and optical modification of the engineering model FLD and the evaluation of peripheral equipment. In Workshop on Applications of Luminescence Techniques to Earth Resources Studies, W.R. Hemphill & M. Settle, eds. LPI Tech Rpt. 81-03. Lunar and Planetary Institute, Houston, TX, pp. 15-18.

Wolfe, W.L. (1997). Introduction to imaging spectrometers, tutorial tests. In *Optical Engineering*, vol. TT25. SPIE Press, Bellingham, WA.

Wrigley, R.C., Hlavka, C.A., Card, D.H., & Buis, J.S. (1985). Evaluation of Thematic Mapper interband registration and noise characteristics. *Photogrammetric Engineering and Remote Sensing*, Vol. 51, No. 9, pp. 1417-1425.

Xiong, X., Sun, J. Q., Esposito, J.A., Guenther, B., & Barnes, W.L. (2003a). MODIS reflective solar bands calibration algorithm and on-orbit performance. *Proceedings SPIE*, Vol. 4891, pp. 392-401.

Xiong, X., Chiang, K., Guenther, B., & Barnes, W.L. (2003b). MODIS thermal emissive bands calibration algorithm and on-orbit performance. *Proceedings SPIE*, Vol. 4891, pp. 95-104.

Yarbrough, S., Caudill, T.R., Kouba, E.T., Osweiler, V., Arnold, J., Quarles, R., Russell, J., Otten, L.J. III, Jones, B.A., Edwards, A., Lane, J., Meigs, A.D., Lockwood, R.B., & Armstrong, P.S. (2002). MightSat II.1 hyperspectral imager: Summary of on-orbit performance. *Proceedings SPIE*, Vol. 4480, pp. 186-197.

This page intentionally left blank

CHAPTER 7
ATMOSPHERIC COMPENSATION: SOLUTIONS TO THE GOVERNING EQUATION

In Chapter 4, we developed the governing equation Eq. (4.63) for the radiance reaching a remote sensing system, the "big equation." In chapter 6, we discussed these remote sensing systems and how they could be calibrated using laboratory or on-board calibrators. As a result of this calibration process, we will assume that, when required, we can convert the remotely sensed signal (i.e., digital count, voltage, or film density) to radiance reaching the front of the sensor. This means that for calibrated sensors, we will assume that we can solve for the radiance reaching the sensor at any image location. This yields the radiance value on the left-hand side of the big equation. In this chapter, we address procedures for solving for the remaining variables in the big equation so that we can associate ground temperatures or reflectance values with each point (pixel) in an image. This process of atmospheric compensation is one of the most difficult tasks facing the remote sensing community. At present, no single approach has proven sufficiently simple, accurate, and robust enough to be widely accepted and operationally used. As a result, a number of methods for atmospheric compensation exist that are useful for a particular type of problem or accuracy level. In this chapter, we will look at a range of compensation approaches aimed at covering most remote sensing situations. These approaches are sometimes referred to in the literature as "atmospheric correction techniques." Because our overall goal is to have a "calibration" system that relates image digital counts to surface parameter (e.g., reflectance or temperature), we will refer to this process as "atmospheric compensation" and treat it as part of the calibration or preprocessing process.

The reader should keep in mind the fundamental assumption behind the value of atmospheric compensation, i.e., that we have identified a relationship between the quantitative value of one or more remotely sensed variables (e.g., temperature and/or spectral reflectance) and one or more parameters of interest (e.g., vegetation vigor). While the functional form of this relationship is very application specific and therefore beyond the scope of this effort, we will cite a few examples here to provide motivation for this process. Gordon et al. (1983) presents a relationship between the reflectance of the ocean and water quality parameters. Piech et al. (1978) cite a correlation between the reflectance of water in the visible region and several water-quality variables (e.g., chlorophyll, suspended solids, and gelbstoffe) and used this relationship to map chlorophyll variations in fresh water using calibrated reflectance values. Piech and Walker (1974) show relationships between soil texture and moisture and reflectance values and suggest how moisture and texture signatures can be separated using broad-band spectral reflectance values. The geologic community has long used laboratory reflectance spectra to assist in the characterization of rock samples, and Marsh and Lyon (1980) show the correlation between surface data and satellite-derived reflectance values. Numerous studies have shown relationships between reflectance and various vegetation parameters, including chlorophyll concentration [cf. Thomas and Gausman (1977)], water content [f. Thomas et al. (1966)] and leaf area index (LAI) [cf. Suits (1972)]. In order to fully utilize these relationships, it is first necessary to remove the effects of the atmosphere.

Similarly, Schott and Wilkinson (1982) discuss how absolute temperature measurements can be related to heat loss from buildings. Heilman and Moore (1980) discuss how temperature is related to soil moisture, and numerous studies have shown relationships between water temperature and water circulation [cf. Legeckis (1978)]. This small sampling of studies shows that laboratory and field reflectance (and/or temperature) values have been correlated with a variety of parameters of interest to applications scientists. The development and use of these relationships are major undertakings, with many application-specific factors confusing these relationships. As a result, we will not address the development of these application-specific relationships here. Instead, we want to stress that one of the goals of the remote sensing community is to be able to make quantitative measurements of reflectance and temperature so that these measurements can be directly used by the application scientists developing laboratory and field correlations between these variables and application-specific parameters (e.g., vegetation condition, and soil moisture). In order to quantitatively measure reflectance or temperature, we must first account for the effect of the atmosphere on the observed radiance.

In this chapter, we consider atmospheric compensation methods employing field measurements (ground truth) at the time of overflight, methods employing measurements from the images themselves (in-scene methods), and methods using radiation propagation models. We also consider some compensation methods used to normalize atmospheric effects between multiple images of the same site from different days (multidate normalization methods) and image-processing methods designed to reduce the apparent effects of the atmosphere on the output image.

7.1 TRADITIONAL APPROACH: CORRELATION WITH GROUND-BASED MEASUREMENTS

A labor-intensive but often effective way to calibrate a remote sensor is to acquire ground truth at the time of image acquisition. Ground truth can take on many forms, but its purpose is to improve the analyst's ability to extract information from the remotely sensed images. In some cases, this means taking measurements so that the atmospheric parameters in the governing equation are known. It can also involve measurements to define the relationship between the remotely sensed variable (e.g., reflectance) and the parameter under study (e.g., stress in a forest canopy). Or, in many cases, ground truth can involve extensive measurements to directly characterize, usually quantitatively, the parameter(s) under study (e.g., soil moisture).

One common way to calibrate an image is to skip the atmospheric compensation step and attempt to develop directly a functional relationship between the remotely sensed values and the parameter under study using observed ground truth values for the parameter. This often takes on the form

$$Y = f(\mathrm{DC}_1, \mathrm{DC}_2 ... \mathrm{DC}_N) \tag{7.1}$$

where Y is the parameter under study (e.g., soil moisture), $\mathrm{DC}_1 \cdots \mathrm{DC}_N$ are the digital count values in spectral bands 1 through N (e.g., red, green, and blue), and f simply denotes that there is a functional relationship between the dependent parameter Y and the observed values ($\mathrm{DC}_1 \cdots \mathrm{DC}_N$). In many cases, no form is known or presumed for the function f, and a number of regressions are tried to minimize the residual errors in equations of the form

$$Y_i = a_0 + a_1 \mathrm{DC}_{1i} + a_2 \mathrm{DC}_{2i} + a_N \mathrm{DC}_{Ni} + e \tag{7.2}$$

$$Y_i = b_0 + b_1 \log \mathrm{DC}_{1i} + b_2 \log \mathrm{DC}_{2i} ... + b_N \log \mathrm{DC}_{Ni} + e \tag{7.3}$$

$$Y_i = c_0 + c_1 \mathrm{DC}_{1i} + c_2 \frac{\mathrm{DC}_{1i}}{\mathrm{DC}_{2i}} + c_3 \mathrm{DC}_{2i} + c_4 \frac{\mathrm{DC}_{2i}}{\mathrm{DC}_{3i}} ... + c_{2N} \frac{\mathrm{DC}_{(N-1)i}}{\mathrm{DC}_{Ni}} + e \tag{7.4}$$

where the subscript i indicates that the data values are coming from the ith ground truth location corresponding to image location i having digital count values DC_{1i} to DC_{Ni}. The values a_0 to a_N, b_0 to b_N, etc., are regression coefficients arrived at by including all sets of parameters and independent digital count values, and e is the residual error of the regression. The regression coefficients are chosen to minimize the residual error (e) for the postulated function, and then the functional form (f) with the lowest mean-square error is selected. In general, a random search for a functional form for the relationship between an independent variable (Y) and digital count values should only be performed if large amounts of data are available and no physical form for the nature of the relationship can be postulated. Even then, the functional form of the relationship should be kept simple, or an overconstrained nonrobust solution can easily result. Such a solution might generate a relatively

small error value (e) for the ground truth data but might generate very large errors when applied to other digital count values elsewhere in the scene. The only way to ensure any degree of robustness to solutions of this form is to have a great deal of data (compared to the degrees of freedom in the regression equation) and to make sure, within reason, that all combinations of digital count values are included in the sample data set. This is generally a prohibitively expensive solution, especially because it is only valid for the data set under study. This is because the regression coefficients will be a function of the atmospheric conditions at the time of image acquisition and will change from acquisition to acquisition. As a result, a major ground truth program, potentially involving hundreds of ground truth samples, is required each time imagery is collected.

This type of regression approach can often be pursued with more confidence and requires less ground truth if the nature of the relationship between reflectance and the parameter of interest (Y) is known. In this case, we can break the overall functional form up into subrelations of the form

$$\mathrm{DC} = f'(L) \tag{7.5}$$

$$L = f''(r) \tag{7.6}$$

$$Y = f'''(r_1, r_2 \ldots r_N) \tag{7.7}$$

or

$$Y = f(\mathrm{DC}_1, \mathrm{DC}_2 \ldots \mathrm{DC}_N) \tag{7.8}$$

In many cases, the functions f' and f'' are approximately linear, such that

$$\mathrm{DC}_1 = m_1' L_1 + b_1' \tag{7.9}$$

$$L_1 = m_1'' r_1 + b_1'' \tag{7.10}$$

where m_1' and b_1' are the linear coefficients relating the radiance in spectral band 1 (L_1) to the digital count in band 1 (DC_1) and m_1'' and b_1'' are the linear coefficients that relate the reflectance in band 1 (r_1) to the radiance in band 1 (L_1). Then, if from laboratory or field studies a functional form for f''' can be postulated, e.g.,

$$Y \cong a_0 + a_1 r_1 + a_2 r_2 \tag{7.11}$$

it can be shown that

$$Y_i = b_0 + b_1 \mathrm{DC}_{1i} + b_2 \mathrm{DC}_{2i} + e \tag{7.12}$$

Again, a least-squares regression can be used to find the coefficients b_0, b_1, etc.; however, in this case the functional form used is defined by the form of Eq. (7.11). The robustness of the solution can be ascertained from the laboratory or field data used to derive Eq. (7.11), and a much smaller ground truth program can be used. This method has been used for a variety of applications. Johnson et al.

(1981), for example, used this general approach to map chlorophyll concentrations for water-quality studies. For this approach to work satisfactorily, it is necessary to have a wide range of ground truth data that corresponds to the dynamic range of the parameter being investigated. Likewise, there is an implicit assumption that the regression coefficients apply equally well over the entire image (i.e., the atmosphere is spatially constant). This general approach also requires the reflectance values to be effectively Lambertian over the sensor's field of view unless special efforts are made to account for any non-Lambertian behavior.

In order to take advantage of a broad range of functional relationships between reflectance and applications-oriented parameters, more general ground truth compensation methods are often employed. These methods are aimed at computation of the necessary atmospheric parameters to solve directly for the value of the surface reflectance in each spectral band. These reflectance values can then be used to study any number of functions of the form of Eq. (7.7) where the nature of the function is known or can be derived at a later time from laboratory or field studies. These approaches are more appropriately called *atmospheric calibration techniques* and are treated in the thermal region in Section 7.3.1 and in the reflective region in Section 7.4.1

Note that while the discussion of the ad hoc image compensation methods in this section (7.1) have emphasized the reflective bands, one need only replace reflectance with temperature as the variable of interest and the approach can be used for calibration of thermal imagery, as well.

7.2 APPROACHES TO ATMOSPHERIC COMPENSATION

Many compensation techniques exist that are targeted at different spectral regions, sensor configurations, or operational constraints. We will discuss a sampling that will provide the reader with a starting point for coping with most collection conditions. The treatment is not intended to be all inclusive, nor were the methods necessarily selected because they are clearly the most accurate or widely used. In fact, there are very few widely used or generally accepted compensation methods, in large part because so few have been objectively tested over a robust data set. This chapter makes a concerted effort to describe all of the methods in terms of a self-consistent nomenclature using standard radiometric terminology to facilitate intercomparison.

Often in the sections that follow (7.3 and 7.4), we are assuming that the image data are from a calibrated sensor. By this we mean that the recorded signal (typically in terms of digital counts, analog voltage, or film density) can be converted to the effective radiance reaching the imaging system. In several cases, the image compensation methodology combines sensor calibration for sensors where radiance and recorded signal are linearly related. As mentioned in chapter 6, increasing care in recent years has gone into the design and implementation of sensor calibration. In addition, operators of satellite sensors have initiated postlaunch, as well as periodic checks, of ongoing sensor calibration, as discussed in Section 6.4. Regrettably, these improvements in sensor calibration can only be fully utilized if

the remaining unknowns associated with atmospheric and illumination effects can be adequately computed or compensated.

There is no clear definition of what constitutes adequate atmospheric compensation. Each user will have a different definition of what calibration level is required for his or her application. Where expected errors associated with a calibration method are clearly demonstrated, we will attempt to cite them. However, the reader should be aware that in many cases the cited numbers are often from a limited study, usually performed by the developer of the method, who may have used several "ad hoc tricks" not clearly defined in the simplified treatment presented here.

Depending on one's perspective, atmospheric compensation can be thought of as either forward or reverse engineering. In the forward-engineering case, the objective is to predict the radiance reaching the sensor and compare that to the sensed radiance to evaluate the calibration of the sensor. This would be our perspective if we were interested in postlaunch or periodic evaluation of on-orbit satellite systems (cf. Sec. 6.4). The more common way to analyze the data would be to assume that the sensor provides accurate radiance values and then to work backward (reverse engineering) to extract information about the absolute or relative reflectance or the absolute or relative temperature of scene elements. Both approaches require similar information about atmospheric conditions. However, the forward-engineering sensor calibration methods can assume that a great deal of ground truth and ancillary data are available, since they only need to be used occasionally and at well-controlled locations. The more common reverse-engineering methods would normally be used under operational conditions where ground truth and ancillary data may be limited or nonexistent. Since most users are in the position of assuming that the sensor is calibrated and need to extract information from the imagery, this chapter emphasizes the reverse-engineering approaches.

The compensation approaches can be further divided based on the type of information required. The most rigorous calibration approaches are aimed at extracting absolute surface reflectance or temperature values. However, in many cases, only relative reflectance or temperature values are needed. For example, in multispectral analysis, it may be sufficient to know the reflectance ratio between two spectral bands without needing to know the reflectance in either band. At other times, it may only be necessary to know the change in reflectance over time. For example, in change detection studies or process monitoring studies, it is often sufficient to know how much the reflectance of an object changed over time or to be able to plot relative reflectance values as a function of time. Finally, there are times when it is only necessary to reduce atmospheric effects to improve the potential for comparing data over time or within a scene. For example, the brightness values of forest pixels on opposite sides of a hill are different, and a method to normalize these illumination variations is required (i.e., neither the actual magnitudes nor change in magnitude must be known). Sections 7.3 and 7.4 address various methods for compensation of image data to temperature and reflectivity, respectively.

In these sections we will, for simplicity, often assume that only small errors are introduced by using effective bandpass values for radiance, transmission, etc. (cf. Sec. 4.5). The reader is cautioned to check the validity of this assumption

against the calibration precision required and use more accurate spectral solutions if necessary and possible.

7.3 APPROACHES TO MEASUREMENT OF TEMPERATURE

This section addresses methods for compensation of thermal images where the objective is temperature measurement. In general, when considering retrieval of temperature, we will assume that we are sensing in the long-wave infrared (LWIR) region or in the midwave infrared (MWIR) region at night, since this is where most quantitative thermal imaging is undertaken (cf. Sec. 4.6). The methods considered will be ground truth, in-scene methods, and atmospheric modeling approaches. In each case, one or more examples of compensation approaches will be presented. These are not intended to be comprehensive, but only an attempt to show some of the more common classes of compensation techniques and to point out the general pros and cons.

7.3.1 Ground Truth Methods (Temperature)

The ground truth methods presented here and in Section 7.4.1 should not be confused with the methods described in Section 7.1. Both approaches use ground truth data; however, the methods presented here are aimed at calculation of atmospheric calibration values (e.g., τ and L_u), whereas the methods in Section 7.1 are aimed at simple one-time ad hoc correlation solutions between digital count values and application-oriented parameters.

The ground truth methods are the most common and straightforward approaches to atmospheric compensation. Simply put, they involve measuring the temperature or radiance from several objects in the study scene and then computing the transmission and path radiance. For a blackbody, the governing equation can be approximated as

$$L = L_T e^{-\delta' \sec\theta} + L_u = L_T \tau_2 + L_u \tag{7.13}$$

For a calibrated sensor, the observed radiance values (L) can be computed from the recorded signal levels and regressed against the radiance due to the temperature (L_T) of the measured objects. This approach was used effectively by Scarpace et al. (1974), who set up several children's wading pools that could be easily resolved and whose water temperatures could be varied and measured (cf. Fig. 7.1). For near-vertical viewing in the LWIR, the emissivity of water is quite high ($\approx$0.986), so the blackbody approximation introduces little error. The assumption is made that the calibration measurements are made on level surfaces free from background effects, so the shape factor F in Eq. (4.70) is set to 1. In addition, since their interest was in studying water temperature alongside the calibration site, any residual error due to reflected downwelled radiance would have been included as part of L_u and corrected for in the calibration, i.e.,

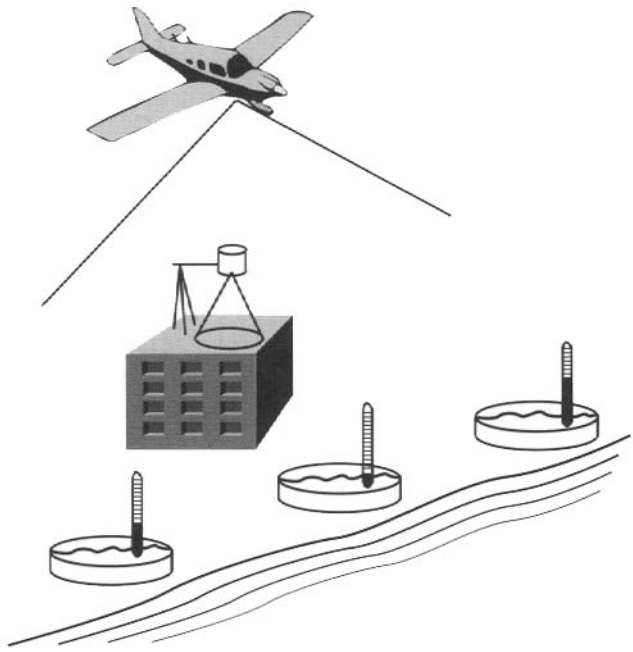


Figure 7.1 Temperature and radiance ground truth measurements can both be used to calibrate infrared images.

$$L = L_T \varepsilon \tau_2 + \tau_2 r L_d + L_u = m L_T + b \tag{7.14}$$

where $m = \tau_2 \varepsilon$ is the slope of the regression and $b = \tau_2 r L_d + L_u$ is the intercept.

In practice, the water temperature in the study area adjacent to the calibration site was computed by solving for L_T, i.e.,

$$L_T = (L - b)/m \tag{7.15}$$

and converting the radiance to temperature through the Planck equation.

This ground truth method will work well for blackbodies and will be self-correcting for gray bodies if it is only applied to objects with the same emissivity as the calibration standards (e.g., in this case the term $\tau_2 r L_d$ is a constant for the pools and the water body under study). On the other hand, a more involved approach is necessary for surfaces having varying emissivity.

In principle, the value of εL_T for ground truth studies can be computed from contact thermistors and emissivity values (ε) determined by laboratory or field measurements, or by using normal "textbook" values [cf. Gubareff et al. (1960) and Salisbury and D'Arian (1992)]. This approach is seldom used in practice because of the difficulty associated with making accurate contact temperature measurements of solids (particularly surfaces with low thermal conductivity) and the limited emissivity database for naturally occurring backgrounds. The problem with using solids as standards is compounded by timing problems associated with taking ground truth. Because the surface temperatures can change rapidly with

environmental conditions, the data must be taken concurrent with the sensor overflight for precise calibration.

To reduce the error associated with temperature and emissivity measurements, ground truth of solid surfaces is more appropriately acquired with a calibrated radiometer (ideally one with the same spectral bandpass as the overflight sensor). By taking several radiometer readings of targets with different radiance values, a simple linear regression can be performed to yield τ_2 and L_u from

$$L = \tau_2(\varepsilon L_T + rL_d) + L_u = \tau_2(L_0) + L_u \tag{7.16}$$

where L_0 is the surface-leaving radiance measured by the radiometer. Regrettably, because of the thermal instability of most surfaces, the surface radiance readings must still be concurrent with the overflight, and care must be taken to avoid background radiation from neighboring objects (including the experimenter), which would change the value of L_d (assumed constant in the above discussion).

Given τ_2 and L_u from Eq. (7.16), L_T and hence T can be computed for any scene object by solving for L_0 and rearranging to yield

$$L_T = (L_0 - rL_d)/\varepsilon \tag{7.17}$$

where ε is determined by ground truth or by using tabulated values. Ideally, L_d can be determined by a radiometer, which integrates the downwelled radiance from the sky in the spectral band of interest. Alternately, nomograms have been developed (Bell et al., 1960) that can be used to approximate L_d based on air temperature and atmospheric conditions. A more rigorous solution for L_d can be obtained by using an atmospheric propagation model such as MODTRAN (cf. Sec. 7.3.3). In general, a fair amount of error in downwelled radiance (L_d) can be tolerated, since the term $r\tau_2 L_d$ is usually small compared to εL_T. For example, if L_d is 20% of L_T and r is 0.1, a 10% error in L_d will result in an error of significantly less than 1% in L_T.

When quality ground truth data are used and emissivities well characterized, surface temperature errors of less than several tenths Kelvin should be achievable. However, because of a variety of logistical problems, it is often very difficult to acquire quality ground truth data. It can be particularly difficult to acquire data for low-resolution systems where the average temperature or radiance of large surfaces must be known with great accuracy. In addition, there are often occasions when, for operational reasons, it is not possible or cost-effective to take ground truth. This forces us to look for alternative methods for calibrating thermal infrared images.

7.3.2 In-Scene Compensation Techniques (Temperature)

A family of atmospheric compensation techniques has emerged that take advantage of characteristics of the data in the images themselves for calibration. These approaches are known as *in-scene* methods and often rely on multiple ways of looking at the same scene. We will look at three of these methods that involve multiple images of the same scene taken at multiple altitudes, multiple view angles, or multiple wavelengths.

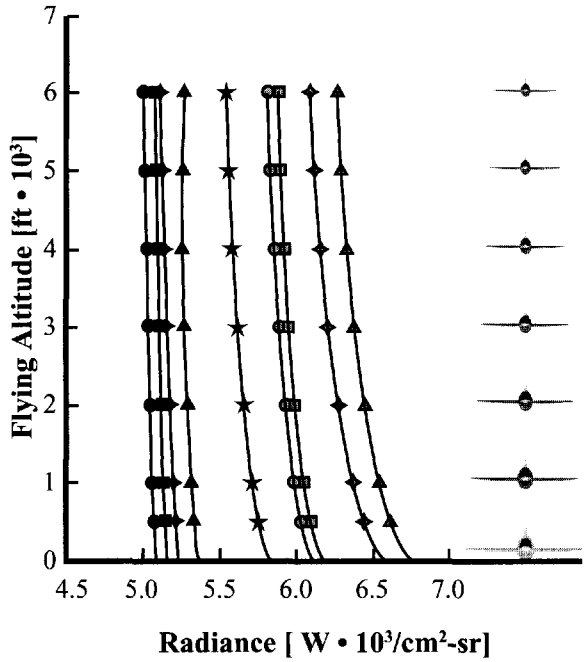


Figure 7.2 Data collected using the multiple-altitude (profile) calibration technique. Symbols represent sampled points for each target, and the line through each point is extrapolated to zero altitude.

7.3.2.1 Profile (Multiple Altitude) Compensation Technique

This technique was developed to eliminate the problems associated with ground truth approaches. It is used on airborne systems and involves flying the imaging system at a series of altitudes above the same ground area. When analyzing the images, targets having a range of radiance values are identified and their radiance measured at each altitude from the flight altitude down to the lowest practical altitude. These radiance values are then plotted as a function of altitude and extrapolated to zero altitude, as shown in Figure 7.2. The data are usually collected near vertical (such that $\theta = 0$). Thus, the radiance at the flight altitude can be expressed as

$$L(h,\theta) = L(h,0) = \tau(h,0)L(0,0) + L_u(h,0) \tag{7.18a}$$

$$= mL(0,0) + b \tag{7.18b}$$

where the extrapolated radiance at the surface is $L(0,0)$, and we have expressed the functional dependence on altitude (h) and view angle (θ) for the relevant variables. If the radiance at altitude $L(h,0)$ is regressed against the radiance for that same object extrapolated to the surface value $L(0,0)$ for each of several objects, then the slope (m) is the atmospheric transmission $\tau(h,0)$ and the intercept (b) is the

upwelled radiance $L_u(h,0)$. In effect, we use the data extrapolated to zero altitude as ground truth. To first order, the transmission at any angle can be approximated as [cf. Eq. (4.2)]

$$\tau(h,\theta) \approx \tau(h,0)^{\sec(\theta)} \tag{7.19}$$

and the upwelled radiance can be approximated as

$$L_u(h,\theta) \approx L_u(h,0)\sec(\theta) \tag{7.20}$$

for clear atmospheres and small θ (less than 60°).

The downwelled radiance can be computed using the MODTRAN model (cf. Sec. 7.3.3). However, an alternative approach is to observe a non-Lambertian surface of known angularly dependent emissivity such as water at two angles. If the reflected downwelled radiance is assumed to be approximately constant, then the following analysis can be performed:

$$L(h,\theta) = \varepsilon(\theta)L_T\tau(h,\theta) + r(\theta)L_d\tau(h,\theta) + L_u(h,\theta) \tag{7.21}$$

and

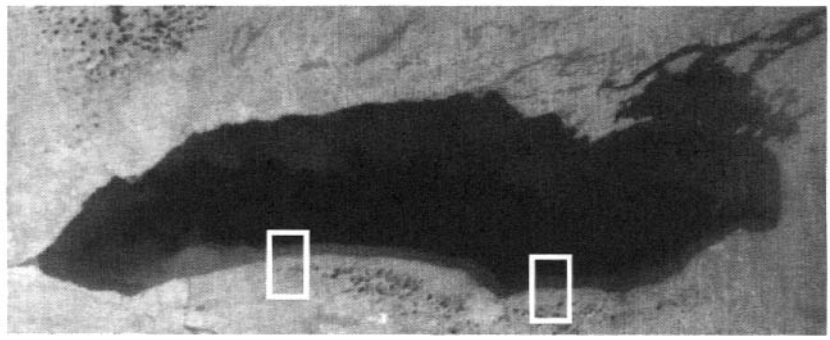
$$L(h,0) = \varepsilon(0)L_T\tau(h,0) + r(0)L_d\tau(h,0) + L_u(h,0) \tag{7.22}$$

Equating the L_T terms for the same object viewed at two angles and solving yields an expression for L_d [cf. Schott, (1979)]. Experiment has shown that this method is not particularly accurate. However, as discussed in Section 7.3.1, for most surfaces the reflected portion of the signal is a small component, so the overall analysis is not particularly sensitive to errors in L_d.

Schott (1979) reports on a blindfold test where the profile technique was used in the 8–14 μm region to compute water surface temperatures. The standard error using the profile method was 0.4 K when compared to simultaneously acquired ground truth values. This technique has also been used to acquire "ground truth" for satellite imaging systems. Schott and Schimminger (1981) utilized aircraft imagery calibrated with the profile technique to measure the radiance of large areas that had been simultaneously measured with a satellite with a 0.5-km GIFOV. The satellite data could then be accurately calibrated to approximately 1.1 K using this approach (cf. Fig. 7.3). The profile technique has the advantage of demonstrated accuracy and no requirements for external data. It has a significant disadvantage in that it cannot be utilized with satellite imaging systems except in the rare case where a simultaneous aircraft underflight is available.

7.3.2.2 Multiple View Angle Techniques

Saunders (1967), McMillan (1975), and Chedin et al. (1982), among others, have reported on methods to extract surface temperatures using some form of multiple-view angle technique. The method presented here was suggested by Byrnes and Schott (1986) and is presented mostly for its consistency with the previous methods. All the methods rely on the assumption that the radiance from a target



(a)

(b)

Figure 7.3 (a) A Heat Capacity Mapping Mission (HCMM) image of Lake Ontario with aircraft underflight coverage shown in white boxes. (b) Thermal map produced after underflight calibration (°C). The warm ring of near-shore water is associated with the spring thermal bar. [Adapted from Schott and Schimminger (1981).]

observed at two different look angles changes only due to the difference in path length through an atmosphere composed of homogeneous layers (cf. Fig. 7.4). If, for convenience, we take one angle to be vertical, then the radiance reaching the sensor from the same target can be expressed for viewing through an angle θ as

$$L(h,\theta) = \tau(h,\theta)L(0,\theta) + L_u(h,\theta) \tag{7.23}$$

and for vertical viewing as

$$L(h,0) = \tau(h,0)L(0,0) + L_u(h,0) \tag{7.24}$$

From Chapter 3, recall that for Lambertian surfaces the surface-leaving radiance into the θ direction $[L(0,\theta)]$ and the radiance normal to the surface $[L(0,0)]$ are the same. Thus, Eqs. (7.23) and (7.24) can be combined by eliminating the surface radiance term to yield

$$L(h,\theta) = \frac{\tau(h,\theta)}{\tau(h,0)} L(h,0) + \left[L_u(h,\theta) - \frac{\tau(h,\theta)}{\tau(h,0)} L_u(h,0) \right] \tag{7.25}$$

Plate 1.1 Photographic image of the earth acquired by Skylab astronauts.

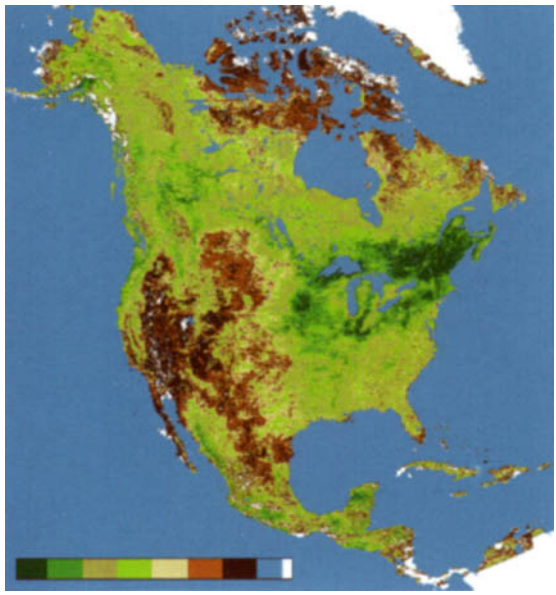


Plate 1.2 Map of the normalized difference vegetative index of North America derived from AVHRR data.

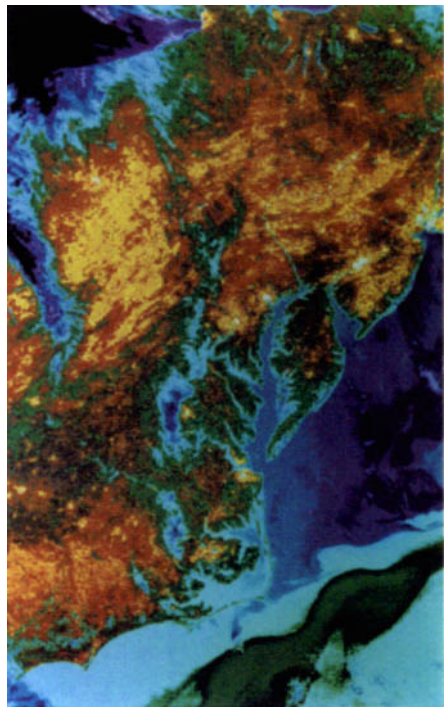
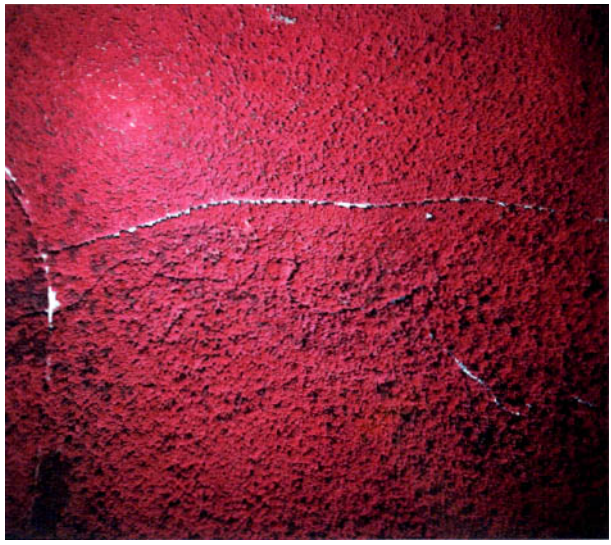


Plate 1.3 Color coded image of eastern U.S. produced from the thermal channel of the Heat Capacity Mapping Mission (HCMM) radiometer. Note the warm urban areas and the warmer water in the Gulf Stream.

(a)



(b)

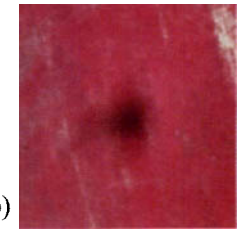


Plate 1.4 (a) Color-infrared aerial photograph of a forested area. Note the "hot spot" to the top left where the line of sight is along the sun-sensor-target line such that the scene elements obscure their own shadow. (b) A zoom of the center of the "hot spot" in a subsequent frame over open fields where the plane's shadow is more defined.

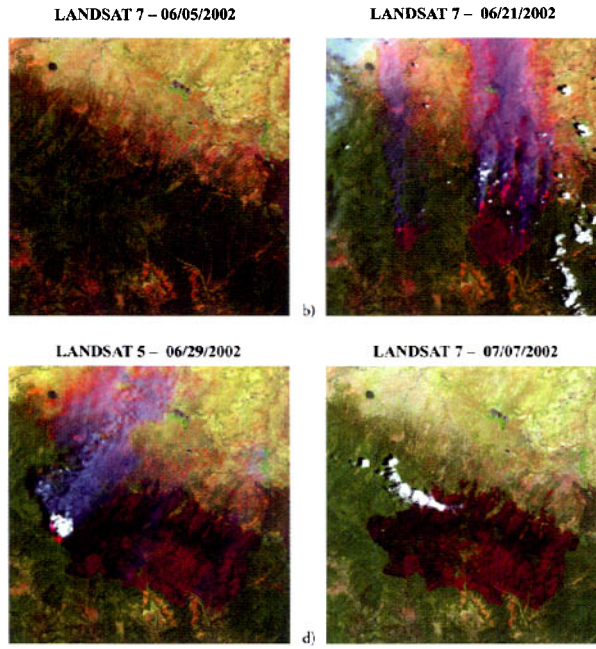


Plate 1.7 Landsat 5 and 7 change sequence showing (a) prefire condition, (b) the Rodeo fire in Arizona (burning for 3 days at this point) and the Chediski fire to the west (burning for 1 day), (c) the combined fires still burning and (d) the extent of the burn scar. The red region shows the burn scar in the band 7, 4, 2 band combination shown.

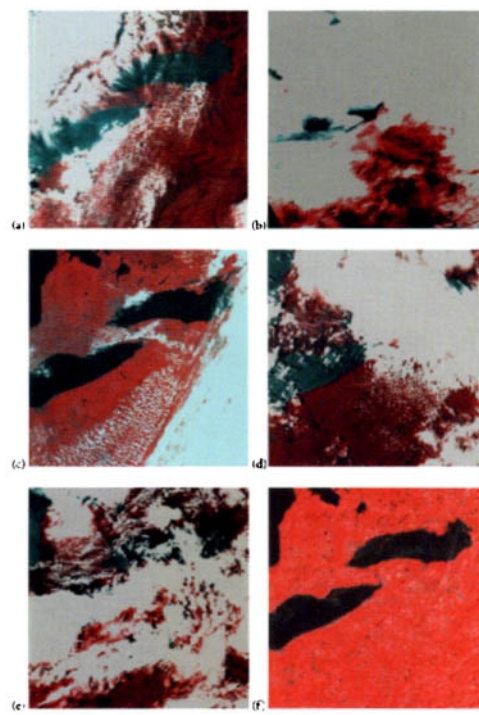


Plate 1.8 Five day sequence of AVHRR images of the Eastern Great Lakes and a cloud-free composite (f) made by combining the five-day sequence.

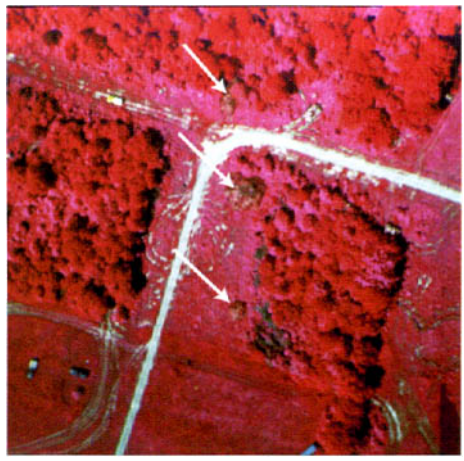


Plate 2.5 Color infrared image of camouflage objects.

Plate 2.6 Color infrared photo showing vegetation condition. A gypsy moth infestation has defoliated most of the trees in the image, "Cosmetic" spray programs along the roadways have prevented severe defoliation in these areas.

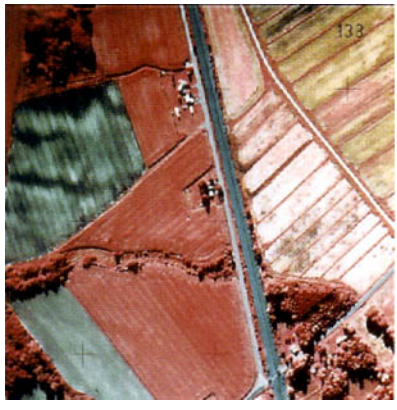


Plate 5.7 Digitized color infrared film image.

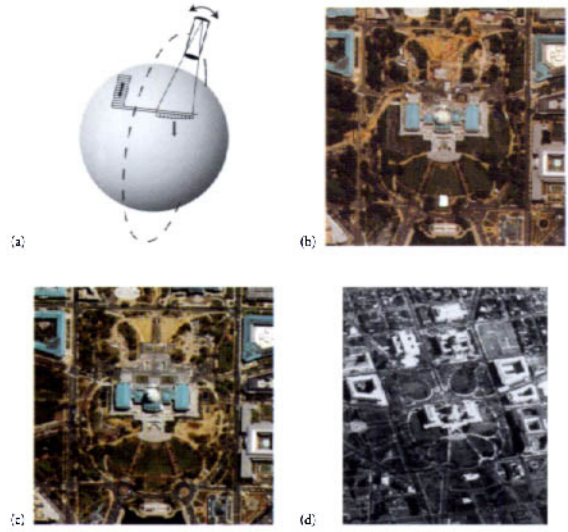


Plate 6.16 Illustration of body pointing push-broom concept (a) and example images: (b) Quickbird panchromatic image of the U.S. Captiol; (c) pan sharpened (see chapter 12 for discussion of pan sharpening) Ikonos color image of the Capitol; (d) For comparison a Corona "spy" satellite photographic image from February 19, 1966 is also included.

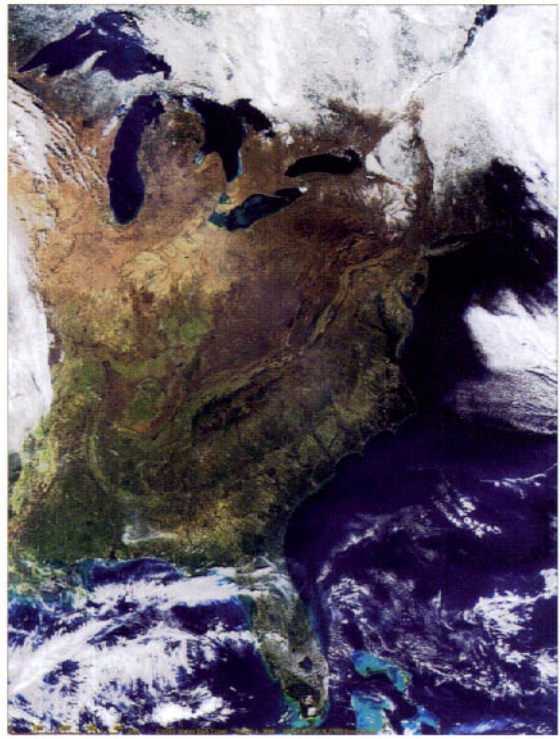


Plate 6.21 True color MODIS image of the East coast of North America acquired March 6, 2000 showing the potential of MODIS to support global and regional Earth resources and Earth process studies.

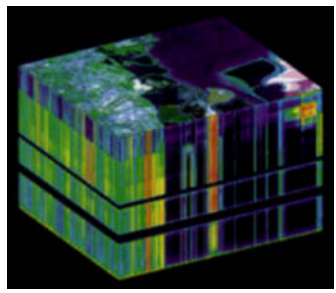


Figure 6.23 AVIRIS image cube of Moffet Field in California. Sensor has 224 channels, from 0.4 μm to 2.5 μm, each with spectral bandwidth of approximately 10 nm.

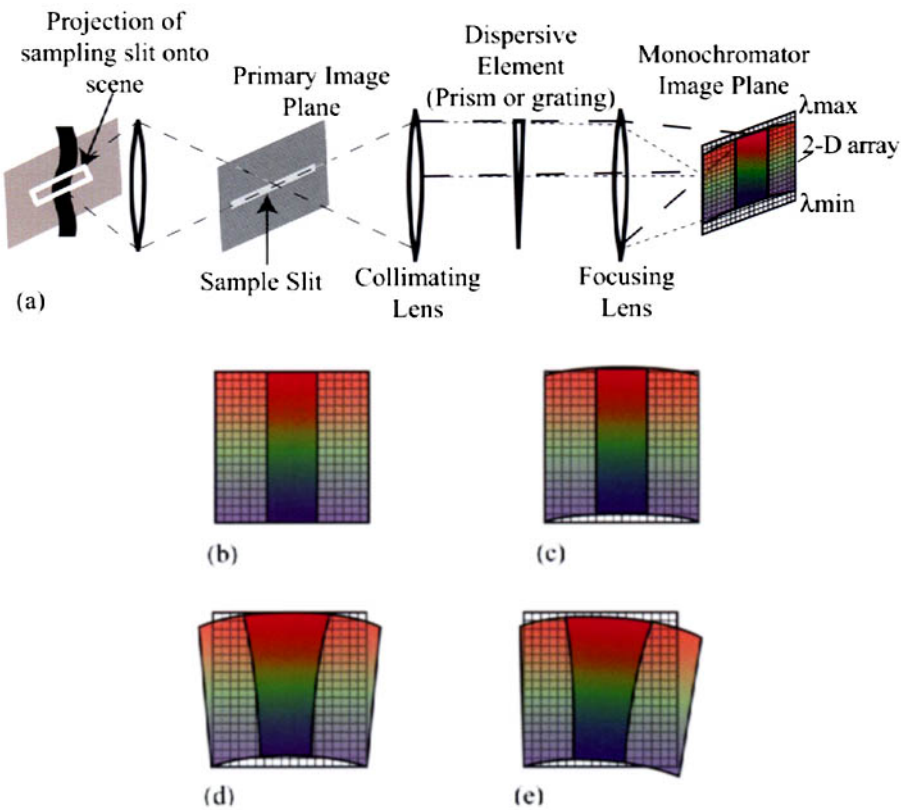
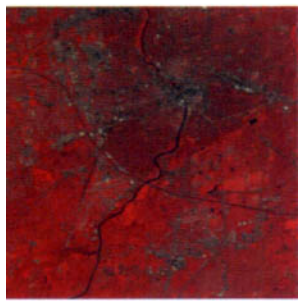
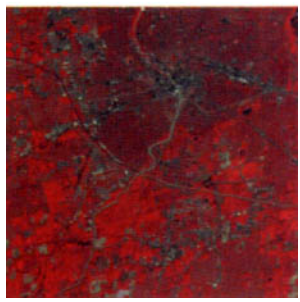


Plate 6.28 Illustration of spectral smile concept and the confounding effect of array misalignment (tilt) (Note that in many designs the collimating and focusing functions use a common lens and the dispersive element is a reflective component.), (a) Source of smile error (note curvature is exaggerated for illustration purposes). (b) Idealized spectrum projected on detector array (c) Spectrum showing smile. (d) Spectrum showing smile and key stoning effects due to optical projection and misalignment. (e) Spectrum showing smile, key stoning, and tilt due to array misalignment.

1982 image



1984 image



1982 image transformed to
"look like" the 1984 image

Plate 7.30 Pseudo-invariant (PIF) transformation applied to a portion of a Landsat scene.

(a)

(b)

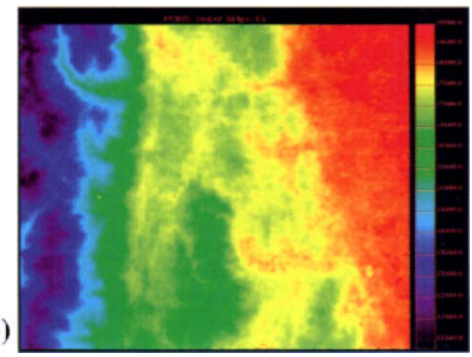


Plate 7.33 (a) AVIRIS image of Jasper Ridge, CA and (b) a map of column water vapor content derived from the 940 nm absorption line data in the AVIRIS spectra.

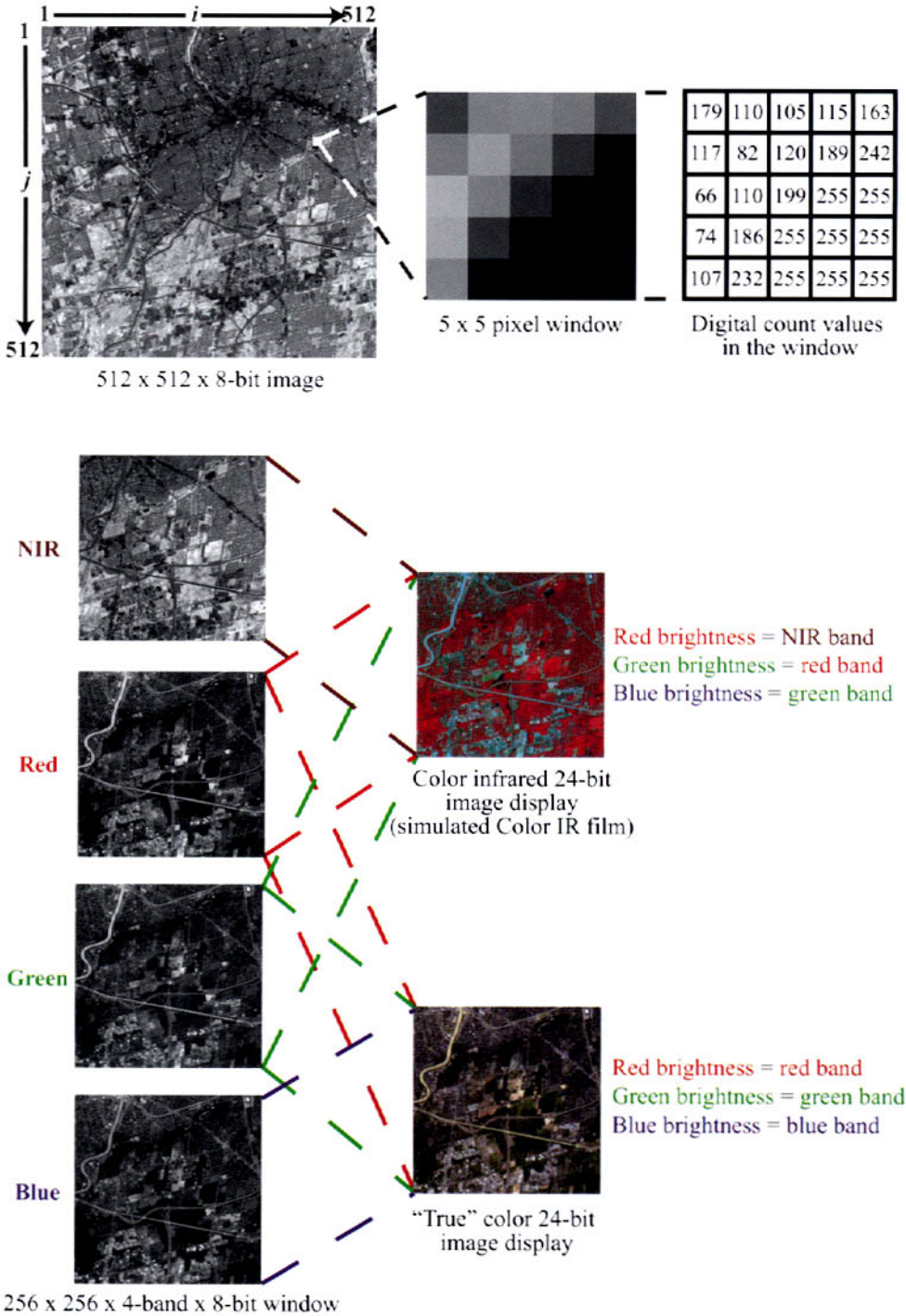


Plate 8.1 Digital image concepts.

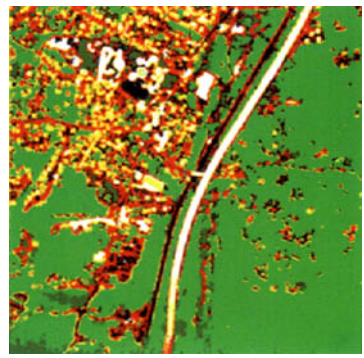


Plate 9.15 Color example of the image shown in Figure 8.10 classified using texture metric images.

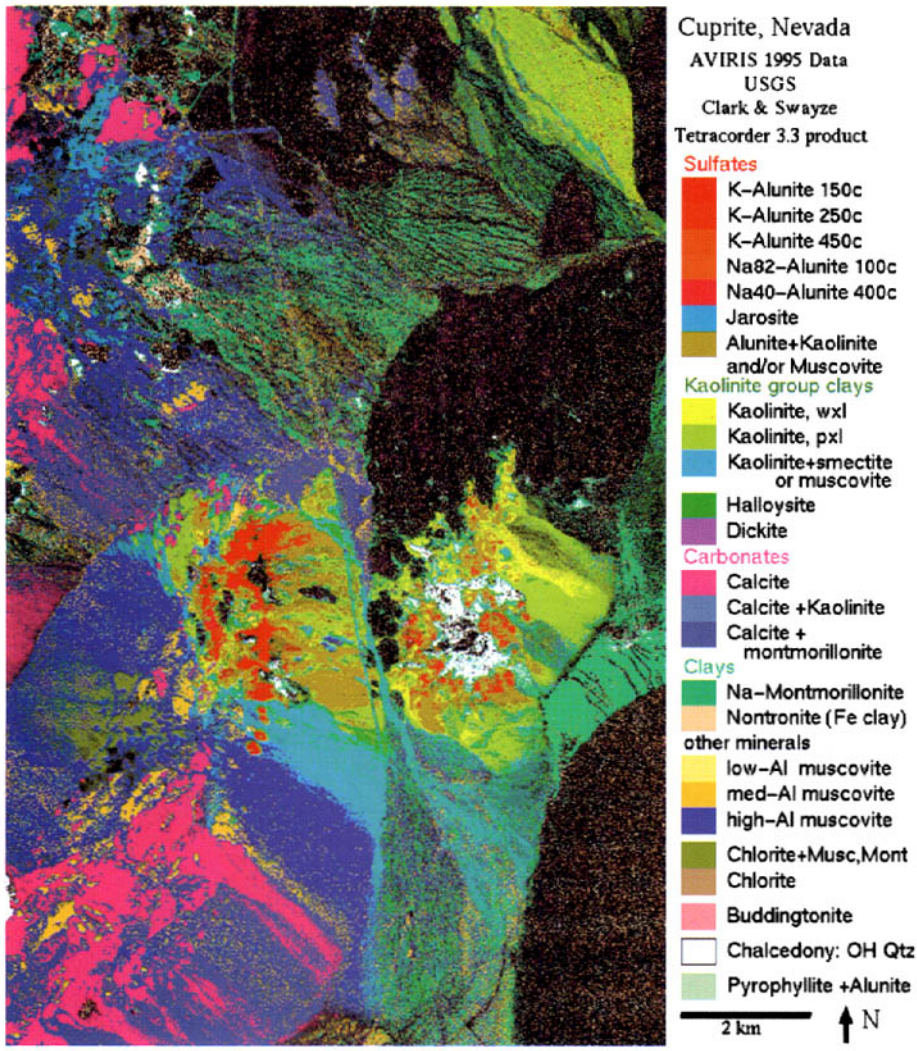


Plate 10.20 Illustration of material map produced by Tetracorder where each pixel is assigned to material with the highest R^2D value.

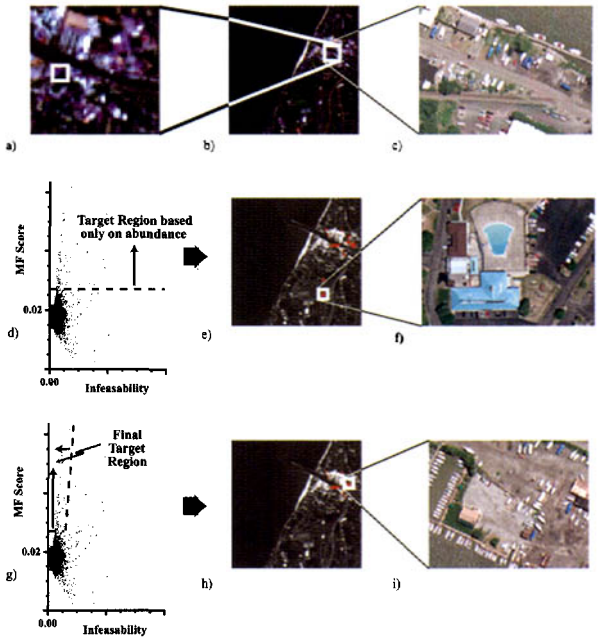


Plate 10.25 Illustration of an application of an MTMF to the detection of blue storm tarps in an AVIRIS image cube of a cluttered scene: (b) a single band of the AVIRIS image with a zoom in (a) and a high-resolution air photo for "truth" in (c). The matched filter results for a target pixel selected from (a) are shown in (e) along with an air photo of false alarms (f) based on matched filter scores (d). (h) Shows the MTMF results for the region shown in (g). An air photo of the selected targets shows the presence of storm tarps.

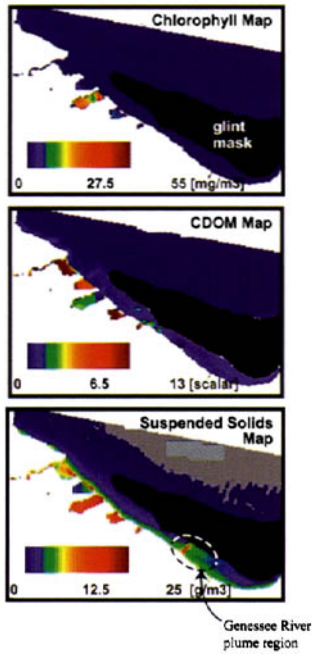


Plate 11.11 Concentration maps of water-quality constituents derived using the model matching method applied to AVIRIS images of the Lake Ontario shoreline. Note the region labeled glint was not analyzed due to solar glint from the surface that can be avoided by proper timing of the image acquisition.

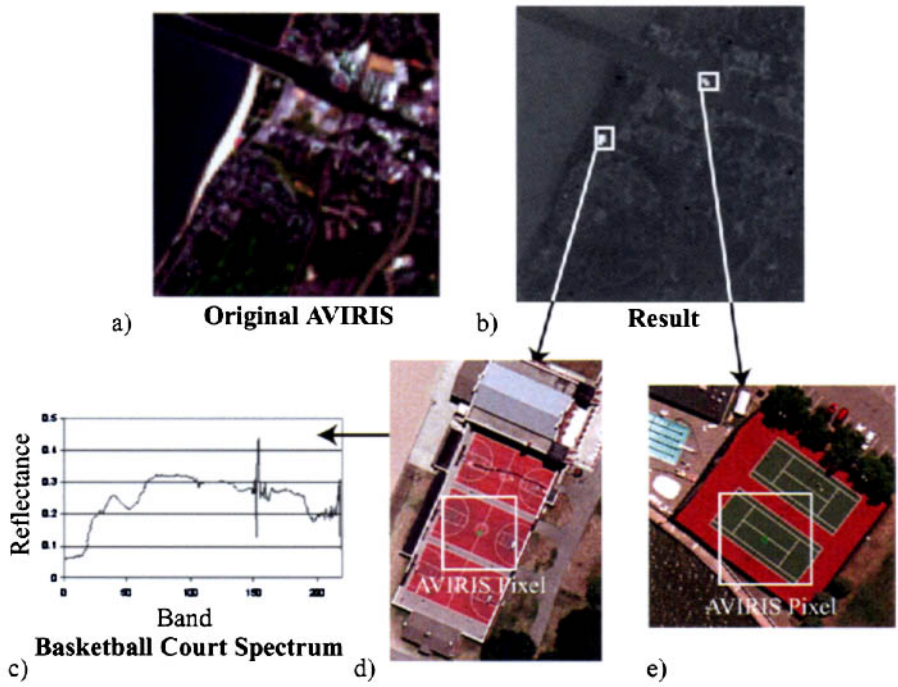


Plate 11.20 Subpixel invariant approach applied to a cluttered scene: (a) a single band from the original image, (b) results of the detection operator, (c) spectrum of the target, (d) high-resolution air photo of the target region showing approximate size of an AVIRIS pixel, and (e) air photo of the second "target" located.

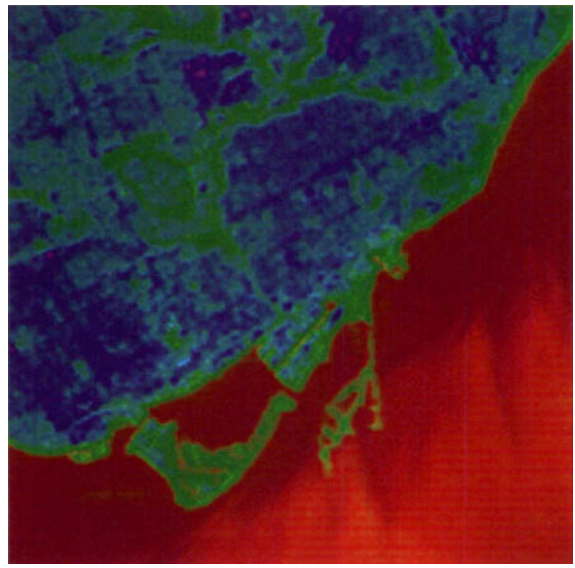


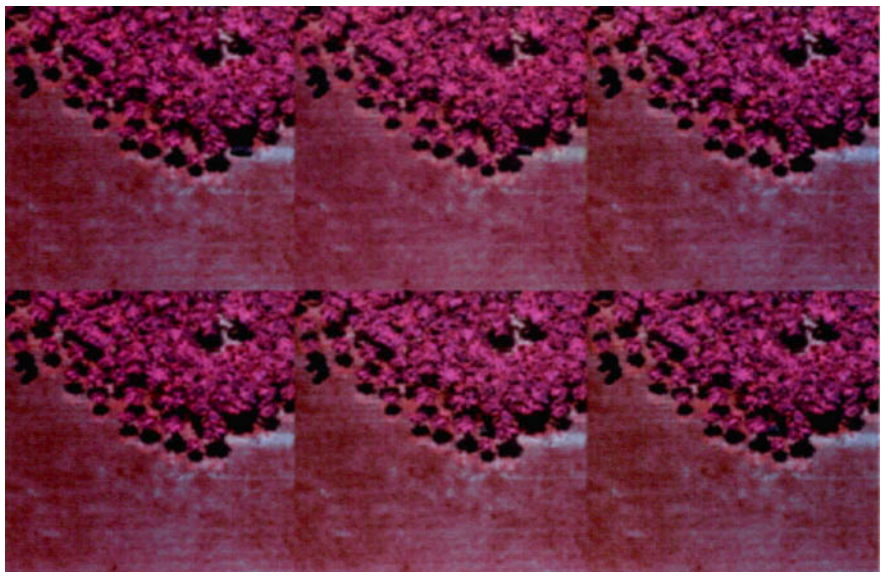
Plate 12.17 Image fusion of the red channel of TM with the LWIR channel using the IHS transform.

(a)

(b)

(c)

Plate 12.18 Fusion of multispectral TM data (a) with geometrically registered and resampled SPOT data (b) to yield the fused product (c).



NASA GODDARD

Plate 14.8 DIRSIG false color infrared image sequence showing a vehicle passing under overhanging trees. These images could be used to predict the vehicle signature under a range of adjacency and obscuration conditions.

Plate 14.16 A true-color synthetic scene produced by DIRSIG showing texture effects. Note this is the same scene shown in the model board in Figure 14.1.

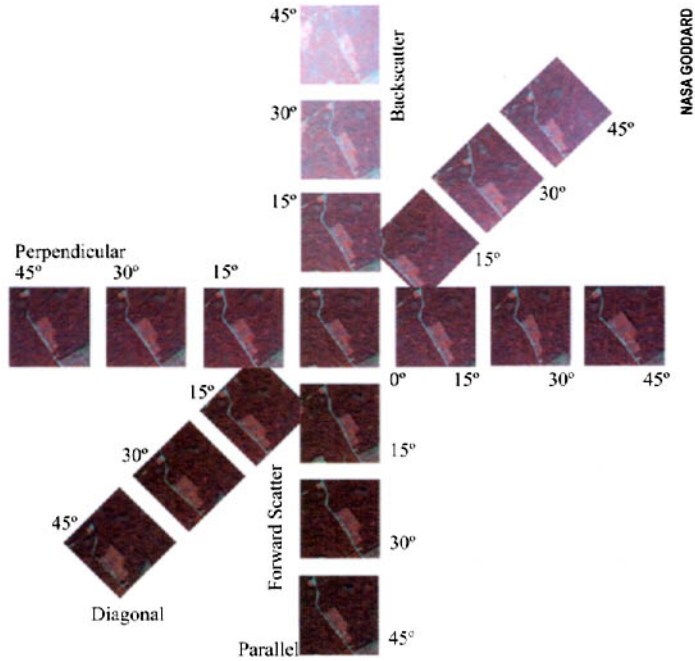


Plate 14.18 Color IR version of ASAS images from three flight lines showing combined effects of atmospheric and bidirectional reflectance variation .

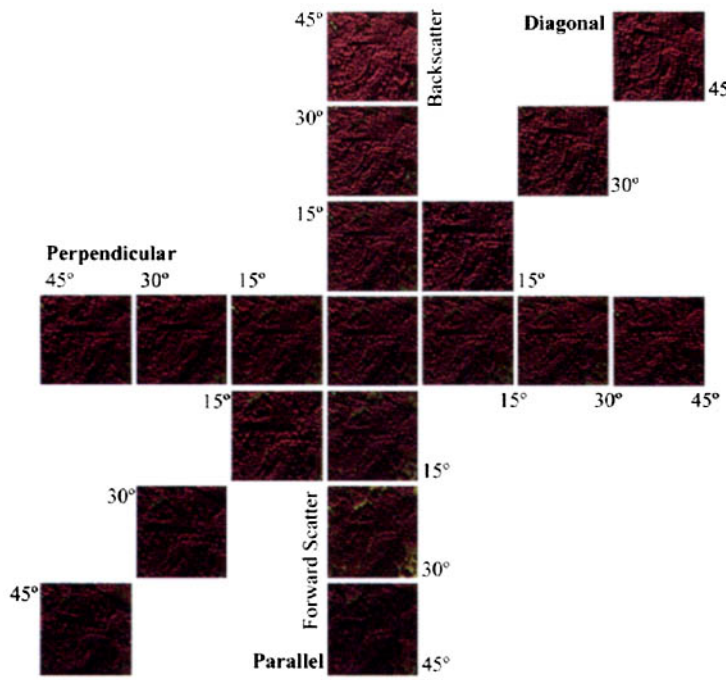


Plate 14.19 DIRSIG simulation of combined atmospheric and bidirectional canopy reflectance, DIRSIG images of a canopy taken using acquisition geometry similar to that used with ASAS to acquire Figure 14.18.

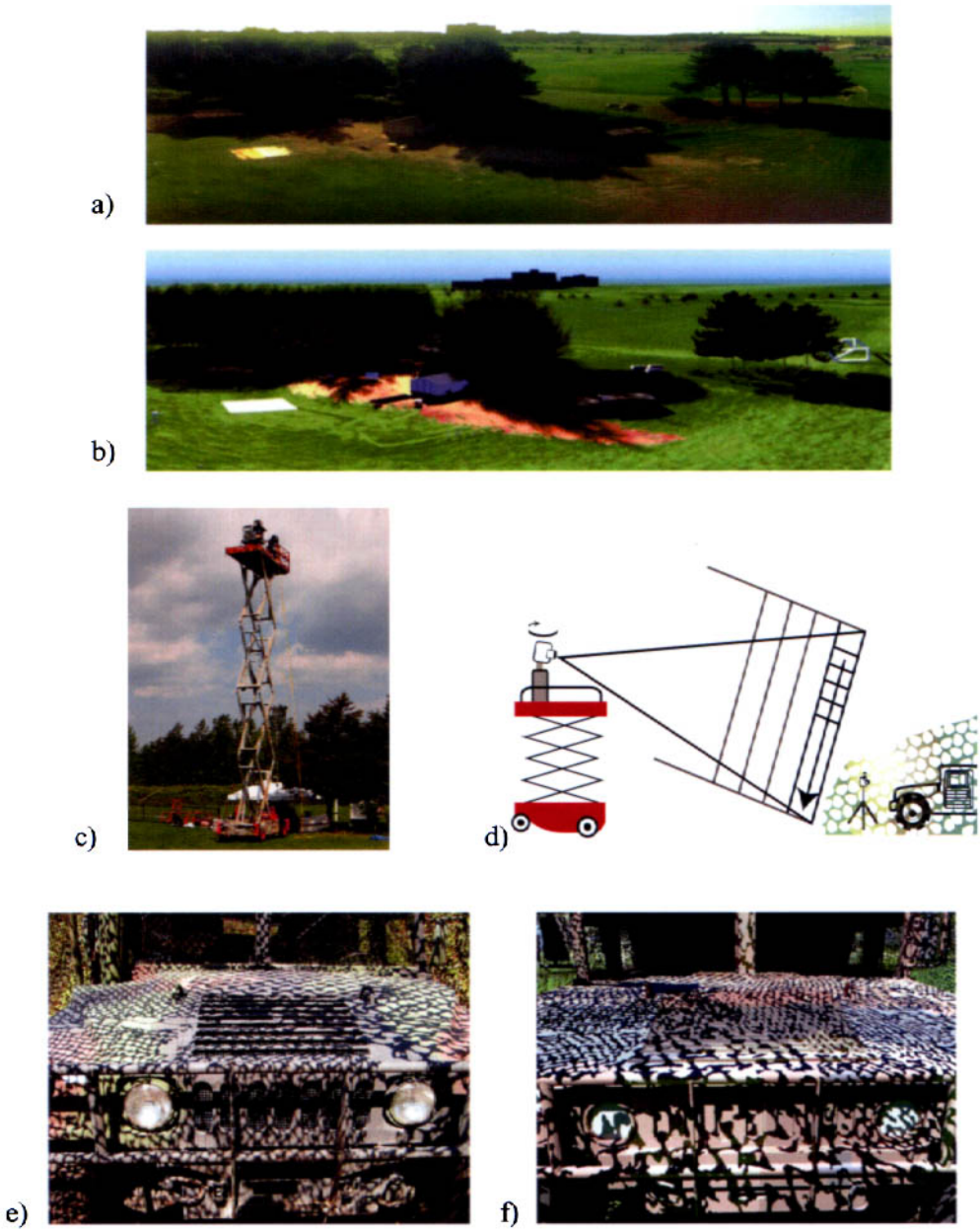


Plate 14.23 Illustration of one band from a real (a) and synthetic (b) hyperspectral line scanner image used for algorithm development. (c) RIT's MISI line scanner used to acquire the real image operating from a rotating table on the scissor truck, as illustrated in (d). (e) and (f) Real and synthetic images from a framing camera on a tripod under the net as illustrated in (d).

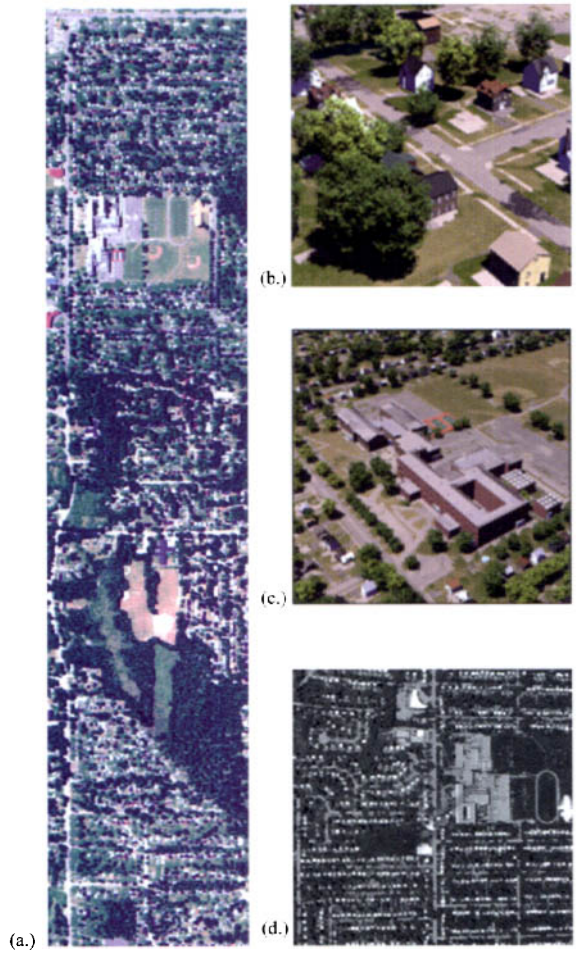


Plate 14.24 (a) True color representation of a portion of a 210-band hyperspectral image of a large synthetic scene built to support hyperspectral algorithm testing and instrument design trades (b) and (c) zooms of a subsection of the scene showing spatial detail available, and (d) portion of the same synthetic scene showing a single band of a synthetic LWIR spectrometer.

Plate 14.25 A synthetic sunset produced by DIRSIG. The effect is possible due to the extensive spectral modeling that DIRSIG incorporates with the help of MODTRAN.

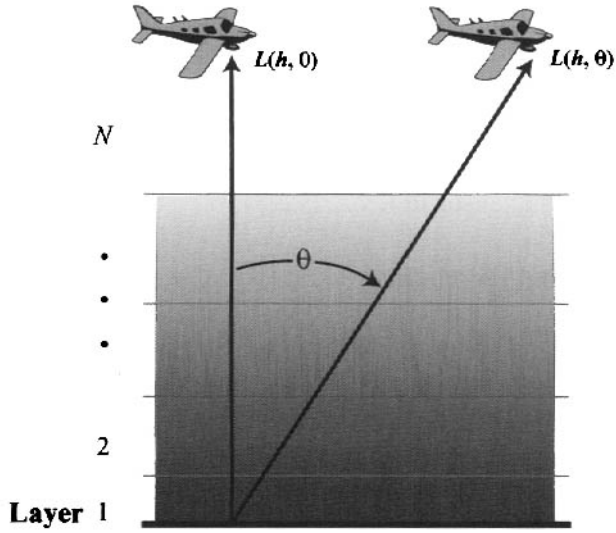


Figure 7.4 The multiple-look-angle calibration technique assumes vertical layering of a horizontally homogeneous atmosphere.

If several Lambertian objects having a range of radiance values are imaged both vertically and at some known angle, then a linear regression of the data will yield a slope term

$$m = \tau(h,\theta)/\tau(h,0) \tag{7.26}$$

and an intercept

$$b = L_u(h,\theta) - \frac{\tau(h,\theta)}{\tau(h,0)} L_u(h,0) \tag{7.27}$$

The slope term can be analyzed using Eq. (7.19) to yield the transmission according to

$$\tau(h,0) = m^{1/(\sec(\theta)-1)} \tag{7.28}$$

In a similar manner using Eq. (7.20), the upwelled radiance for small angles ($\theta = 0°$ to 60°) and relatively clear atmospheres can be approximated as

$$L_u(h,0) = \frac{b}{\sec(\theta) - m} \tag{7.29}$$

Most variations of the multiple-view angle technique use clear atmosphere approximations similar to those of Eqs. (7.19) and (7.20). As a result, the use of this method is often restricted to certain altitudes, atmospheric conditions, or atmospheric types. Nevertheless, the potential value of not requiring ground truth data or low-flying aircraft is a considerable advantage. Chedin et al. (1982) and Byrnes and Schott (1986) both point to potential errors of several degrees Kelvin using simplified approaches to the multiple-view angle technique in the 8–14 μm

bandpass. Mericsko (1993) has suggested empirically derived refinements to this approach that indicate that considerably improved results can be obtained using better approximations to Eqs. (7.19) and (7.20). This approach is still often limited to aircraft use, because most currently operational space-based sensors cannot readily obtain multiple views of the same object through the same atmosphere (see the discussion of the MTI sensor in Sec. 7.3.2.3).

7.3.2.3 Multiple-Bandpass Technique

Prabhakara et al. (1974) and Anding and Walker (1975) discuss a method for computation of surface radiance based on differential absorption in two spectral bands (channels) within a broader atmospheric transmission window. This is very similar in theory to the multiple-angle technique where the differential atmospheric effects are assumed to be well defined due to path length changes. In the multichannel approach, the basic radiative transfer equation is expressed as

$$L(h,\theta) = L(0,\theta)\tau(h,\theta) + L_{TA}[1-\tau(h,\theta)] \tag{7.30}$$

where $L(h,\theta)$ is the radiance at the sensor, $L(0,\theta)$ is the radiance at the Earth's surface headed toward the sensor, L_{TA} is the radiance from the air column between the target and the sensor due to its mean effective temperature, and $[1 - \tau(h,\theta)]$ is the effective emissivity of the air column (note: $L_u(h,\theta) = L_{TA}[1 - \tau(h,\theta)]$). From Section 3.4, recall that for atmospheres dominated by absorption effects, the transmission can be expressed as

$$\tau(h,\theta) = e^{-C_{ext}mz} \tag{7.31}$$

where C_{ext} is the extinction cross section [m^2], m is the number density [m^{-3}], and z is the path length [m] between the Earth and the sensor.

For clear atmospheres, this can be expanded using a Taylor series and truncated to yield as a good approximation

$$\tau(h,\theta) \approx 1 - C_{ext}mz \tag{7.32}$$

Then Eq. (7.30) becomes

$$L(h,\theta) = L(0,\theta) - [L(0,\theta) - L_{TA}]C_{ext}mz \tag{7.33}$$

Expanding the Planck radiance equation about temperature and keeping only linear terms yields

$$T_i(h,\theta) \cong T(0) - [T(0) - T_A]C_{exti}mz \tag{7.34}$$

where $T_i(h,\theta)$ is the apparent temperature at the sensor in the ith bandpass, $T(0)$ is the apparent temperature at the surface, T_A is the apparent temperature of the atmosphere corresponding to L_{TA}, and C_{exti} is the absorption cross section in the ith bandpass.

Thus, apparent temperature and the extinction cross section in the bandpass are seen to be approximately linearly related for atmospheric windows where the

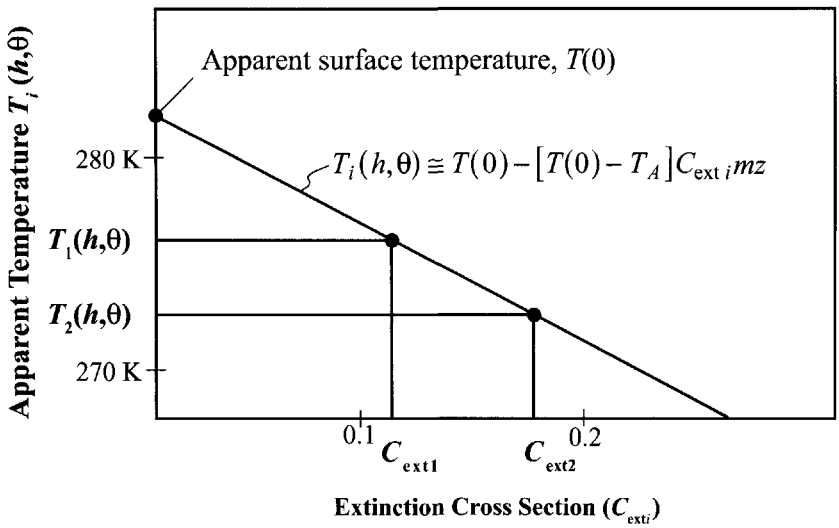


Figure 7.5 The multiple bandpass calibration technique relies on the difference in the extinction coefficient between two usually adjacent spectral bands.

apparent temperature of the Earth and the apparent temperature of the atmosphere (T_A) are each constant. For many adjacent regions (i.e., subwindows) within an atmospheric window, this is approximately true. Then if the extinction cross sections are known for two or more spectral bands, the apparent temperatures at the sensor in each band need only be plotted against the corresponding extinction cross section to yield the apparent surface temperature as the intercept (cf. Fig. 7.5). When only two spectral bands are used, the apparent surface temperature can be expressed in terms of the ratio of the extinction cross sections between the bands, which is often easier to obtain than an absolute value. The apparent surface temperature can then be expressed as:

$$T(0) = \frac{T_1(h,\theta) - RT_2(h,\theta)}{1 - R} \tag{7.35}$$

where $T_1(h,\theta)$ and $T_2(h,\theta)$ are the apparent temperatures in bands 1 and 2, respectively, and $R = C_{\text{ext1}}/C_{\text{ext2}}$. The apparent temperature can be converted to the surface-leaving radiance value using the Planck equation, and corrected for emissivity and downwelled radiance as described in Section 7.3.1 for Eq. (7.17). (Note that this analysis can also be done in terms of the extinction coefficient β_{ext}, where $\beta_{\text{ext}i} = mC_{\text{ext}i}$.)

For this approach to be effective, the two windows need to have significantly different extinction cross sections. Anding and Walker (1975) discuss methods for computation of the extinction coefficients using radiation propagation models. They go on to indicate the results of tests of this method using multiple bands in the 8–14 μm window. These results indicate errors of the order of 1 K in apparent surface temperature. They also indicate that better results can be obtained using more than two bandpasses.

The advantages of this approach are that no ground truth is required and that one does not have to image from multiple angles or at multiple altitudes. It does, however, require having multiple images of the target in different bands. The NOAA TIROS N series of satellites were specifically designed for this type of calibration using multiple bands in the 10.5–12.5 μm window [cf. Price (1984) and Walton et al. (1990)]. Because these sensors have a 1 km ground spot, it is difficult to evaluate calibration accuracy except over very large uniform temperature regions. Where this has been done for sea surface temperatures, errors of less than 1 K are reported.

Borel et al. (1999) describe an approach for atmospheric compensation that in effect combines the multiple angle and multiple band techniques. They take advantage of the unique capabilities of the Multispectral Thermal Imager (MTI) satellite to acquire multiple looks in several thermal bands to develop advanced atmospheric compensation approaches.

7.3.3 Atmospheric Propagation Models (Temperature)

The logistics of ground truth or in-scene calibration techniques often make their use impractical or impossible. As a result, atmospheric propagation models are often a useful solution when other calibration methods cannot be utilized. These models are attractive in that they can be applied to any image (i.e., there are no multiple look angle, multiple altitude, etc., requirements). In addition, they can be applied to existent imagery where little or no ground truth is available. As a result, they offer a very attractive solution that only involves the running of a propagation model. The disadvantage is that they can only effectively model the propagation through an atmosphere that is well characterized in terms of its temperature and the distribution of constituent elements. In general, they are designed to yield solutions, even if only rough estimates of atmospheric conditions are available. However, accurate results require a detailed knowledge of the atmosphere. (The adage "garbage in equals garbage out" is very applicable here.)

The LOWTRAN model (Kneizys et al., 1983, 1988), which has evolved into the higher spectral resolution MODTRAN model (Berk et al., 1989), is probably the most widely used and readily available of the propagation models, so we will use it here to exemplify this approach. Generally, these models assume that the atmosphere is divided into a number of homogeneous layers as shown in Figure 7.6. The temperature of each layer can be determined from radiosonde data acquired at the time of the data collection or from generic profiles stored in the MODTRAN model. Similarly, the concentration of the permanent gases and water vapor can be estimated from radiosonde air pressure and relative humidity data as a function of altitude. User-supplied data on meteorological visibility, season, and air mass type (continental, maritime, etc.) can be used to estimate aerosol numbers and size distributions. Once again, MODTRAN provides several standard atmospheric profiles if no detailed data are available from the time of the collection.

At any wavelength the transmission through the ith layer of the atmosphere can be approximated as

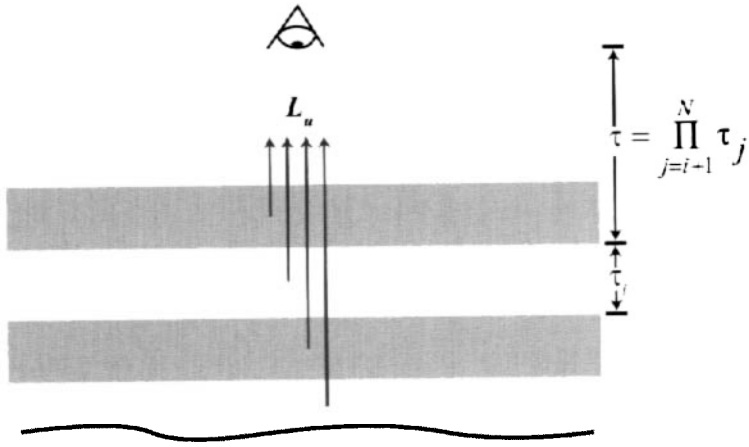


Figure 7.6 Summation of the contributions from each layer in the atmosphere to obtain the cumulative path radiance reaching the sensor.

$$\tau_{ik} = e^{-m_{ik} C_k z_i} \tag{7.36}$$

where m_{ik} is the number density of the kth atmospheric constituent (e.g., CO_2) in the ith layer, C_k is the spectral extinction cross section of that constituent, and z_i is the path length for propagation through the ith layer.

This expression is only rigorously true for discrete wavelengths and non-scattering media. In practice, MODTRAN uses an empirical approximation to Eq. (7.36) and solves for τ_{ik} based on number densities derived from atmospheric profile data (e.g., radiosonde) and database values for the spectral extinction cross sections of the atmospheric constituents. The transmission ($\tau_{i\lambda}$) through the ith layer along a path length z_i at wavelength λ is simply the product of all the τ_{ik} values, i.e.,

$$\tau_{i\lambda} = \prod_k \tau_{ik} \tag{7.37}$$

In wavelength regions where solar reflection effects are negligible (e.g., the 8–14 μm window or 3–5 μm window at night), the MODTRAN-style propagation equation can be approximated as (cf. Fig. 7.6)

$$L_\lambda = \varepsilon L_{T\lambda} \prod_{i=1}^{N} \tau_{i\lambda} + \sum_{i=1}^{N} \left[(1 - \tau_{i\lambda}) L_{Ti\lambda} \prod_{j=i+1}^{N} \tau_{j\lambda} \right] \tag{7.38a}$$

$$= \varepsilon L_{T\lambda} \tau_\lambda(h, \theta) + L_{u\lambda}(h, \theta) \tag{7.38b}$$

where $L_{Ti\lambda}$ is the blackbody equivalent spectral radiance associated with the temperature (T_i) of the ith layer.

The effective radiance reaching the sensor is then expressed as

$$L = \int R'(\lambda) L_\lambda d\lambda \tag{7.39}$$

where $R'(\lambda)$ is the normalized spectral response function of the sensor. Similarly, the effective spectral radiance can be derived from the MODTRAN estimates of spectral radiance according to

$$L_{\lambda eff} = \int L_\lambda \frac{R'(\lambda)}{\int R'(\lambda) d\lambda} d\lambda \tag{7.40}$$

Most radiation propagation models will allow the user to input a target surface temperature T and emissivity ε, to be used in generating the surface-leaving spectral radiance ($\varepsilon L_{T\lambda}$), and solve an equation similar to Eq. (7.38a) for the spectral radiance reaching the sensor. Equation (7.39) or (7.40) would then be numerically integrated to yield the effective radiance reaching the sensor. To solve for the effective transmission and path radiance reaching the sensor for use in calibrating thermal images, it is often necessary to modify the radiation propagation source code or "cheat" the code into providing the required values. In our case, we require the effective bandpass transmission $\tau(h,\theta)$, upwelled radiance $L_u(h,\theta)$, and downwelled radiance L_d. The $\tau(h,\theta)$ and $L_u(h,\theta)$ terms can be found by inserting several target temperatures T for blackbodies that cover the range of temperatures in the scene and generating the corresponding effective radiance values at the sensor $L(h,\theta)$. (MODTRAN allows the user to define the sensor location and view angle.) If one also solves for the effective bandpass radiance from these same targets on the ground, $L(0)$, through numerical solution of Eq. (7.39), then a simple linear regression of $L(0)$ with $L(h,\theta)$ can be performed according to

$$L(h,\theta) = mL(0) + b \tag{7.41}$$

where $m = \tau(h,\theta)$ is the effective bandpass transmission and $b = L_u$ is the effective upwelled radiance in the bandpass. Note that, in general, the effective transmission $\tau(h,\theta)$ will *not* equal the MODTRAN-derived average transmission. The MODTRAN propagation model can be used to compute the downwelled radiance using the hemispheric integration approach introduced in Chapter 4 [cf. Eq. (4.55)]. The model will allow a sensor on the ground to look to space, allowing a computation of the directional downwelled radiance $L_d(\theta,\phi)$ for a zenith angle θ from the zenith and an azimuthal angle ϕ. Repeated runs of MODTRAN looking at a range of angles will allow generation of sufficient data to compute the integrated downwelled radiance L_d by numerical integration of (cf. Fig. 7.7)

$$L_d = \frac{E_d}{\pi} = \int \pi^{-1} L_d(\theta,\phi) \cos\theta d\Omega \tag{7.42a}$$

Substituting $d\Omega = \sin\theta \, d\theta d\phi$ yields

$$L_d = \int_{\phi=0}^{2\pi} \int_{\theta=0}^{\frac{\pi}{2}} \pi^{-1} L_d(\theta,\phi) \cos\theta \sin\theta d\theta d\phi \tag{7.42b}$$

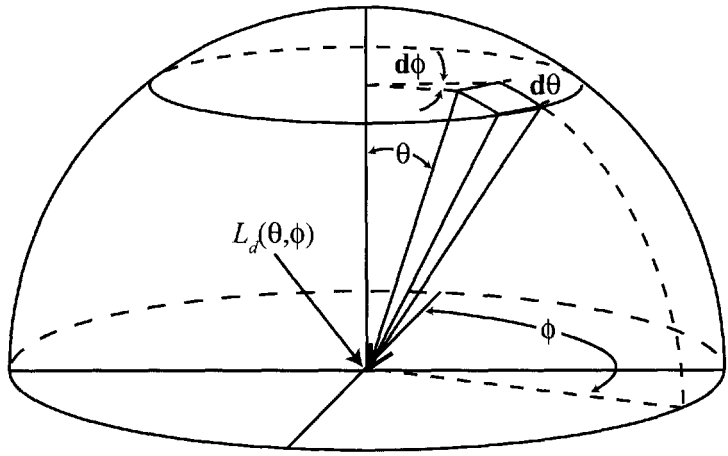


Figure 7.7 Illustration of Eq. (7.42) used for the computation of the total downwelled radiance from angular values supplied by multiple MODTRAN runs.

and for $L_d(\theta, \phi)$ independent of ϕ (i.e., no azimuthal variations in the sky),

$$L_d = 2\pi \int_{\theta=0}^{\frac{\pi}{2}} \pi^{-1} L_d(\theta,\phi)\cos\theta\sin\theta d\theta \tag{7.42c}$$

$$= 2\int_{\theta=0}^{\frac{\pi}{2}} L_d(\theta,\phi)\cos\theta\sin\theta d\theta \tag{7.42d}$$

where E_d is the irradiance from the sky [W/m^2] onto the Earth's surface, $d\Omega$ is the element of solid angle [sr], and the integral is over the hemisphere above the target at the Earth's surface. In essence, MODTRAN can be made to perform a numerical approximation to the downwelled radiance computations described in Section 4.4.2.

Thus, the standard radiation propagation models can be manipulated to yield the required inputs to a governing equation of the form

$$L(h,\theta) = [\varepsilon L_T + (1-\varepsilon)L_d]\tau(h,\theta) + L_u(h,\theta) \tag{7.43}$$

where $L(h,\theta)$ is derived from image data, and the only unknowns are ε and L_T. If the user can estimate emissivity, then L_T and, hence, T can be found.

The most serious limitation to this approach is that the accuracy of the solution is very dependent on the quality of the input data on atmospheric conditions. Thus, if a radiosonde is launched at the study site at the time of the image data acquisition, the results should be very good [Schott (1993) suggests that errors of less than 1.0 K can be expected in the 8–14 μm window]. Where this is not possible, and standard atmospheres are used or radiosonde data are interpolated across space and time, substantial errors can occur (>7 K). As a result, when concurrent

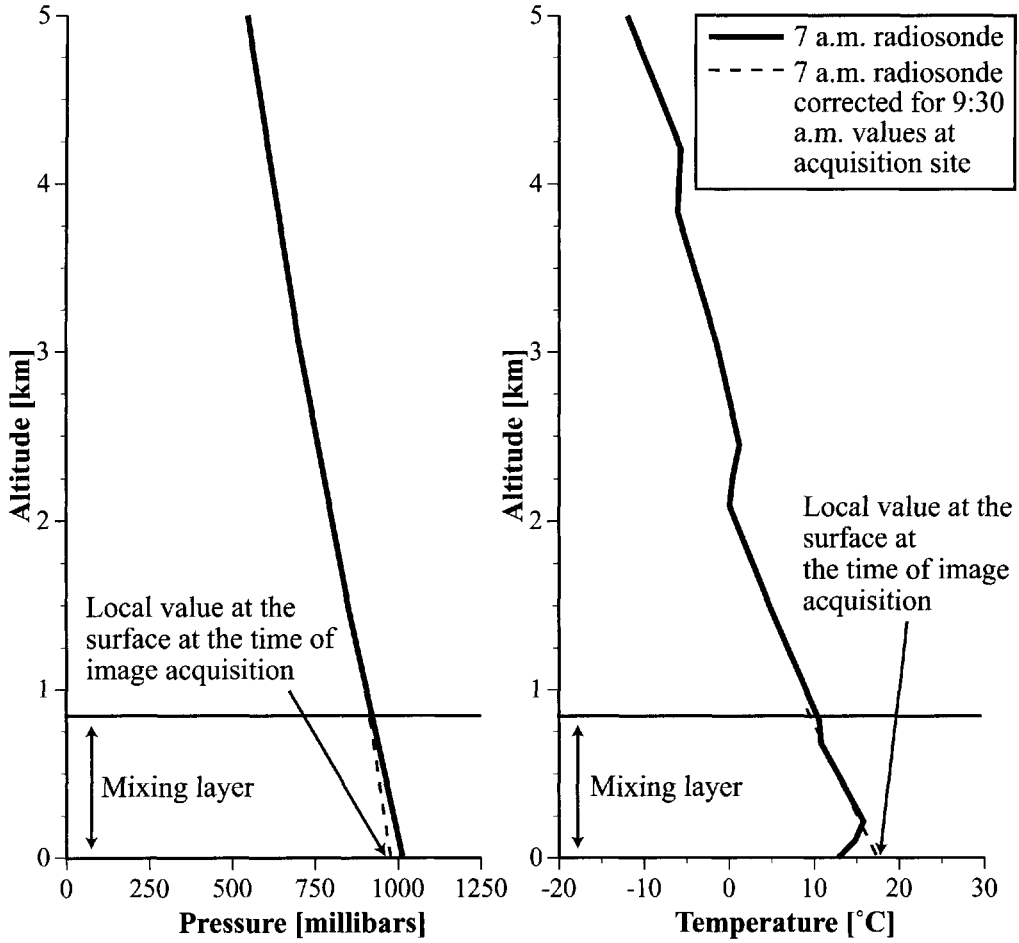


Figure 7.8 A simple method of correcting radiosonde data for local conditions. The radiosonde data were collected in Buffalo, NY, at 7:00 a.m. The data were corrected for an acquisition over Rochester, NY, (100 km away) using surface data acquired at the time of the overflight (9:30 a.m.). The correction involved a simple straight-line extrapolation from the top of the near-surface mixing layer to the surface.

and coincident radiosonde data are not available, methods to improve atmospheric interpolation or to augment radiosonde data must be employed. Byrnes and Schott (1986) describe simple techniques for adjusting the lower atmospheric profiles as a function of time using surface temperature, humidity, and visibility data (cf. Fig. 7.8).

Barsi et al. (2003) describe an approach to operationally use MODTRAN to compensate Landsat thermal data for atmospheric effects. The method employs radiosonde data that has been interpolated to 1° centers (i.e., 1° of latitude and longitude) and allows the user to provide local surface values to improve the estimates of radiosonde data using methods similar to those proposed by Byrnes and Schott (1986). A web-based interface then lets the user submit a batch job that runs MODTRAN for the relevant meteorological conditions and Landsat sensor response range and computes the appropriate atmospheric compensation param-

eters. In addition, much work is now concentrating on the use of satellite-based multispectral infrared and microwave sounders to extract temperature and humidity data as a function of altitude [cf. Hanel et al. (1992)]. While these methods are not yet as widely used as radiosonde data, they may be a valuable tool in assisting us to interpolate sparse radiosonde data over space and time. Chapter 11 discusses methods to use thermal IR imaging spectrometer data to compensate for the atmosphere using a sounder-like approach.

As pointed out earlier, the MODTRAN-style approach also has the advantage of relative ease of use, so it can supplement other atmospheric calibration techniques by generating any of the variables they do not. For example, the multiple-bandpass techniques do not generate a value of L_d, and this can be obtained from MODTRAN.

7.3.4 Emissivity

All of the approaches presented above have a common problem in that they cannot separate temperature from emissivity effects. In order to separate out what component of the surface-leaving radiance is due to temperature and what is due to emissivity, we need to estimate the emissivity. This can be quite difficult because only limited data are available on the emissivity of surfaces, particularly irregular surfaces as they are commonly found in imagery (e.g., road surfaces, grass, gravel, etc.). Where data are available [cf. Gubareff et al. (1960)], they are often not spectral; they do not cover the correct spectral bandpass, or only total hemispheric data are available. Becker et al. (1985) describe an instrument for measuring the spectral and directional emissivity of surfaces, and Schott (1986) describes a pair of instruments that can be used in the field to measure both normal hemispheric and angular hemispheric emissivity values. More recently, Salisbury and D'Arian (1992) describe an instrument and measurements of the spectral emissivity of a variety of natural surfaces. Many of these spectral emissivity values have been collected in databases maintained by the NASA Earth remote sensing community and available over the Internet. Knowledge of emissivity is very important, since an error in emissivity of one one-hundredth referred to as one unit (in the 8–14 μm window) can commonly result in temperature errors of approximately 1 K. The importance of emissivity and emissivity error is downplayed in most of the quantitative results cited in the literature. This is because water is commonly used as the test target (particularly for satellite studies). The emissivity of water for near-normal viewing is well known and very high. This minimizes any error in temperature that would result due to emissivity effects and underestimates errors that will occur for objects with lower or less well-known emissivities. Lowe (1978) demonstrated an interesting approach for dealing with the emissivity problem. Airborne images were acquired with two sensors. An active system imaged the ground where it was illuminated with a CO_2 laser operating at 10.6 μm, and passive thermal images were collected in the 10–12.6 μm region (cf. Fig. 7.9). The brightness in the active images was assumed to vary primarily as a function of emissivity (reflectivity), thus producing a linear emissivity map that could be calibrated from two or more

known targets. These emissivities could then be used with the passive images to solve for temperature under the assumption that the objects in the scene behaved as gray bodies over the 10–12.6 μm region. Lowe (1978) points out that this may be a poor assumption for the silicates due to strong Restrahlen effects from quartz in this spectral band.

7.3.5 Summary of Thermal Atmospheric Compensation

In summary, numerous approaches have been identified for measurement of surface temperature. Most of them have been shown under certain circumstances to yield accurate temperature measurements to better than 1 K. However, very few have undergone rigorous tests of robustness under operational conditions. Furthermore, some of the most accurate and best proved, such as the multiple-altitude or split-band approaches, require specific platform or sensor characteristics that can be prohibitive. We are thus often left with a situation where adequate compensation approaches still do not exist, or where the possible approaches have not been tested to the point where the expected results are known. In short, the old adage of *caveat emptor* applies, and the user should become quite familiar with a compensation technique in known situations before trusting the results.

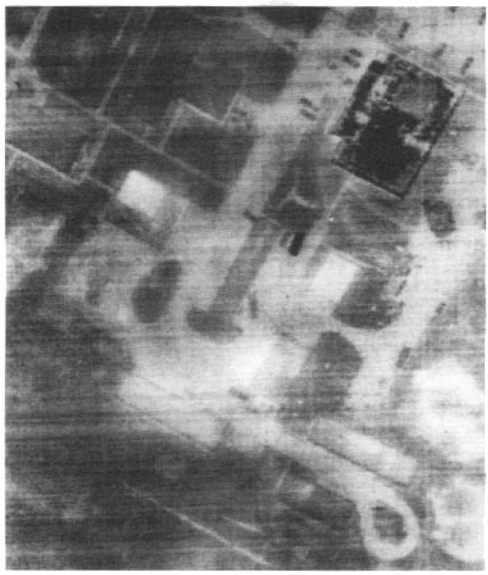
The atmospheric compensation techniques discussed in this section are generally not applicable in regions of the EM spectrum where reflective solar flux dominates the observed radiance. There are, however, a parallel set of compensation methods applicable in the solar reflective region, as we will see in the next section.

7.4. APPROACHES TO MEASUREMENT OF REFLECTIVITY

In the VIS-NIR and SWIR, the scientific questions have revolved around how to measure reflectance or, barring that, how to measure the relative reflectance between spectral bands. In this section, we will parallel the treatment used for calibration of images for temperature measurement by looking at ground truth, in-scene, and atmospheric propagation models for calibration of image data to reflectance. The underlying premise is that the type or condition (e.g., vegetation health, water quality, soil moisture) of the targets of interest is a function of reflectance. Therefore, if the image data can be calibrated to reflectance, they can be directly compared to laboratory and field measurements of reflectance to yield improved classification and condition assessment.

7.4.1 Ground Truth Methods (Reflectance), a.k.a. Empirical Line Method (ELM)

The calibration of aerial and satellite systems using ground truth has generally relied on the use either of control panels or ad hoc control surfaces of known reflectance. The control panel approach is attractive because the range of reflectance values can be controlled by the fabrication process, and care can be taken to ensure



(a) A passive thermal infrared image in the 10–12.6 μm band.

(b) A reflective image obtained by illuminating the same scene as in (a) with an active CO_2 laser at 10.6 μm.

Figure 7.9 A pair of images used to produce a temperature map.

that the samples are approximately Lambertian to minimize any errors that could be introduced by sensor view angle effects. The use of panels is the most attractive calibration method if quality ground truth panels three or more times the ground instantaneous field of view of the sensor can be imaged at the time of data collection (cf. Fig. 7.10). The calibration is then a simple matter of regressing observed radiance values against known reflectance values in each band according to

$$L = (E_s \pi^{-1} + L_d)\tau_2 r_d + L_u \tag{7.44a}$$

$$L = m r_d + b \tag{7.44b}$$

where $E_s \cong E_s' \cos \sigma' \tau_1$, $\tau_2 \cong e^{-\delta' \sec\theta}$, $m = E_s \pi^{-1} \tau_2 + L_d \tau_2$ is the slope of the regression, and $b = L_u$ is the intercept. We will assume for most of this section that calibration sites are chosen such that the shape factor (F) is unity. In addition, we assume that τ_1, τ_2, and L_u are constant over the scene. When this is not the case (e.g., sensors with a wide field of view or a large area of coverage), corrections must be made for changes in the atmosphere over the scene. In many cases, this involves recalibrating the scene at several locations.

The reflectance of any Lambertian object or the apparent Lambertian reflectance of any object in the scene can then be computed by rearranging Eq. (7.44b). In fact, if the sensor's radiance is a linear function of digital count, the sensor need not be radiometrically calibrated to use this approach. The panel digital count values can be regressed with the reflectance values according to

$$\mathrm{DC} = gL + b_o \tag{7.45a}$$

$$= g m r_d + g b + b_o \tag{7.45b}$$

$$= m' r_d + b' \tag{7.45c}$$

where g and b_o are the unknown sensor gain and bias for the bandpass, and m' and b' are the slope and intercept of the regression of image DC values with known panel reflectance values. This method of using a linear regression in each band to relate digital counts or radiance to reflectance has become known as the *empirical line method* (*ELM*).

In cases where logistics preclude the use of ground truth panels, the same effect can be achieved by measuring the reflectivity of several Lambertian objects in the scene with a range of reflectance values and again regressing against the observed radiance or digital count. Care must be taken in the selection and measurement of these ad hoc control surfaces to ensure that they are sufficiently diffuse over all relevant angles, and that they are uniform enough that the field sample will be characteristic of the image sample. This can be particularly difficult with satellite images where the *ground sample distance* (*GSD*) may be tens or hundreds of meters. Figure 7.11 shows an image of the "reflectomobile" used by Slater et al. (1987) in characterizing the reflectance of large surfaces for satellite sensor calibration as described in Section 6.4.

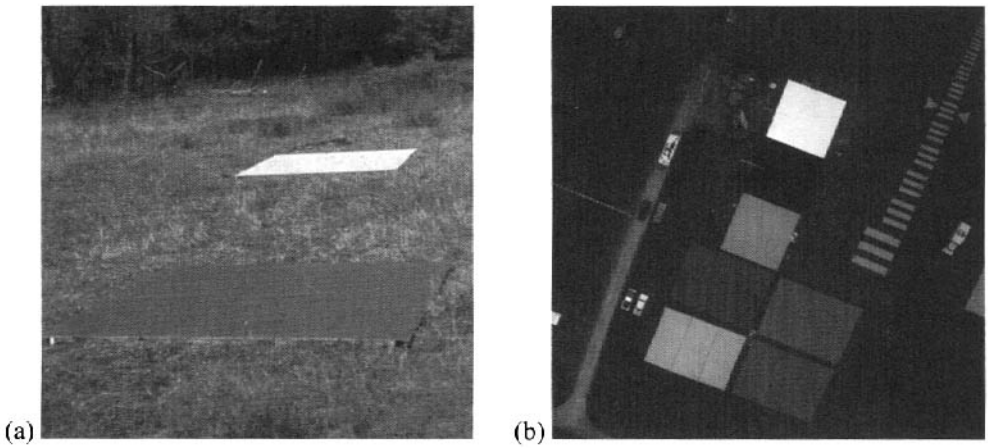


Figure 7.10 Ground control panels used for atmospheric calibration. (a) Control panels. (b) Images of control panels painted on a runway apron.

It is important to note the assumptions implicit in the use of these standard ground truth or ELM methods. First, the calibration targets are assumed flat and level, with no neighboring obscuration; second, they are assumed homogeneous; and third, they are Lambertian. If these conditions are met and the calibration target reflectance values are well known, then the solution that results should be very accurate. However, they will only be completely valid for targets with similar characteristics. Sloped surfaces and non-Lambertian surfaces will have errors that will be a function of how significantly they deviate from the calibration assumptions.

The slope effects can easily be corrected for Lambertian surfaces if the relative amount of skylight to total illumination (l) is known. In this case, it can be shown that the radiance from any angled surface is

$$L_i = \left[(m - lm)\cos\sigma_i' / \cos\sigma' + lm\right] r_d + b \tag{7.46}$$

where m and b are the slope and intercept from Eq. (7.44b), l is the ratio of skylight to total illumination in the bandpass $L_d/(E_s\pi^{-1}+L_d)$, and σ'_i is the angle from the normal of the ith target to the sun (recall that σ' is the angle from the normal of the calibration targets to the sun). Rearranging Eq. (7.46) yields the diffuse reflectance for a target at any orientation. The skylight to total illumination ratio in the bandpass can be determined with a radiometer located in the scene that records the irradiance onto a horizontal surface. This yields the total irradiance. By simply casting a shadow on the irradiance sensor with a disk whose shadow just blocks the sun, the skylight irradiance can be determined and the ratio computed (cf. Fig. 7.12). Note that the instrument used need only be linear and have no dark-level bias. No careful calibration is required. Since the skylight-to-total-illumination ratio (l) can vary significantly with time and sky conditions, care must be taken to make these measurements coincident with the image acquisition and in the spectral

Figure 7.11 The reflectomobile used by Slater et al. (1987) in calibration of satellite sensors. The radiometers were calibrated in the field with reflectance standards and then driven over a study site to generate mean reflectance over a large, approximately uniform surface.

band (s) of interest. The value of l can also be determined using in-scene methods as described in Section 7.4.2.

For non-Lambertian surfaces, solving for reflectance values requires assumptions regarding the shape of the bidirectional reflectance function. That treatment is beyond our scope, so we will assume that we will be reporting the apparent diffuse reflectance of all samples (i.e., reflectance factors as described in Sec. 4.2.1).

7.4.2 In-Scene Methods (Reflectance)

In many cases, the logistics associated with acquiring good ground truth are impractical or impossible to meet. As a result, many methods of in-scene compensation have been developed.

One of the most interesting approaches takes advantage of the difference in radiance levels observed at shadow edges. Piech and Walker (1974) have shown how this radiance difference can be used to compute the upwelled radiance term (L_u) and the skylight to total illumination factor (l). The radiance in the bandpass observed just outside a shadow edge (L_s) and just inside a shadow (L_{sh}) cast on the same diffuse material can be expressed as

$$L_s = \left[E_s \pi^{-1} \tau_2 + F L_d \tau_2 \right] r_d + L_u \tag{7.47}$$

$$L_{sh} = F L_d \tau_2 r_d + L_u \tag{7.48}$$

where F is the fraction of the hemisphere above the target area that is sky [i.e., (1–F) is the fraction obscured by neighboring objects as shown in Fig. 7.13]. Combining Eqs. (7.47) and (7.48) by substitution of r_d and rearranging yields

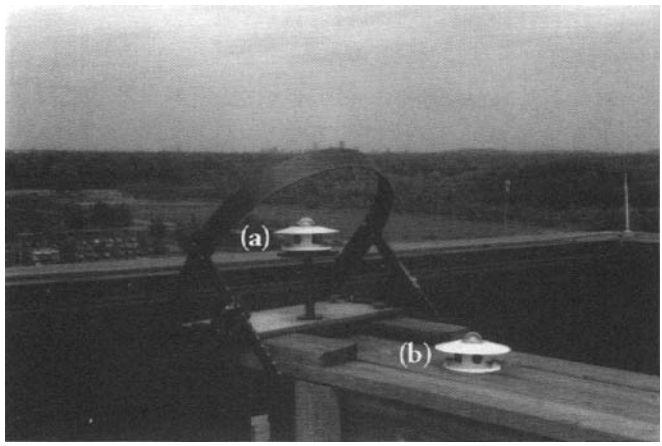


Figure 7.12 Apply total irradiance pyroheliometer (a) and skylight pyroheliometer with sun band (b). These instruments measure the total (a) and skylight (b) irradiance onto a horizontal surface. The devices shown are for broad-band measurements. To obtain bandpass values, similar instruments with appropriate filters are used.

$$L_s = \frac{E_s \pi^{-1} + FL_d}{FL_d} L_{sh} - \frac{E_s \pi^{-1} + FL_d}{FL_d} L_u + L_u \tag{7.49a}$$

$$L_s = mL_{sh} + b \tag{7.49b}$$

where $m = (E_s \pi^{-1} + FL_d)/FL_d$ and $b = (1-m)L_u$.

For narrow fields of view, L_u can be assumed constant, and if objects of similar shape are selected, then F is also approximately constant. Therefore, m and b in Eq. (7.49b) are approximately constant and can be solved for by a linear regression of the radiance just in the shadow (L_{sh}) versus the radiance just in the sun (L_s) for several targets of varying reflectance. L_u can then be found from the intercept (b) and slope term (m) to be

$$L_u = \frac{b}{1-m} \tag{7.50}$$

Rearranging the intercept term yields an expression for the skylight to total illumination (l):

$$l = \frac{1}{1+F(m-1)} \tag{7.51}$$

where F can be computed from geometry for similarly shaped shadow-casting objects used in the regression (e.g., buildings).

Piech et al. (1978) suggest that, given the upwelled radiance (L_u) from Eq. (7.50), the total incident radiance term can be solved for using a statistical estimate of the mean observed radiance for a class of objects whose mean reflectance in the

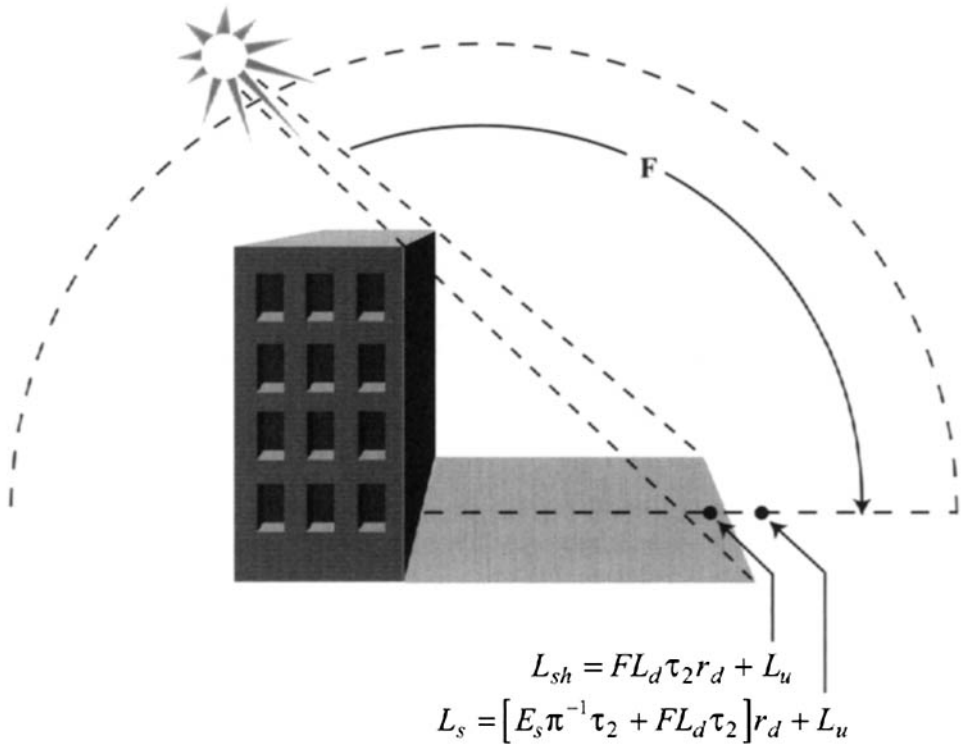


Figure 7.13 The radiance from the same surface just beyond a shadow edge L_s and just inside the shadow L_{sh}. The shape factor F is the fraction of the hemisphere above the shadow edge illuminated by the skylight.

bandpass can be well estimated (cf. Fig. 7.14). In their studies, they commonly chose concrete. The necessary equations take on the form

$$m = \left[E_s\pi^{-1}\tau_2 + L_d\tau_2\right] = \frac{L_{avg} - L_u}{r_{davg}} \tag{7.52}$$

where L_{avg} is the mean radiance observed for many samples of the standard material (concrete) sampled in the scene, and r_{davg} is the mean reflectance based on numerous laboratory or field measurements of the standard material (concrete).

Piech (1980) found that, in applying this "scene color standard" approach to color aerial photography, reflectance errors were typically less than 1 unit below 10 reflectance units, and less than approximately 10% of the reflectance value above 10 reflectance units. (I.e., The additive error due to L_u dominated below about 10%, and the multiplicative error associated with the slope term dominated above about 10%.) Regrettably, this method cannot be used with imagery where the resolution is insufficient to show sharp shadow edges (i.e., most satellite data). As a result, alternative approaches to solve for the upwelled radiance are required.

One of the simplest methods called the *dark object subtraction* or *histogram minimum method* (*HMM*) is to set the minimum scene radiance to be the upwelled

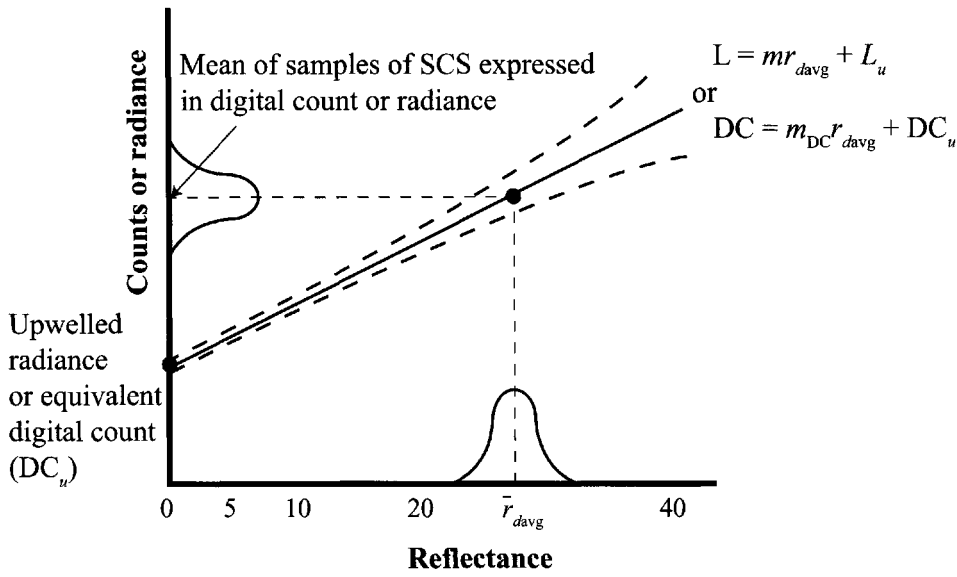


Figure 7.14 Illustration of how the radiance or counts to reflectance relationship is generated using a known upwelled radiance value and the scene color standard (SCS) concept. The dashed lines on either side of the two-points fit show the error bounds as estimated by Piech (1980).

radiance on the assumption that it represents the radiance from a scene element with near zero reflectivity [cf. Chavez (1975) and Switzer et al. (1981)], i.e.,

$$L = (E_s \pi^{-1} + L_d)\tau_2 r_d + L_u \cong L_u \text{ for } r_d \approx 0 \tag{7.53}$$

This method works reasonably well in spectral bands where near-zero reflectors are available in close proximity to the study region, i.e., within a region where the atmosphere may be assumed constant, and when applied to brighter objects where any errors are relatively small. For example, in the NIR, clear water has very low reflectance, and most studies are aimed at vegetation whose reflectance is quite large. On the other hand, this method often breaks down in the visible region where minimum scene reflectances may still be a few percent, which is also the reflectance of water and vegetation in this region. This limitation can be overcome if one considers multispectral data where the upwelled radiance (L_u) can be found with confidence in one spectral band (e.g., the NIR using clear water). Then an estimate for L_u in other bands can be found if we assume that the reflectance in the two bands are approximately correlated with zero bias such that

$$r_2 = Cr_1 + e \tag{7.54}$$

where r_1 and r_2 are the reflectance values in band 1 and band 2, C is approximately a constant, and e is the error due to the lack of perfect correlation between r_1 and r_2. If we then write the expression for the radiance observed from level objects in spectral bands 1 and 2 as

$$L_1 = (E_{s1}\pi^{-1} + L_{d1})\tau_{21} r_{d1} + L_{u1} = m_1 r_{d1} + L_{u1} \tag{7.55a}$$

$$L_2 = (E_{s2}\pi^{-1} + L_{d2})\tau_{22} r_{d2} + L_{u2} = m_2 r_{d2} + L_{u2} \tag{7.55b}$$

and combining using Eq. (7.54), we have

$$L_2 \cong \frac{m_2 C}{m_1}(L_1 - L_{u1}) + L_{u2} + e \tag{7.56}$$

If we let band 1 be the NIR band where the upwelled radiance is known using the dark-object method, then for many objects [preferably level and of materials where we believe Eq. (7.54) is valid (i.e., we might use pixels from a single material class)], we can regress the radiance in band 2 (e.g., any visible band) against the radiance in band 1 after dark-object subtraction (i.e., L_1 - L_{u1}). The regression minimizes the error term (e), such that the intercept is approximately the upwelled radiance in band 2 (L_{u2}). This process can be repeated for each band of interest.

To complete the full calibration, one must still solve for m_1 and m_2 using a method such as the scene color standard approach described above. However, many investigations only require the variation in the relative spectral reflectance ratio. This value can be expressed as

$$\frac{L_1 - L_{u1}}{L_2 - L_{u2}} = \frac{m_1 r_{d1}}{m_2 r_{d2}} \propto \frac{r_{d1}}{r_{d2}} \tag{7.57}$$

where m_1/m_2 can be assumed constant (e.g., in level terrain). Although not rigorously valid in sloped terrain, Eq. (7.57) is approximately valid when the skylight term is small compared to the sunlight term (i.e., the cos σ' effects—which vary from pixel to pixel—approximately cancel), i.e.,

$$\frac{L_1 - L_{u1}}{L_2 - L_{u2}} = \frac{E'_{s1}\pi^{-1}\tau_{11}\tau_{21}\cos\sigma' + \tau_{21}L_{d1}}{E'_{s2}\pi^{-1}\tau_{12}\tau_{22}\cos\sigma' + \tau_{22}L_{d2}} \frac{r_{d1}}{r_{d2}} \propto \frac{r_{d1}}{r_{d2}} \tag{7.58}$$

Equation (7.58) has been used in geologic studies to minimize pixel-to-pixel terrain effects and increase variations due to material changes as characterized by reflectance ratios. Crippen (1988) discusses the importance of correcting for the additive (path radiance) effects when using band ratios in regions with topographic relief. This is illustrated in Figure 7.15, in which the simple band ratio image is dominated by topographic effects, though reversed in contrast from the original images. After subtracting for additive effects, the topographic effects are greatly reduced, and a mine and tailings show up as distinctly bright (high ratio) objects.

Another approach introduced by Crippen (1986) to solving for the upwelled radiance that does not require a band where we can assume a zero reflectance object exists also takes advantage of the within-class correlation described by Eq. (7.54). In this case, the radiance in band 2 is related to the radiance in band 1 for a class of objects whose spectral reflectance ratio can be approximated as $C_A = r_{d2}/r_{d1}$. The radiance relationship can be expressed as

$$L_2 = \frac{m_2 C_A}{m_1} L_1 - \frac{m_2 C_A}{m_1} L_{u1} + L_{u2} + e \tag{7.59}$$

Note that this is the same as Eq. (7.56) except we have added a subscript to the reflectance ratio (C_A) to indicate that all the pixels used to solve Eq. (7.59) come from a class designated A. If we could identify a second class of objects with a different reflectance ratio C_B in the same two bands (see K means approach in Section 9.2.3 for a method of class selection), we could write a corresponding equation:

$$L_2 = \frac{m_2 C_B L_1}{m_1} - \frac{m_2 C_B}{m_1} L_{u1} + L_{u2} + e \tag{7.60}$$

By plotting the radiance in band 2 versus the radiance in band 1 and regressing by class, we can estimate the form of the straight lines described by Eqs. (7.59) and (7.60) (cf. Fig. 7.16). A little algebra shows that the intersection of these lines should occur when $L_1 = L_{u1}$ and $L_2 = L_{u2}$, yielding estimates for the upwelled radiance. By including additional classes in the process, more estimates of upwelled

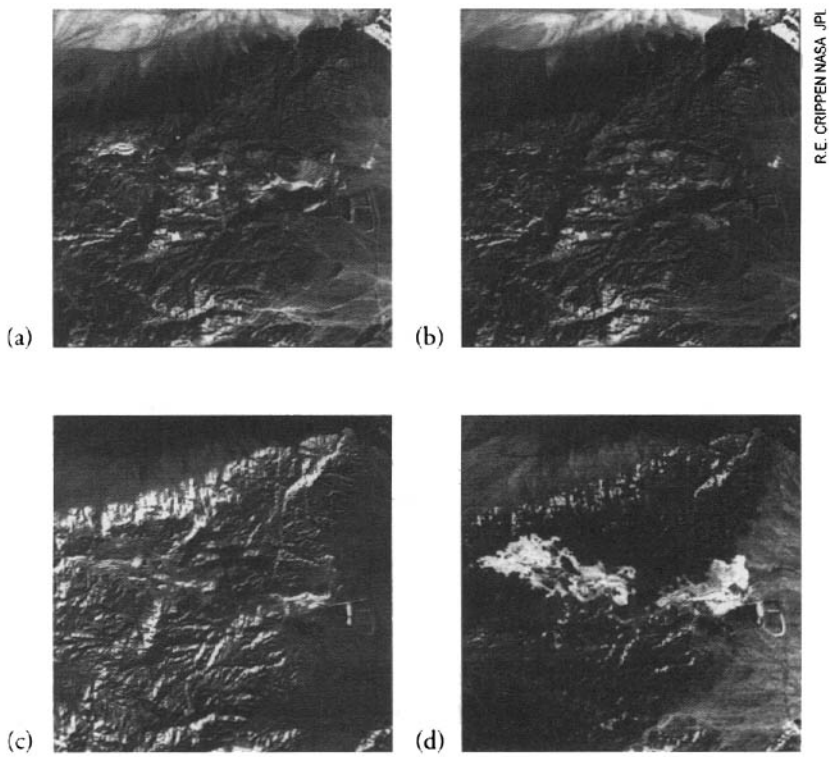


Figure 7.15 Images showing how illumination effects can be reduced in ratio images if additive effects due to the sensor bias and path radiance are removed. (a) Landsat TM band 1 subscene of the Eagle Mountains Iron Mine, (b) band 4, (c) band 1/band 4 (note reversal of topographic effects), (d) band 1/band 4 after subtraction for additive effects (note how the topographic effects are reduced and the high-ratio signature of the mine and tailings become apparent).

radiance can be obtained. Then a decision about the best estimate can be based on several criteria [e.g., L_u must be less than L_u from the histogram minimum, keep only the best regressions (high r^2 values), and keep classes with greatest difference in C_i values]. For images with several bands, this approach can be used in a similar fashion to generate multiple estimates of L_u in a band, which can then be averaged or a quality metric can be used to identify the best estimate.

Note that most of these in-scene methods do not require absolute sensor calibration if the sensor output (e.g., digital count) is a simple linear function of incident radiance. In these cases, the relationships of Eq. (7.45) can be used. Instead of using radiance values, digital count values can be used directly in the in-scene methods presented here. For example, Eq. (7.56) reduces to

$$DC_2 = \frac{g_2 m_2 C}{g_1 m_1}(DC_1 - DC_{u1}) + DC_{u2} + e \tag{7.61}$$

where $DC_{ui} = g_i L_{ui} + b_{oi}$ is the digital count in the ith band that would be produced by the upwelled radiance in the ith band (L_{ui}), and the magnitude of the error term is modified by the sensor gain. DC_{u1} can be approximated in the NIR band using the minimum digital count or dark-object method. The regression of DC_2 values versus (DC_1 - DC_{u1}) values produces DC_{u2} as the intercept. Substitution of digital

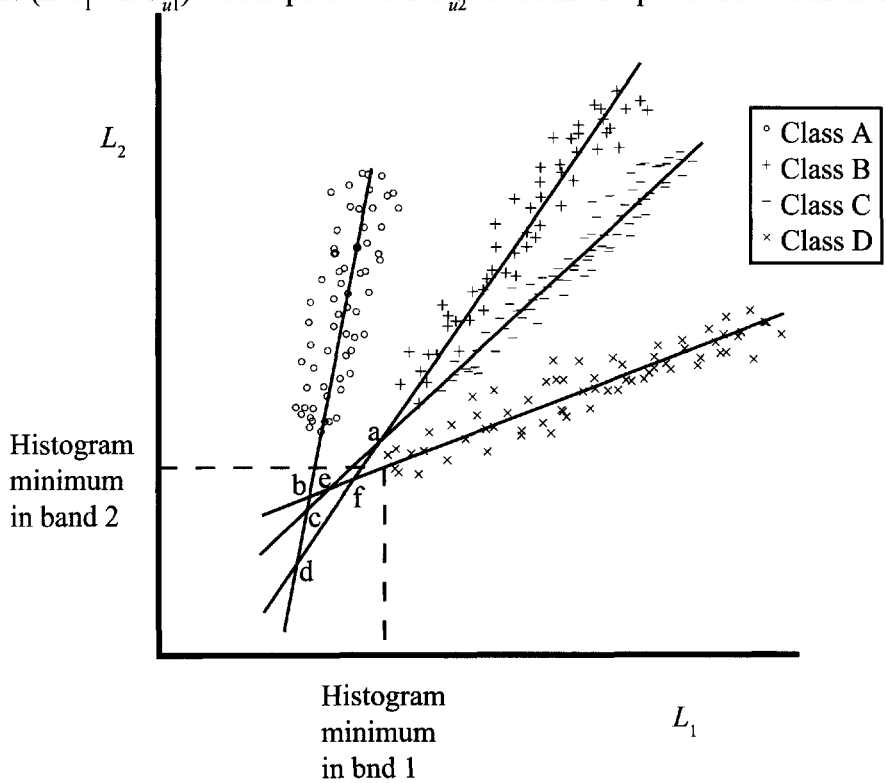


Figure 7.16 Illustration of the use of the regression intersection method to find upwelled radiance as well as several of the decision criteria introduced in the text to decide on the "best" solution(s).

count values in the left hand side of Eq. (7.58) results in a solution for the right hand side that is similar to Eq. (7.58):

$$\frac{DC_1 - DC_{u1}}{DC_2 - DC_{u2}} = \frac{g_1}{g_2} \frac{(E'_{s1}\pi^{-1}\tau_{11}\tau_{21}\cos\sigma' + \tau_{21}L_{d1})}{(E'_{s2}\pi^{-1}\tau_{12}\tau_{22}\cos\sigma' + \tau_{22}L_{d2})} \frac{r_{d1}}{r_{d2}} \propto \frac{r_{d1}}{r_{d2}} \tag{7.62}$$

A similar analysis can often be performed using digital count values with the sensor calibration incorporated into the atmospheric calibration. In many cases, combining the calibrations can reduce the overall errors in the calibration process by simultaneously minimizing errors.

With all of these methods, residual errors in low-reflectance targets are still of concern both when full calibration to reflectance is required and when using reflectance ratios. The need for a reflectance standard in the scene is also often a problem in regions without manmade features or where the resolution precludes isolating unique materials (e.g., pure pixels of concrete). As a result, atmospheric propagation models must sometimes be considered.

7.4.3 Atmospheric Propagation Models (Reflectance)

In this section, we want to address how a radiative transfer (RT) code (e.g., the MODTRAN code introduced in Sec. 7.3.3) can be used as a means to estimate the key terms in the reflective radiance equation and thus allow us to invert the equation and solve for the reflectance of the target. It turns out that these RT codes are widely used to support a number of research and development studies, as well as in support of operational programs, so it is worth our while to delve at least a little way into their general characteristics. To this end, we need to first take a closer look at ways to characterize the atmosphere (Sec. 7.4.3.1) before we address some of the RT mechanisms that are estimated by the computational codes. Our discussion is based on consideration of several RT codes [cf. Vermote et al. (1997), and Herman et al (1995)]; however, we have tended to draw most directly from the approach used in the LOWTRAN, MODTRAN, FASCODE family of codes developed by the U.S. Air Force research labs [cf. Kneizys et al. (1983), Berk et al. (1989) and Smith et al. (1978)].

7.4.3.1 Characterization of the Atmosphere in a Radiative Transfer Context

This treatment draws on the issues we raised in Chapters 3 and 4 to define what parameters are needed to characterize the atmosphere as it impacts radiation propagation. As a reminder, we determined that absorption, scattering, and self-emission by gases required knowledge of their number density and temperature profiles. In addition, absorption, scattering, and self-emission by aerosols require us to know their origin, temperature, number density, and particle size distribution at any relevant altitude. Generally speaking, the radiative transfer codes treat the atmosphere as being made up of homogeneous layers wrapped around the globe or stacked above a flat world (plain parallel layers), depending on the level of fidelity required. For our purposes, we can describe the atmosphere as indicated

in Figure 7.17, with most of our interest focused on the lower (denser, warmer, more variable) layers below 25 km. The concentration or number density of the well-mixed gases is then just a function of the predetermined mixing ratio and the temperature and pressure of each layer. In some models, optically important gases (e.g., ozone) may be introduced in only a small number of layers (stratosphere for ozone), and the concentration may be adjusted seasonally and/or geospatially (or controlled by the user). In addition to the dry or well-mixed gases, water vapor varies substantially with elevation and horizontally around the globe. The World Meteorological Organization characterizes the atmosphere using radiosondes that are launched at 0 and 12:00 Greenwich Mean Time from stations all over the globe. These radiosondes are balloon-borne instrument packages that provide the temperature, pressure, and relative humidity, as a function of altitude, to altitudes of the order of 20 km (10 mb). From the relative humidity (note that dew point is

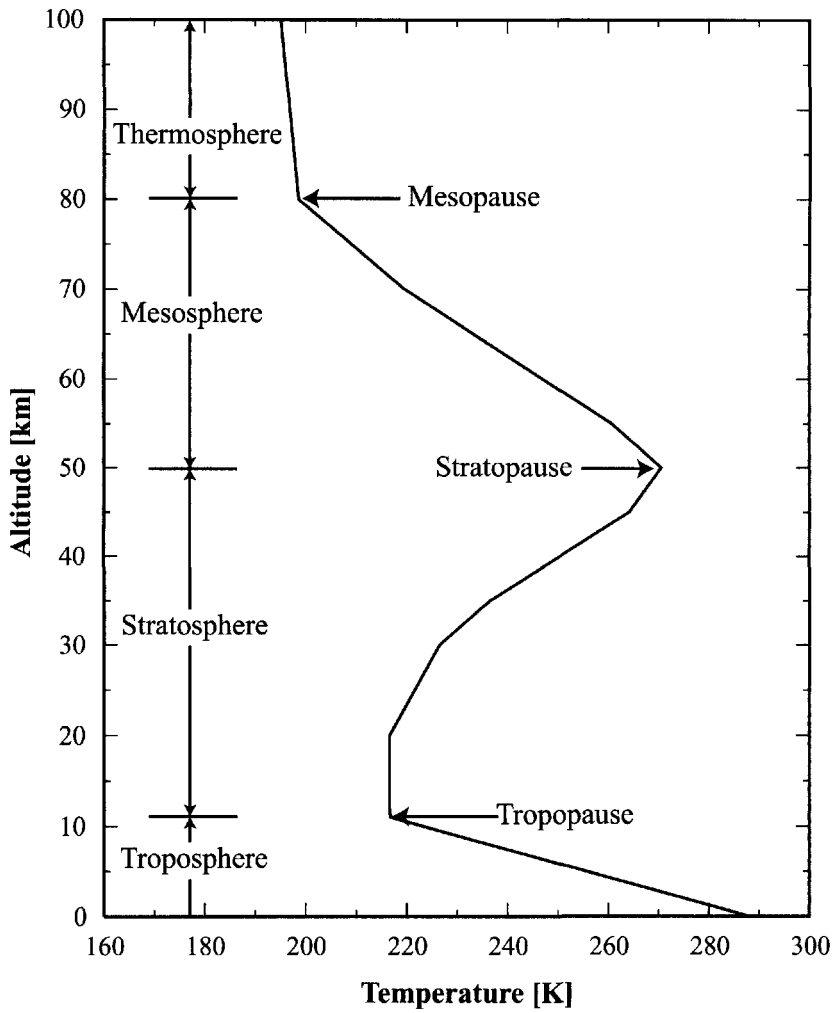


Figure 7.17 Description of the atmosphere for use in the radiative transfer calculations. [Adapted from Fenn et al. (1985).]

often plotted to fit conveniently on graphs with air temperature), temperature, and pressure, the water vapor number density can be computed for each layer using the gas laws. Sometimes we will be interested in the total amount of water along a path. This is often referred to as the column water vapor amount. It is obtained by computing the depth the water would have if all the water in an air column along that path were precipitated out. It can be computed from the temperature, pressure, relative humidity, and path length through each layer along the path. In many cases, standard atmospheres are used as reference or baseline cases. These standard atmospheres define the temperature, pressure, and relative humidity from Earth to space for each layer in addition to defining the relevant gases and the mixing ratios at each layer. Six standard atmospheres defined in Fenn et al. (1985) are commonly used in the literature and are available in MODTRAN (1976 U.S. standard, mid-latitude summer, mid-latitude winter, subarctic winter, subarctic summer, and tropical). The radiosonde data can then be taken as a means to adjust the most appropriate standard atmosphere for the specific temperature, pressure, and relative humidity available to the point where the radiosonde stopped sampling. Figure 7.18 shows some example characteristics of standard atmospheres and the use of specific radiosonde data to adjust the lower layers. Figure 7.19 shows the radiosonde sample stations for the United States. and environs indicating that while there is reasonable coverage, the likelihood of having local radiosonde data (i.e., less than 30 miles or less than one hour of time difference) is essentially zero. As a result, most users must interpolate across space and time and/or use local surface data to adjust the available data (see the discussion in Sec. 7.3.3).

Having characterized the gases and the atmospheric temperature using standard atmospheres and local radiosonde data, we now turn to aerosols. In general, we will be concerned with three types of aerosols. Those in the mixing layer (1-2 km), those widely distributed above the mixing layer referred to as tropospheric or background aerosols, and those due to volcanic events, which can distribute aerosols relatively high (15-20 km) in the atmosphere in bands that are spread around the globe by upper atmospheric circulation patterns. There are seldom (never) sufficient resources to characterize local aerosols. However, aerosols can often be characterized into a relatively small number of classes without introducing too much error. Shettle and Fenn (1979) describe several classes of aerosols, including rural, urban, maritime, tropospheric, and fog. The characterization is based on the inputs needed to run a Mie scattering code to generate the angular scattering coefficient. Each aerosol class is characterized by one or two log-normal particle size distributions (the histogram of the number of particles in each size bin), the relative number of particles in each distribution, and the optical properties (complex index of refraction) of the particles in each distribution. The two size distributions are for particles less than a micron and particles roughly a micron in size (cf. Fig. 7.20). Note that Shettle and Fenn (1979) point out that the particle size distributions are also a function of humidity, as illustrated in Figure 7.20. Choosing the class of aerosols defines the scattering phase function (derived from a Mie scattering code) and the relative spectral shape of the absorption and scattering, but not the absolute magnitude, which is defined by the total number of particles. In the MODTRAN code, this is commonly done by using the horizontal visibility parameter. This tells

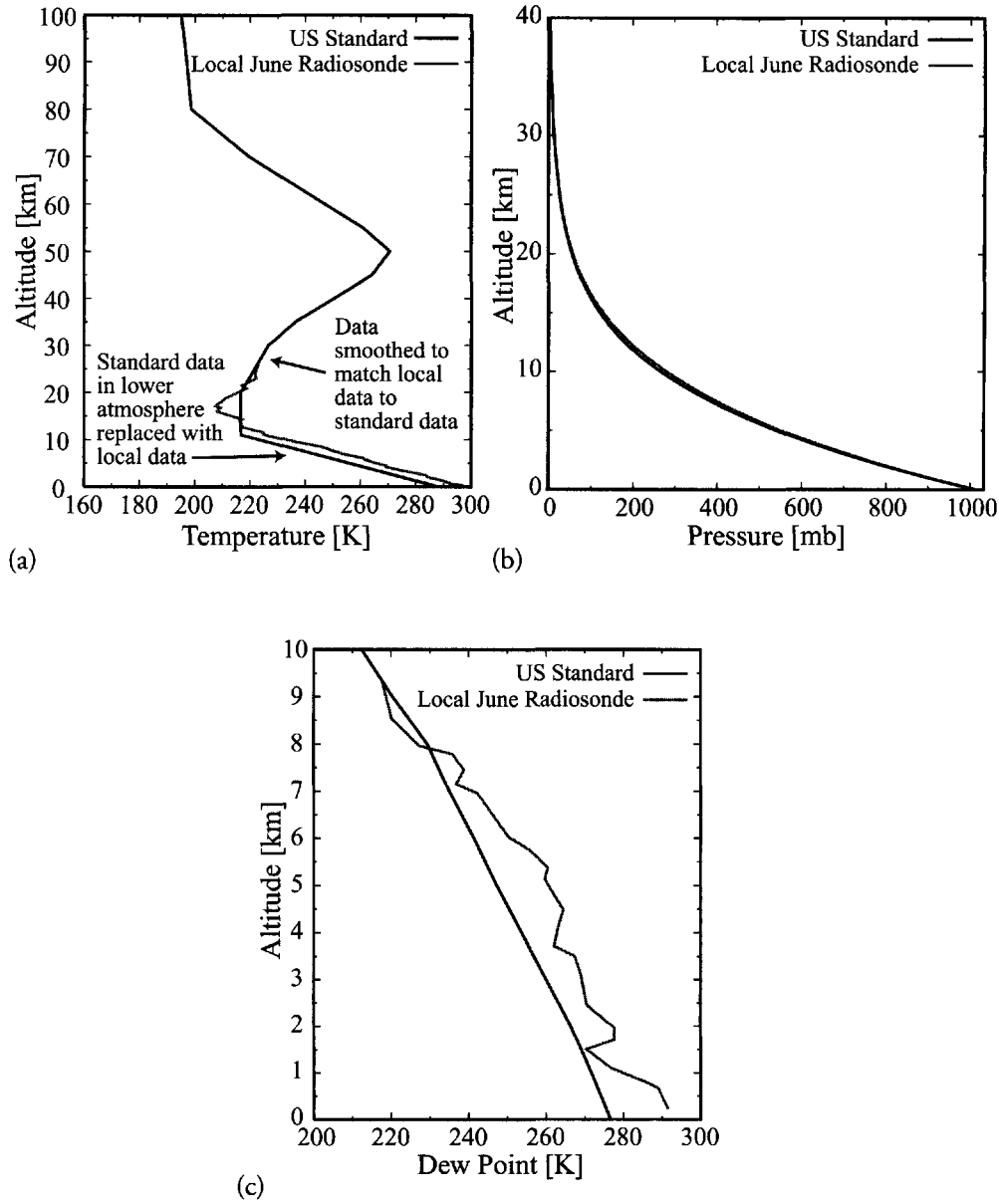


Figure 7.18 Example of standard atmosphere and use of radiosonde data to adjust tropospheric layers [Adapted from Fenn et al. (1985).]

the code the horizontal optical depth at 550 nm. This in turn fixes the magnitude of the extinction coefficient (conceptually through the number density) in the mixing layer. The number density of mixing layer aerosols, of whatever type specified, decays to the tropospheric background level above the mixing layer, where the tropospheric aerosols continue to decay with elevation. Note that in MODTRAN the tropospheric aerosols have two user selectable distributions, spring-summer and fall-winter, reflecting the greater numbers of background aerosols in the spring and summer. The volcanic aerosol models used in MODTRAN are described by

Figure 7.19 Map of radiosonde stations. Balloons are released twice daily at 12:00 and 24:00 Greenwich.

Fenn et al. (1985) in terms of their attenuation coefficients, particle size distribution, spectral optical properties, and height distribution. The user can select one of four levels (background, moderate, high, and extreme) and two classes of volcanic aerosols (fresh or aged). The final atmospheric constituents we may need to consider are clouds. While Earth observing remote sensors prefer to work under severe clear conditions, we must often deal with clouds. In general, we will characterize clouds by their elevation, thickness, type, and density. The type and density information typically provide information on the mixture of water droplets and ice crystals and on the optical absorption and angular scattering coefficients (magnitude and phase).

With the atmosphere characterized in this fashion, one can proceed to run one of the radiative transfer codes.

7.4.3.2 Modeling the Optical Properties of the Atmosphere

The goal of the radiation propagation models is to convert the atmospheric characterization data described in the previous section into optical properties and then to use the radiation propagation principles introduced in Chapters 3 and 4 to propagate the solar and self-emitted radiance along the relevant optical paths to eventually reach the sensor. Most of the basic principles employed by these models were introduced in Chapters 3 and 4, so we will tend to emphasize implementation issues and some advanced concepts in this section. Furthermore, each model uses slightly different approaches to implement the concepts treated here. The approach presented here is similar to that used in the MODTRAN code. However, the reader should not treat this as a one-to-one mapping onto any of the radiative transfer codes and should always refer to the relevant documentation for the specific implementation. Furthermore, until this point in this chapter, we have tended to isolate the reflective and thermal treatment of RT. For completeness, the rest of the discussion of RT codes will include both self-emitted and solar scattered terms. Note that this treatment will yield more rigorous results in the thermal infrared than the simplified (nonscattering) treatment in Section 7.3.3. However, except in cases of clouds, fogs, and large particulates, the scattering is quite weak, resulting in small changes in the computed radiance.

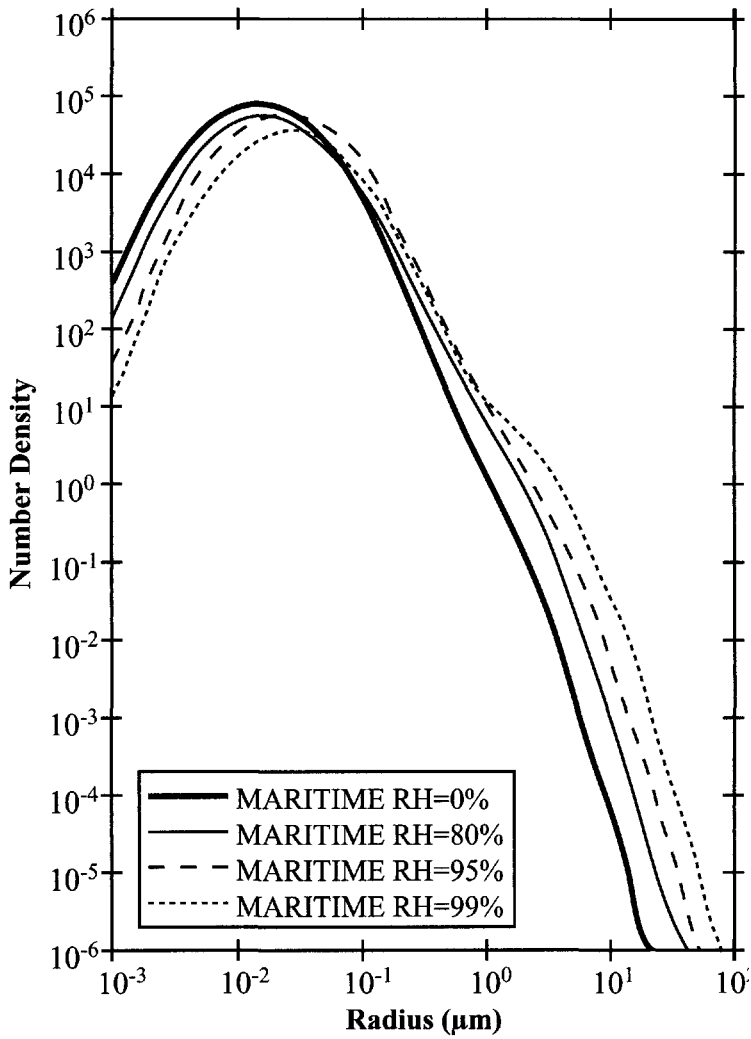


Figure 7.20 Example particle size distribution for maritime aerosols for different relative humidities. Each distribution is the weighted sum of two distributions, one from small particles (less than 1 μm) and one from large particles (approximately 1 micron). [Adapted from Fenn et al. (1985).]

Based on the treatment introduced in Chapter 4, we will focus on how the RT model gathers the relevant data about any layer and then consider propagation through the layers. We first begin with the gases. As indicated in Chapter 3, the gaseous optical depth due to absorption is a function of number density, absorption cross section, and path length. The number density for each constituent at each layer is readily calculated based on the temperature, pressure, and mixing ratio for the well-mixed gases and based on the temperature, pressure, and relative humidity for water vapor. Any other gases of interest are typically modeled for a range of conditions of interest (e.g., ozone) or can be adjusted by an advanced user. The absorption cross sections as a function of wavelength and temperature are typically computed from database values by interpolation between temperatures

and spectral convolution of detailed databases of the absorption line structure for each constituent along with line shape estimates of the background continuum. The MODTRAN code, for example, has a database based on convolution of the HITRAN database [cf. Rothman et al. (2003)]. FASCODE, on the other hand, is a line-by-line model that uses the full fidelity of the line structure in the HITRAN database [cf. Smith et al. (1978)]. As a result, FASCODE is more appropriate where very high spectral resolution is critical (e.g., for laser propagation). However, the run times are quite long, making the band models (MODTRAN) more appropriate where lower spectral resolution will suffice. In addition to absorption, we have to compute the optical depth due to Rayleigh scatter, which, as discussed in Chapter 3, is a function of number density and tabulated values for the index of refraction. The Rayleigh angular scattering functions will also be used to compute the energy directly scattered to the sensor, as well as the energy diffusely scattered up and down from each layer. The output from this process is the optical depth (and therefore the transmission) vertically through the layer, as well as along the solar beam and the target sensor beam (the L_1 and L_2 paths, respectively) for the layer as a function of wavelength for absorption and scattering by the well-mixed gases, water vapor, ozone and any other gas that needs to be treated separately due to unusual spatial distributions (e.g., HNO_3). In addition, at this point we can know the angular scattering function ($\beta(\theta)z$ [sr^{-1}]) due to the gases for any beam pair (e.g., L_1 and L_2) and path length (z) through the layer. Furthermore, by integration of the angular scattering function over a hemisphere, we can know the fractional amount of energy scattered upward or downward by the layer, i.e.,

$$r_i^+ = r_i^- = \int \beta_r(\theta) z d\Omega = 2\pi \int_{\theta=0}^{\frac{\pi}{2}} \beta_r(\theta) z \sin\theta d\theta \tag{7.63}$$

where r^+ or r^- indicate the fraction scattered upward or downward, respectively, by the ith layer. Note that the symmetry of the Rayleigh phase function dictates that the fractional amount scattered into any two hemispheres will be equal.

The next consideration for each layer is the impact of aerosols. From an optical standpoint, the treatment is very similar to that of the gases. Based on the concentration (number density), absorption coefficient, scattering coefficient, and path length, the transmission and scattering characteristics of each layer can be computed. In the case of the aerosols, a Mie scattering code must be employed (often prerun and the results accessed via lookup tables) that computes the wavelength-dependent angular scattering coefficient based on the complex index of refraction of the aerosol class, the particle size distribution, and the number density.

The gas and aerosol terms can then be combined for each layer and wavelength to generate the transmission due to absorption (τ_α), the transmission due to scattering (τ_{scat}), and the effective scattering function ($\beta(\theta)[m^{-1}sr^{-1}]$). Then the single scattering solution can be computed in a straightforward fashion based on the principles introduced in Chapters 3 and 4. The single scattering solution accounts for those photons that interact, at most, once with the target or the atmosphere. Referring to Figure 7.21, we can express the single scattered solution as:

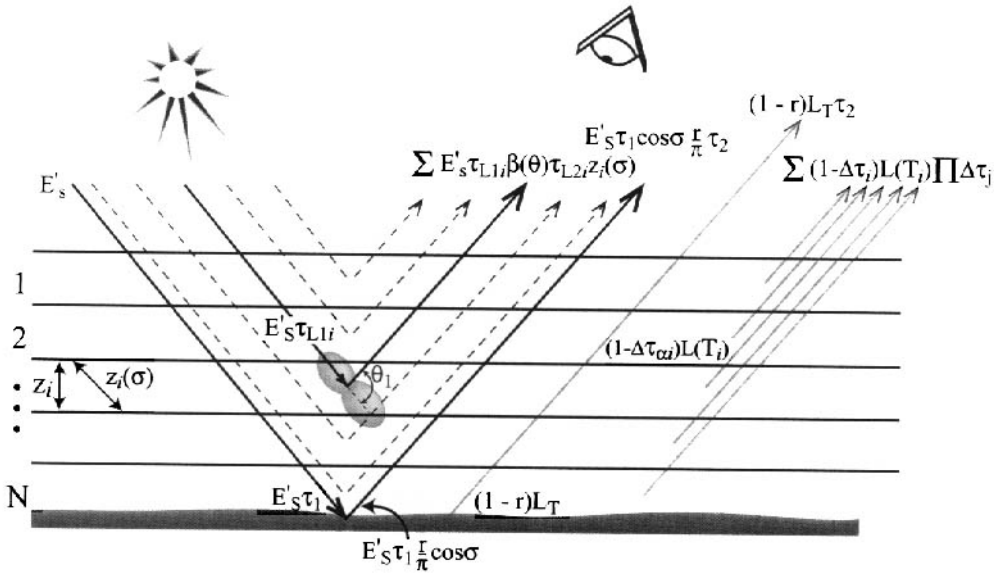


Figure 7.21 Illustration of single scattering and direct beam terms in radiative transfer code calculations [see Eq. (7.64)].

$$L_{ss\lambda} = E'_{s\lambda}\tau_1(\lambda)\tau_2(\lambda)\frac{r(\lambda)}{\pi}\cos\sigma + [1-r(\lambda)]L_\lambda(T)\tau_2(\lambda)$$
$$+\sum_{i=1}^{N} E'_{s\lambda}\tau_{L1i}(\lambda)\beta_{scai}(\lambda,\theta_v)\tau_{L2i}(\lambda)z_i(\sigma)$$
$$+\sum_{i=1}^{N}\left[[1-\Delta\tau_{\alpha i}(\lambda,\theta)]L_\lambda(T_i)\prod_{j=1}^{i-1}\Delta\tau_j(\lambda,\theta)\right] \quad (7.64a)$$

$$=\left[E'_{s\lambda}\tau_1(\lambda)\cos\sigma\frac{r(\lambda)}{\pi}+[1-r(\lambda)]L_\lambda(T)\right]\tau_2(\lambda)+L_{u\lambda} \quad (7.64b)$$

where all the terms are the same as described in Chapter 4 but now are derived from the databases and atmospheric descriptions associated with the RT code. For example, $\tau_1(\lambda)$ is the transmission along the sun target line of site incorporating the losses due to absorption and scattering by the gas molecules and aerosols in each layer corrected for the path length through the layer due to solar zenith angle and beam curvature due to refraction in the atmosphere, $z_i(\sigma)$ is the geometric path length through the *i*th layer along the solar illumination line, $\Delta\tau_{\alpha i}(\lambda,\theta)$ is the transmission loss through the *i*th layer along the sensor line of site due to absorption, such that $(1-\Delta\tau_{\alpha i})$ is the effective emissivity of the *i*th layer, and $r(\lambda)$ is the bidirectional reflectance factor or the diffuse reflectance for codes that do not accommodate bidirectional reflectance. The summation is over the layers from just below the sensor (*1*) to just above the Earth (*N*).

Inspection of Eq. (7.64) indicates that it is missing the downwelled radiance terms included in our discussion in Chapter 4. These can be estimated even with a

single scattering RT code by placing the sensor on the ground and looking to space for a range of zenith and azimuthal directions that span the sky dome. For any single run, the radiance calculation is essentially an inverted version of the calculations leading to Eq. (7.64) except there is no Earth reflection or self-emission term since deep space provides the boundary for these calculations.

The radiance downwelled from any direction is then expressed as

$$L_{dss}(\sigma_q,\phi_q)=\sum_{i=1}^{N}E'_{s\lambda}\tau_{L1i}(\lambda)\beta_{scai}(\lambda,\theta_v)\tau_{L2i}z_i(\sigma)$$
$$+\sum_{i=1}^{N}[1-\Delta\tau_{\alpha i}(\lambda,\sigma_q)]L_\lambda(T_i)\prod_{j=1}^{i-1}\Delta\tau_j(\lambda,\sigma_q) \tag{7.65}$$

where σ_q,ϕ_q are generic zenith and azimuthal direction angles and the numbering convention on the layers has been inverted to start at the lowest layer and proceed to space.

The bidirectionally reflected radiance at the surface headed toward the sensor could then be expressed as

$$L_{surf}(\sigma_s,\phi_s,\theta,\phi_d)=E'_s\tau_1\cos\sigma\frac{r(\sigma_s,\phi_s,\theta,\phi_d)}{\pi}+(1-r)L(T_o)$$
$$+\int L_{dss}(\sigma_q,\phi_q)\cos\sigma_q\frac{r(\sigma_q,\phi_q,\theta,\phi_d)}{\pi}d\Omega \tag{7.66}$$

where we have explicitly used the subscript s to indicate the dependence of the bidirectional reflectance factor on the solar illumination angles σ_s,ϕ_s to differentiate them from the generic direction angles σ_q,ϕ_q used in the integration of the downwelled radiance, and we have dropped the explicit wavelength dependency for compactness. Note that the first two terms in Eq. (7.66) are the directly reflected solar term and the thermal emission due to the target temperature T. This surface-leaving radiance can then be coupled with the target-sensor transmission (τ_2) and the upwelled radiance terms [cf. Eq. (7.64)] to generate the sensor-reaching radiance.

Incorporation of multiple scattering into radiative transfer codes is quite involved in both computational and theoretical terms. The details of this process are more involved than warrants thorough treatment here for most remote sensing users. However, a reasonably complete conceptual treatment will be presented as multiple scattering plays an important role in most radiance calculations in the visible region and at nearly all wavelengths where clouds are important.

As shown in Figure 7.22, multiple scattering includes several ray paths not adequately covered by any of our discussions thus far, as well as more rigorous methods to account for some of the approaches we have introduced (e.g., reflected downwelled). To provide a sense of the approaches used in multiple scattering codes, we will provide a simplified treatment of a two-stream model. A more complete discussion of the "Isaacs" two-stream model used in MODTRAN can be found in Isaacs et al. (1987). Our interest is in treatment of the diffuse scattering or diffusely emitted radiance from each layer. This is distinct from the radiance

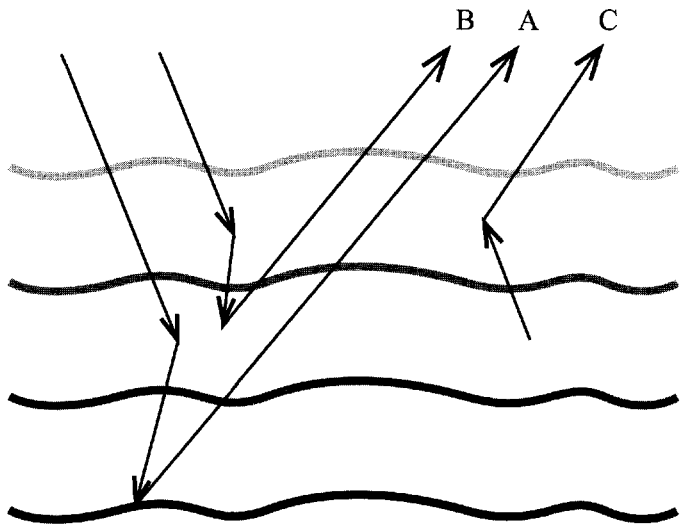


Figure 7.22 Examples of the many multiple scattering ray paths not fully treated by simple single scattering approaches. (a) Solar flux scattered by the atmosphere and then by the Earth. (b) Solar photons scattered by the atmosphere more than once. (c) Self-emitted photons scattered one or more times by the atmosphere. (Note that although in the simplest case this only involves one scattering event, it has not been dealt with in our single scattering treatment.)

scattered or emitted directly into the beam headed to the sensor that was characterized by our earlier treatment. In this case, we can think of each layer as a source of diffuse illumination that can be scattered from one or more other layers and eventually reach the sensor. Note that from this perspective, the target plane (the Earth in our case) defines a boundary layer that must be included in our multiple scattering interactions. Figure 7.23 illustrates this concept of *source functions* where we can treat the diffuse flux headed upward or downward from a layer as if it originated in that layer. The solar scattered radiance originating in the *i*th layer headed into any direction can be expressed as

$$L_{si}(\sigma_q,\phi_q) = E_s'\tau_{L1i}\beta_{scai}(\sigma_q)z_i(\sigma) \tag{7.67}$$

and the corresponding self-emitted radiance as

$$L_{\varepsilon i}(\sigma_q,\phi_q) = (1-\Delta\tau_{\alpha i})L(T_i) \tag{7.68}$$

The irradiance into the upper and lower hemispheres can then be expressed as

$$E_i^{\uparrow} = E_{si}^{\uparrow} + E_{\varepsilon i}^{\uparrow} = \int_{+} L_{si}(\sigma_q,\phi_q)\cos\sigma_q d\Omega + \int_{+} L_{\varepsilon i}\cos\sigma_q d\Omega \tag{7.69}$$

and

$$E_i^{\downarrow} = E_{si}^{\downarrow} + E_{\varepsilon i}^{\downarrow} = \int_{-} L_{si}(\sigma_q,\phi_q)\cos\sigma_q d\Omega + \int_{-} L_{\varepsilon i}\cos\sigma_q d\Omega \tag{7.70}$$

where the integral with the "+" indicates the upper hemisphere (solid hemisphere in Figure 7.23 and the "–" indicates the lower hemisphere (the dashed hemisphere in Fig. 7.23). Also, we need to adopt a convention that $\cos\sigma_q = |\cos\sigma_q|$ (i.e., negative values for $\cos\sigma_q$ imply irradiance headed into the lower hemisphere).

Referring to Figure 7.24, if we assume the multiscattered component of the flux will tend toward isotropic (Lambertian) behavior, we can compute the component of the downwelling diffuse irradiance that will be back-scattered by each layer according to

$$E_{backi} = \int_{back} {}^{i-1}E^{\downarrow}\beta_i(\sigma_q)z_i d\Omega = {}^{i-1}E^{\downarrow}\int_{\phi_q=0}^{2\pi}\int_{\sigma_q=0}^{\frac{\pi}{2}}\beta_i(\sigma_q)z_i \sin\sigma_q d\sigma_q d\phi_q \quad (7.71)$$

This is often expressed in the literature as

$$E_{backi} = {}^{i-1}E^{\downarrow}\omega_{0i}B_i = {}^{i-1}E^{\downarrow}R_i \quad (7.72)$$

where ${}^{i-1}E^{\downarrow}$ is the total diffuse irradiance from the layers above incident on the ith layer, ω_{0i} is the *single scattering albedo* of the ith layer (i.e., the fraction of the total loss in the ith layer due to scattering), B_i is the backscatter fraction (i.e., the fraction of the scattered energy into the solid hemisphere in Fig. 7.24 as dictated by the scattering phase function), E_{backi} is the component of the incident irradiance from above scattered backward. and R_i is the effective diffuse reflectivity of the ith layer.

Note that

$$\omega_{oi} = \int\beta_i(\sigma_q)z_i d\Omega = \int\frac{\beta_{scai}}{4\pi}p_i(\sigma)z_i d\Omega \quad (7.73)$$

and

$$B_i = \frac{1}{4\pi}\int_{back}p_i(\sigma_q)d\Omega = \frac{1}{4\pi}\int_{\phi_q=0}^{2\pi}\int_{\sigma_q=0}^{\pi/2}p_i(\sigma_q)\sin\sigma_q d\sigma_q d\phi_q \quad (7.74)$$

where β_{scai} and $p_i(\sigma_q)$ are the scattering coefficient and the scattering phase function, respectively for the ith layer as introduced in Chapter 3. Similarly, the diffuse irradiance scattered forward can be expressed as

$$E_{sca\ forward\ i} = \int_{forward} {}^{i-1}E^{\downarrow}\beta_i(\sigma_q)z_i d\Omega = {}^{i-1}E^{\downarrow}\int_{\phi_q=0}^{2\pi}\int_{\sigma_q=\frac{\pi}{2}}^{\pi}\beta_i(\sigma_q)z_i \sin\sigma_q d\sigma_q d\phi_q \quad (7.75)$$

$$= {}^{i-1}E^{\downarrow}\omega_{0i}(1-B_i) \quad (7.76)$$

where $E_{\text{sca forward}}$ is the component of the incident irradiance from above scattered forward, if only scattering is considered, and Eq. (7.76) neglects the absorption by the layer and the direct beam transmission, which can be included to yield

$$E_{forward\ i} = E_{sca\ forward\ i} + {}^{i-1}E^{\downarrow}\Delta\tau_{\alpha i} = {}^{i-1}E^{\downarrow}\omega_{0i}(1-B_i) + {}^{i-1}E^{\downarrow}\Delta\tau_{\alpha i} = {}^{i-1}E^{\downarrow}T_i \quad (7.77)$$

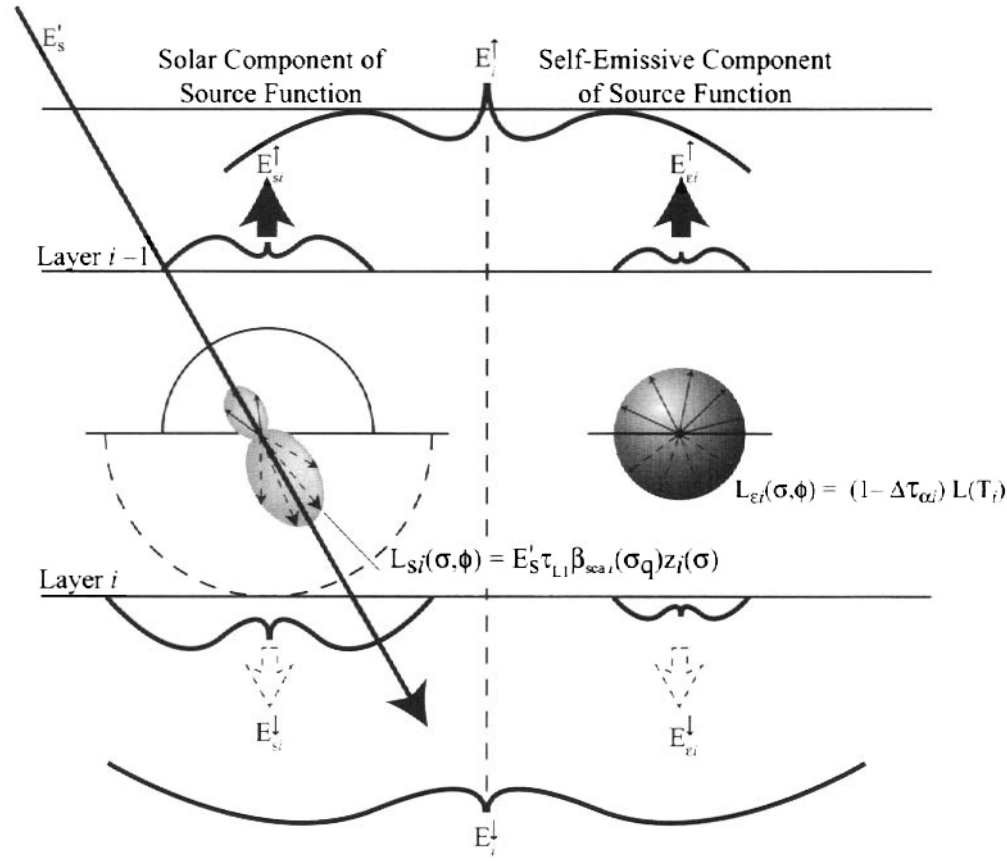


Figure 7.23 Illustration of the diffuse components of the source functions for the radiance (irradiance) from the *i*th layer of a multiple scattering code. The combined solar and self-emitted components represent the diffuse irradiance originating in the layer and emerging from the *i*th layer headed up ($E_i^{\uparrow}$) and down ($E_i^{\downarrow}$).

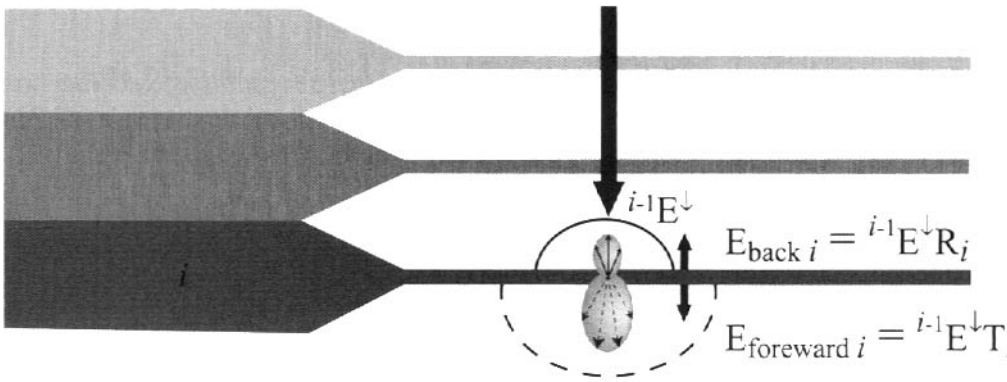


Figure 7.24 Illustration of diffuse irradiance reflectance and transmission concepts. For convenience, we have conceptually compressed the atmospheric layers into planer diffusely reflecting and transmitting layers.

where $E_{\text{forward } i}$ is the component of the diffuse incident irradiance transmitted through the ith layer and T_i is the effective diffuse transmittance of the ith layer. The definition of R_i and T_i is most obvious if we rearrange Eqs. (7.72) and (7.77) to yield

$$R_i = \frac{E_{back\ i}}{{}^{i-1}E^{\downarrow}} = \omega_{0i} B_i \tag{7.78}$$

and

$$T_i = \frac{E_{forward\ i}}{{}^{i-1}E^{\downarrow}} = \frac{{}^{i-1}E^{\downarrow}\omega_{0i}(1-B)}{{}^{i-1}E^{\downarrow}} + \Delta\tau_{\alpha i} = \omega_{0i}(1-B_i) + \Delta\tau_{\alpha i} \tag{7.79}$$

With these preliminaries, we can now describe the radiation transfer using layer adding as described by Isaacs et al. (1987). Let's begin by considering the downwelling radiance reaching the Earth starting at the top of the atmosphere, as illustrated in Figure 7.25. This is a convenient point because at the topmost or boundary layer, we know there is no diffuse irradiance incident on the layer from above and none of the flux scattered upward is scattered back down. Thus, the irradiance leaving the first layer headed down in the absence of other layers can be expressed as the downward source function from Eq. 7.70, i.e.,

$${}^{1}E^{\downarrow} = E_1^{\downarrow} \tag{7.80}$$

where we have introduced the convention that a left-hand superscript will indicate the aggregate behavior of the layer indicated and all those above (or below depending on the context). If we consider two layers, we need to propagate the flux from 1 down through 2, account for the multiple scattering of downward flux between layers 2 and 1, include the downward source function for layer 2, and include the multiple scattering between layer 2 and layer 1 for flux originating in 2 and heading up. Thus, the total irradiance leaving layer 2, headed down, when layers 1 and 2 are combined in the absence of others can be expressed as

$$\begin{aligned}
{}^{2}E^{\downarrow} &= E_2^{\downarrow} + {}^{1}E^{\downarrow}T_2 + {}^{1}E^{\downarrow}T_2R_2R_1 + {}^{1}E^{\downarrow}T_2(R_2R_1)^2 + \cdots \\
&\quad + E_2^{\uparrow}T_2R_1 + E_2^{\uparrow}T_2R_1^2R_2 + E_2^{\uparrow}T_2R_1^3R_2^2 \cdots \\
&= E_2^{\downarrow} + {}^{1}E^{\downarrow}T_2[1 + R_2R_1 + (R_2R_1)^2 + \cdots] + E_2^{\uparrow}T_2R_1[1 + R_1R_2 + (R_1R_2)^2 + \cdots] \\
&= E_2^{\downarrow} + {}^{1}E^{\downarrow}T_2\left(\frac{1}{1-R_1R_2}\right) + E_2^{\uparrow}T_2R_1\left(\frac{1}{1-R_1R_2}\right)
\end{aligned} \tag{7.81}$$

where we recognize that R_1R_2 will always be less than one and use a series representation for the infinite sum. Before adding a third layer, it is useful to introduce the composite diffuse reflectance of layers 1 and 2 to flux from below layer 2 headed up (i.e., reflected back down) as

$${}^{2}R^{\downarrow} = R_2 + T_2^2R_1 + T_2^2R_1^2R_2 + T_2^2R_1^3R_2^2 \cdots = R_2 + T_2^2R_1\left(\frac{1}{1-R_1R_2}\right) \tag{7.82}$$

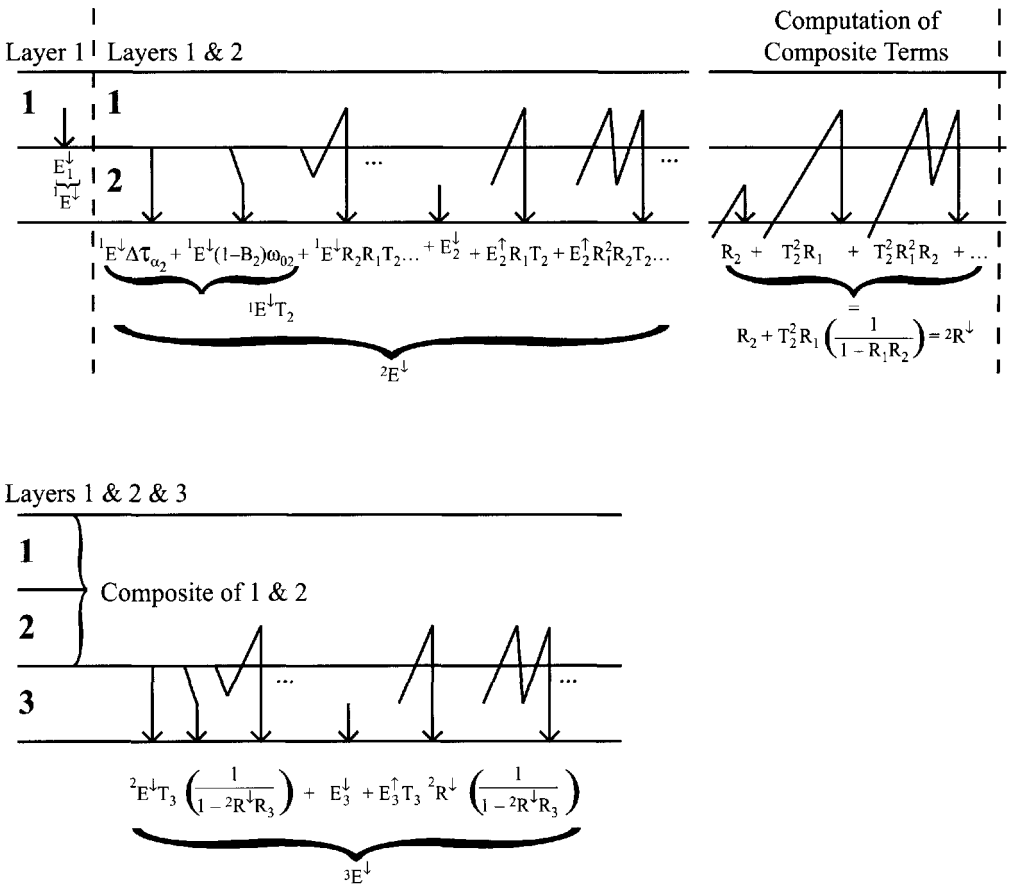


Figure 7.25 Illustration of concepts involved in combining source terms and multiple scattering terms to generate the composite irradiance from multiple layers.

or in general,

$$ {}^{i}R^{\downarrow} = R_i + T_i^2 \, {}^{i-1}R^{\downarrow} \frac{1}{1 - {}^{i-1}R^{\downarrow} R_i} \tag{7.83} $$

Taking advantage of this composite term, the diffuse irradiance leaving the third layer headed down can then be expressed as

$$ {}^{3}E^{\downarrow} = E_3^{\downarrow} + {}^{2}E^{\downarrow} T_3 \left(\frac{1}{1 - {}^{2}R^{\downarrow} R_3} \right) + E_3^{\uparrow} T_3 \, {}^{2}R^{\downarrow} \left(\frac{1}{1 - {}^{2}R^{\downarrow} R_3} \right) \tag{7.84} $$

Clearly, we can continue to add layers in this fashion until we reach the bottom using the general form of Eq. 7.84, i.e.,

$$ {}^{i}E^{\downarrow} = E_i^{\downarrow} + {}^{i-1}E^{\downarrow}T_i\left(\frac{1}{1-{}^{i-1}R^{\downarrow}R_i}\right) + E_i^{\uparrow}T_i\,{}^{i-1}R^{\downarrow}\left(\frac{1}{1-{}^{i-1}R^{\downarrow}R_i}\right) \tag{7.85} $$

At this point, we have the diffuse irradiance from the sky leaving the *n*th or bottom most layer expressed as ${}^{N}E^{\downarrow}$ and the effective diffuse reflectivity of the sky expressed as ${}^{N}R^{\downarrow}$. This term is more commonly referred to as the spherical albedo of the sky using the variable S, i.e.,

$$ S = {}^{N}R^{\downarrow} \tag{7.86} $$

Recognizing the Earth as the final or lower boundary layer, we can compute the surface leaving radiance from a Lambertian Earth incorporating multiple scattering as

$$ L_{surf\ ms} = \left[E_s'\tau_1\cos\sigma\frac{r_d}{\pi} + (1-r_d)L(T) + {}^{N}E^{\downarrow}\frac{r_d}{\pi}\right]\left(\frac{1}{1-Sr_d}\right) \tag{7.87} $$

The upwelled radiance from the air column (L_u), including multiple scattering, can be computed using the same source functions and layer adding approach described above for downwelled radiance. In general, the single scattering and direct beam component of the upwelled radiance can be computed using the single scattering approach described in Eq. (7.64). This preserves the strong directional component of the upwelled radiance with the multiply scattered radiance captured by the two-stream method added in to yield the total radiance. This requires that the single scattered contribution to the multiple scattering solution be subtracted out since it is being replaced by the upwelled term from Eq. (7.64). This yields an expression from the upwelled radiance of the form

$$ \begin{aligned} L_u = &\left[{}^{N}E^{\uparrow} - \sum_{i=1}^{N}E_i^{\uparrow}\prod_{j=1}^{i=1}\Delta\tau_j\right]/\pi + \sum_{i=1}^{N}E_s'\tau_{L1i}\beta_{scai}(\theta_v)\tau_{L2i}z(\sigma) \\ &+\sum_{i=1}^{N}[1-\Delta\tau_{\alpha i}(\theta)]L(T_i)\prod_{j=1}^{i-1}\Delta\tau_i(\theta) \end{aligned} \tag{7.88} $$

Note that because we are assuming approximately isotropic behavior, we can simply take the multiply scattered irradiance from the air column reaching the top of the atmosphere and divide by π to estimate the multiple scattered component of the upwelled radiance reaching the sensor.

The reader should recognize that as involved as this treatment of RT codes may seem, it is, nevertheless, a highly simplified version compared to many of the actual codes (e.g., MODTRAN). Many of the codes involve a more complete treatment of multiple scattering to better deal with directional effects such as the discrete ordinates (DISORT) approach available in MODTRAN [cf. Stamnes et al. (1988)]. In addition, the treatment presented here is only valid for discrete wavelengths. Band model codes such as MODTRAN use bandpass values for the absorption coefficients in their spectral databases. In bands that include narrow ab-

sorption lines, the transmission over a path is not a simple product of the spectrally integrated transmission values in the layers along the path (cf. Fig. 7.26). Instead, computational tools are used to generate effective absorption coefficients for each spectral band [cf. Berk et al. (1998)] that account for the line structure within the band. Figure 7.27 shows some of the utility of the advanced RT codes in terms of their ability to generate estimates of all of the parameters involved in the governing equation. This allows the use of the RT codes to support not only the atmospheric compensation process we have been discussing in this section, but also phenomenology and image analysis tasks. In particular, in Chapters 11 and 14 we will use the RT codes extensively in support of model-based algorithms and synthetic scene generation.

Most atmospheric propagation models use numerical solutions to radiance equations similar to those described here, coupled with tables of data on the spectral properties of the atmospheric constituents. The main limitation with MODTRAN and other similar atmospheric propagation models in the wavelength regions where solar scattering is important is the difficulty in adequately characterizing the make-up of the atmosphere (particularly aerosols). As a result, the angular scattering coefficients associated with the atmospheric layers are often poorly characterized.

Slater et al. (1987) have shown how atmospheric propagation models can be used to characterize the atmosphere when supplemented with ground-based sun photometer readings. The sun photometers are used over time to measure the optical depth of the atmosphere using the *Langley plot* method [cf. Slater (1980)]. The total atmospheric optical depth δ is computed as minus the slope of the logarithmic relation between the observed irradiance (E_{obs}) and the solar zenith angle (σ), i.e.,

$$\ln E_{\text{obs}} = \ln(E_s' e^{-\delta \sec \sigma}) = \ln E_s' - \delta \sec \sigma \tag{7.89}$$

where E_{obs} is the irradiance recorded by a sensor aimed directly at the sun with a field of view just large enough to fully encompass the solar disk. To simplify the treatment, the total optical depth can be treated as

$$\delta = \delta_g + \delta_w + \delta_a \tag{7.90}$$

where δ_g is the optical depth associated with the permanent gasses, δ_w is the optical depth associated with water vapor, and δ_a is the optical depth associated with aerosols. If δ_g and δ_w can be well estimated by atmospheric propagation models where good radiosonde data are available, then aerosol optical depth (δ_a) can be found from Eq. (7.90) using model-derived values for δ_g and δ_w and the experimental value of the overall optical depth (δ). The values of aerosol concentration input to the RT model can then be adjusted such that the model predicts the observed value (by working at several wavelengths significant improvements in the characterization of the water vapor and aerosols is possible). The "calibrated" model can then be used to compute the necessary terms in the governing equation. Slater et al. (1987) have shown that this type of approach can work quite well, particularly in low-turbidity atmospheres. Ongoing limitations are associated with the need for good radiosonde data, surface irradiance measurements, and stable atmospheres, all of which can be prohibitive. It is particularly important to realize that if one

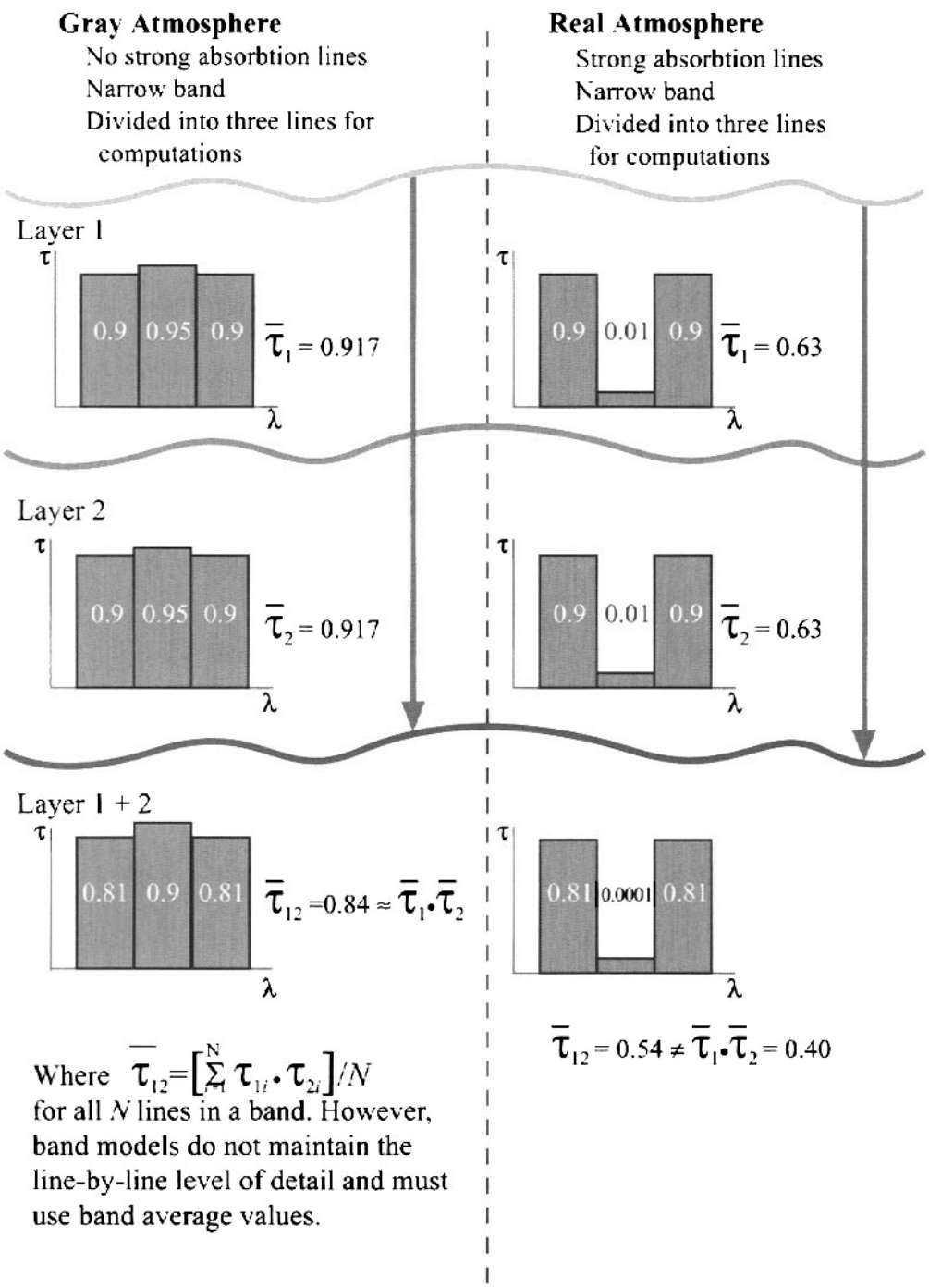


Figure 7.26 Example illustrating that the product of band average transmissions to compute the total transmission through multiple layers fails when strong absorption features are present in a band. To overcome these limitations, effective band average values are used in radiative transfer codes using band models[see Berk et al. (1998) for the MODTRAN correlated k treatment].

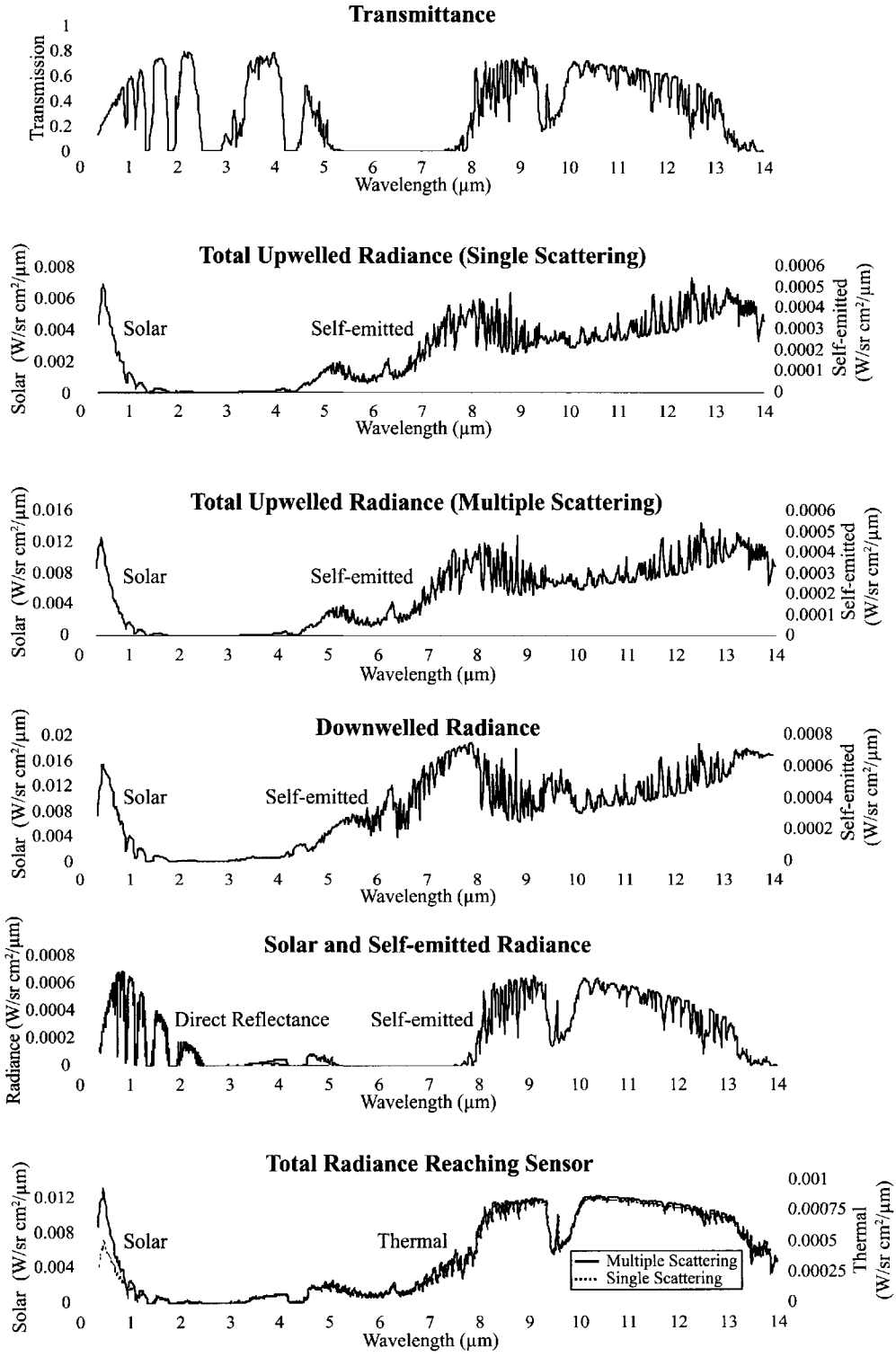


Figure 7.27 Examples of MODTRAN output showing the contribution of various terms to the sensor reaching spectral radiance. This example is for mid-latitude summer, urban aerosols, 12 km visibility, at noon looking from space at a 10% reflecting, 300 K target in Rochester, NY, on June 22.

does not use some method of enhancing the calibration of the atmospheric models (e.g., the Langley plot method), significant errors can occur, particularly at shorter wavelengths where scattering is important. This is because scattering is such a significant component of the transmission, illumination, and path radiance terms, and the radiosonde data, upon which these models rely, is insufficient to predict the magnitude and phase of the aerosol scattering function.

The radiation propagation models are very attractive, because they will yield a complete solution for all of the parameters in the governing equation at all wavelengths. In addition, these models include corrections for many of the variations in the atmospheric parameters due to changes in elevation, slant range, and view angle. Most of the other methods assume that these values are constant over the study area. However, the errors from the models in spectral regions where scattering is important can be very large if the models are not somehow adjusted for the magnitude and often the shape (phase function) of the scattering coefficient. Note that this is not a fundamental limitation of the models, but of our knowledge of the input parameters the models require. As methods evolve to calibrate the inputs to these models, they will undoubtedly play an increasing role in atmospheric compensation. At present, essentially the only way to properly calibrate the models is with extensive ground truth, e.g., field radiometry measurements that are impractical for most operational uses. Ongoing research studies are attempting to design and test alternative methods for characterizing atmospheric parameters that could be used to improve the input to RT models. The *Multiangle Imaging Spectro Radiometer (MISR)* program, for example, uses multiple looks at the same targets from space to attempt to characterize the type and density of aerosols [cf. Martonchik et al. (2002)].

Before leaving this treatment of RT models, we should also note their potential to aid in compensation for the impact of thin *cirrus clouds*. These thin cirrus clouds are often referred to as subvisual cirrus because they can be difficult to see in the visible region. They are made up of ice crystals in the upper atmosphere (usually just below the tropopause) and can reduce the transmission and increase the path radiance even when they are not visually discernable in the image data. They are of particular concern because their probability of occurrence is quite high. Because the cirrus clouds are so high, they are generally more of a problem for high-altitude (7 km or more) and satellite sensors. The cirrus clouds can sometimes be detected as cold features in thermal band imagery. However, many new sensors are taking advantage of the 1.37 μm spectral region to observe and characterize cirrus clouds. In this region, the water vapor in the atmosphere is highly absorptive, and essentially no sunlight reaches the Earth. In fact, most of the solar irradiance is absorbed before any significant scattering occurs such that the Earth is nearly black in this spectral region when observed from high-altitudes. However, when cirrus clouds are present, the ice crystals scatter solar irradiance that can be used as an indicator of the presence of cirrus clouds (cf. Fig. 7.28), and the magnitude of the return may be an indicator of the impact of the cirrus clouds in other spectral regions. This approach is being used to generate cirrus cloud masks using MODIS data and to compensate for cirrus effects in the VNIR region [cf. Gao et al. (1997)]. The relationships between the observed behavior in the 1.37 μm region

and the expected behavior in other spectral regions is largely driven by RT models such as MODTRAN, which is often employed in supporting the generation of empirical relationships used in compensating for cirrus effects [cf. Adler-Golden et al. (1999)].

Because of the ongoing difficulties in achieving a full calibration to reflectance, many users have opted for various types of relative calibration techniques. Also, in many cases, our knowledge of the relationship between reflectance and application-specific parameters is still poorly characterized and is often only expressed in relative terms (e.g., spectral reflectance ratios or changes in reflectance over time). Thus, in many applications relative calibration may provide much of the data that can be readily used given our current knowledge base.

7.5 RELATIVE CALIBRATION (REFLECTANCE)

Of all the various relative calibration or normalization approaches, we will concentrate on two general categories. The first is the use of ratios or relative spectral information between multiple spectral bands of the same scene. The second involves the use of normalization techniques to reduce atmospheric variations between multiple images of the same area acquired at different times.

7.5.1 Spectral Ratio Techniques

As introduced earlier (Sec. 7.4.2), in certain cases where the absolute reflectance is not required, the relative reflectance between spectral bands can provide significant information. Piech et al. (1978) point out how the blue-to-green reflectance ratio can be very useful in the study of water quality, and many investigators have used near infrared to red reflectance ratios to characterize vegetation condition. Piech et al. (1978) describe how reflectance ratios can be estimated given that one knows the upwelled radiance (L_u) in each spectral band (e.g., by using the shadow method or the minimum radiance method). The radiance in spectral bands 1 and 2 can then be expressed as

$$L_1 = (E_{s1}\pi^{-1}\tau_{21} + L_{d1}\tau_{21})r_{d1} + L_{u1} = \alpha_1 r_{d1} + L_{u1} \tag{7.91}$$

$$L_2 = (E_{s2}\pi^{-1}\tau_{22} + L_{d2}\tau_{22})r_{d2} + L_{u2} = \alpha_2 r_{d2} + L_{u2} \tag{7.92}$$

where L_1 and L_2 are the observed radiance values in bands 1 and 2 for an object with diffuse reflectivity values r_{d1} and r_{d2}. These equations can be combined to yield

$$L_1 = \frac{\alpha_1}{\alpha_2}\frac{r_{d1}}{r_{d2}}L_2 - \frac{\alpha_1}{\alpha_2}\frac{r_{d1}}{r_{d2}}L_{u2} + L_{u1} \tag{7.93}$$

Let's assume a class of objects can be identified whose spectral reflectance ratios approximate a constant of known value. For example, Piech et al. (1978) observed that the class of manmade "gray" objects (e.g., concrete, asphalt, gravel

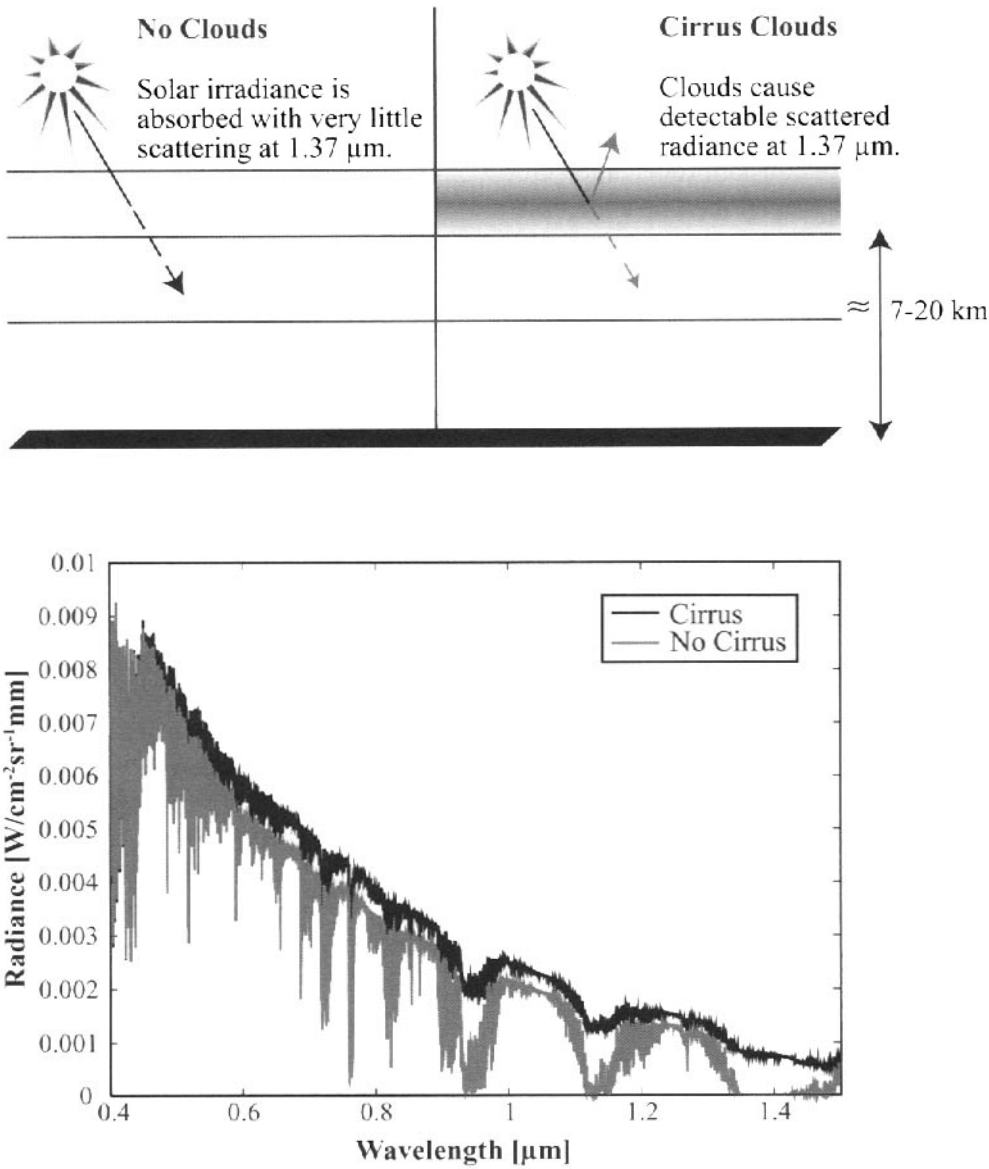


Figure 7.28 Illustration of the use of the 1.37 μm spectral band to detect and potentially compensate for the presence of cirrus clouds. The spectral radiance plots are for a vertical viewing sensor in space looking at a 10% reflecting Earth.

roofs) exhibited this behavior. They generated average reflectance ratio values for the visible and near-infrared reflectance regions based on many laboratory and field samples. For such a class of objects, Eq. (7.93) becomes a linear equation of the form

$$L_1 = \frac{\alpha_1}{\alpha_2} k_{12} L_2 - \frac{\alpha_1}{\alpha_2} k_{12} L_{u2} + L_{u1} \tag{7.94}$$

$$L_1 = mL_2 + b \tag{7.95}$$

where m and b are the slope and intercept of a linear regression of the radiance observed in band 2 against the radiance observed in band 1 for samples from the class of objects with constant reflectance ratio $k_{12} \cong r_{d1}/r_{d2}$. The reflectance ratio for any unknown sample can then be expressed as

$$\frac{r_{d1}}{r_{d2}} = \frac{L_1 - L_{u1}}{L_2 - L_{u2}} \frac{k_{12}}{m} \tag{7.96}$$

The precision of this method was observed to be approximately 10% of the measured ratio when the upwelled radiance values (L_u) were computed using the shadow method.

Many investigators have found that by simply ratioing raw digital counts, they can crudely approximate reflectance ratios with significant, but often tolerable errors. The commonly used *normalized difference vegetation index* (*NDVI*), suggested by Rouse et al. (1973), is a slightly more sophisticated version of this approach, which attempts to reduce atmospheric and illumination effects by using a differencing and ratioing method applied to observed digital count values, i.e.,

$$\text{NDVI} = \frac{\text{DC}_{\text{IR}} - \text{DC}_{\text{R}}}{\text{DC}_{\text{IR}} + \text{DC}_{\text{R}}} \tag{7.97}$$

where DC_{IR} and DC_{R} are, respectively, the digital count values in the near infrared and red spectral bands. Figure 1.2 illustrates how several AVHRR images can be combined to generate an NDVI image over very large regions. By producing sequences of NDVI images over time, large-scale changes in biomass can be monitored.

7.5.2 Scene-to-Scene Normalization

In cases where only the relative change between two images is of interest, or where we only want to improve the appearance of an image, radiometric normalization techniques can be used.

The simplest of these is a method called *histogram matching* or *histogram specification* [cf. Gonzalez and Wintz (1987)]. The method involves passing one image through a lookup table that attempts to adjust the histogram to match some specified histogram (cf. Sec. 8.1). Schott et al. (1985) utilized this approach to visually enhance satellite images by removing the apparent atmospheric effects. This was accomplished by matching the histogram of each spectral band to the corresponding histogram of a very low-altitude image of similar scene content. Since the low-altitude scene had almost no atmospheric effects, the resulting satellite image "appears" to have the atmospheric effects removed. This is a nonquantitative approach aimed solely at improving the visual appearance.

This same approach has been used to normalize images of the same scene taken at different times. The process removes the dominant brightness variations

between the scenes caused by illumination and sensor response variations. The resultant scenes "appear" as if they were taken under the same imaging conditions (i.e., same sensor response, same atmospheric effects, and same illumination levels). However, because the histogram matching technique is a simple numerical fit, it can mask real changes between the images. For example, given a scene where a major portion of the image area was water that was turbid in the first scene and clear in the second, the histogram matching approach would try to force the two scenes to look the same. As a result, after modification, the second scene would not only look like the first in terms of the imaging conditions but also what was clear water would appear turbid, thus masking rather than enhancing real differences between the scenes. Because the transforms are based on whole-image statistics, this masking effect will be minimized when the differences between the scenes only represent a small number of pixels. Small-scale effects like new buildings or roads will be clearly differentiated. Large-scale subtle changes, e.g., water quality effects and vegetation condition, will tend to be masked by the process.

Building on the histogram matching approach, several authors have suggested a more quantitative scene-to-scene normalization built on radiation propagation principles and designed to eliminate the undesirable effects of the global histogram matching approach. To derive the normalization transform, the radiance on each day is expressed as [cf. Eq. (7.91)]

$$L_i = \alpha_i r_{di} + L_{ui} \tag{7.98}$$

The digital count on each day can then be expressed as

$$\mathrm{DC}_i = g_i L_i + b_{oi} \tag{7.99}$$

$$= g_i \alpha_i r_{di} + g_i L_{ui} + b_{oi} \tag{7.100}$$

if the sensor's response is linear with radiance. For each band, the relationship between the digital count on day 1 (DC_1) and day 2 (DC_2) for pixels with the same reflectance can be written as

$$\mathrm{DC}_1 = \frac{g_1\alpha_1}{g_2\alpha_2}\mathrm{DC}_2 - \frac{g_1\alpha_1}{g_2\alpha_2}(g_2 L_{u2} + b_{o2}) + g_1 L_{u1} + b_{o1} \tag{7.101a}$$

$$= m\mathrm{DC}_2 + b \tag{7.101b}$$

where m and b are the coefficients that would transform a digital count on day 2 to have the same digital count on day 1 for pixels that have the same reflectance on each day (cf. Fig. 7.29). That is, this transform would make the day-2 image appear quantitatively as though it were taken under the same sensor response (gain and bias), illumination, and atmospheric conditions (α and L_u) as the day-1 image. As a result, after the transform any differences between the images would be real since pixels with the same reflectance on the two days would have the same digital count.

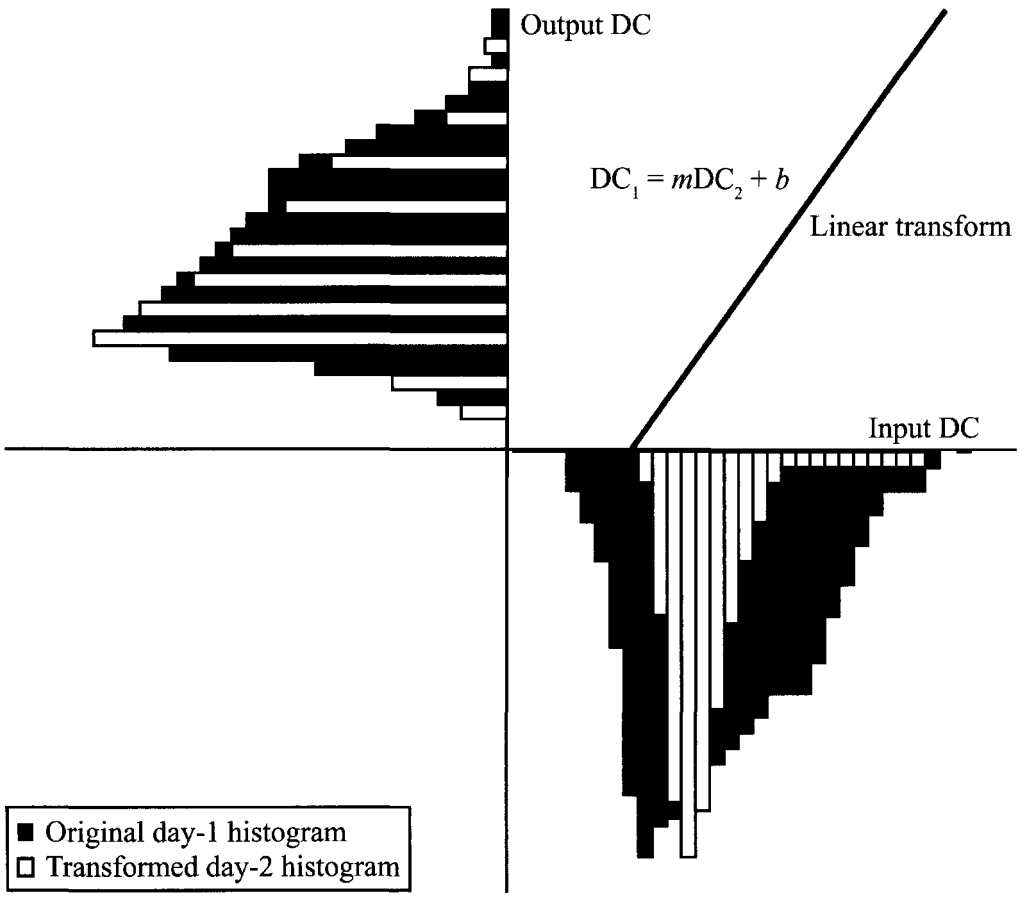


Figure 7.29 Linear lookup table utilized to transform a day-2 image to appear as if imaged under the same conditions as the day-1 image.

The normalization process reduces to solving for the linear transform coefficients [m and b in Eq. (7.101b)] for each spectral band of interest. The simplest solution [cf. Hall et al. (1991)] is to identify two objects, or classes of objects, whose reflectance is assumed constant between the two images. To reduce errors, a dark object and a bright object are chosen, and the digital counts (or mean digital counts from a sample) for the objects are acquired from each image. These "dark object–bright object" samples yield a two-point solution to the simple linear equation relating the digital counts on the two days [Eq. (7.101b)]. This method assumes that the objects chosen (or their mean if sampling is done) will have the same reflectance on both days. This method can be difficult to apply in scenes where unique objects with invariant reflectances are difficulty to identify. Even where invariant objects exist, due to the small sample size, a slight error in one term can lead to large normalization errors (e.g., if the dark object is water and the water on day 2 is more turbid than on day 1). A regression method used by Jensen (1983) overcomes this sample limitation by spatially registering the two scenes and then performing a simple linear regression of day-1 digital counts against day-2 digital counts to find the coefficients (m and b) in Eq. (7.101b). This method

implicitly assumes that the average reflectance values on day 1 and day 2 are the same and that the large sample size and least-squares method will reduce the errors due to those pixels whose reflectance has changed. This can be a problem when changes occur over a large portion of an image, since these real changes will tend to be normalized out by the regression process. Problems can also be introduced if significant changes occur between the two dates in a feature class that represents sufficient pixels to perturb the regression result.

Yuan and Elvidge (1993) have suggested a method for interactively selecting a subset of the total scene for input to the regression calculation. This method is designed to reduce effects due to class differences between the scenes, including effects such as clouds, snow, and shadows. However, subtle changes in vegetation or water covering large portions of the scene will still tend to be masked by this enhanced regression method. Schott et al. (1988) proposed a method that uses the invariant reflectance concept employed by the "dark object–bright object" method but employs large sample sizes (like the regression method) to reduce errors. It is premised on the user's ability to isolate a class of pixels called *pseudo-invariant features* (*PIF*) whose reflectance distribution is assumed to remain nearly constant over time (e.g., manmade surfaces such as roads and rooftops). The values of m and b in Eq. (7.101b) can then be expressed as

$$m = \frac{\sigma_1}{\sigma_2} \tag{7.102}$$

$$b = \mathrm{DC}_{1\mathrm{avg}} - \frac{\sigma_1}{\sigma_2}\mathrm{DC}_{2\mathrm{avg}} \tag{7.103}$$

where σ_1 and σ_2 are the standard deviations of the class of invariant pixels segmented on day 1 and day 2, respectively, and $\mathrm{DC}_{1\mathrm{avg}}$ and $\mathrm{DC}_{2\mathrm{avg}}$ are the means of these same classes. Salvaggio (1993) indicates that changes of more than about one reflectance unit should be above the noise level of this process in Landsat TM scenes where a good sampling of PIFs is possible. This method is limited, however, to scenes where a reasonable number of invariant features can be isolated. Figure 7.30 shows an example of the PIF transformation applied to a portion of a Landsat TM scene.

In addition to the simple ratios and NDVI methods presented in this section, a number of image processing techniques have been employed to reduce the effect of the atmosphere on image analysis. Among the most common of these is the principal component transform described in Section 9.3.

7.6 COMPENSATION OF IMAGING SPECTROMETER DATA FOR ATMOSPHERIC EFFECTS

A large proportion of the new work in the field of atmospheric compensation is focused on the potential of new sensors. In particular, hyperspectral imaging (introduced in Sec. 6.2), in addition to providing a new way to sample information about

surfaces, also provides much more information about the atmosphere. The narrow bandpass data allow us to sample the absorption lines of several of the relevant atmospheric constituents, and thereby back out their effects on the overall scene. These sensors offer the potential for directly measuring the detailed condition of the atmosphere using overhead imagery.

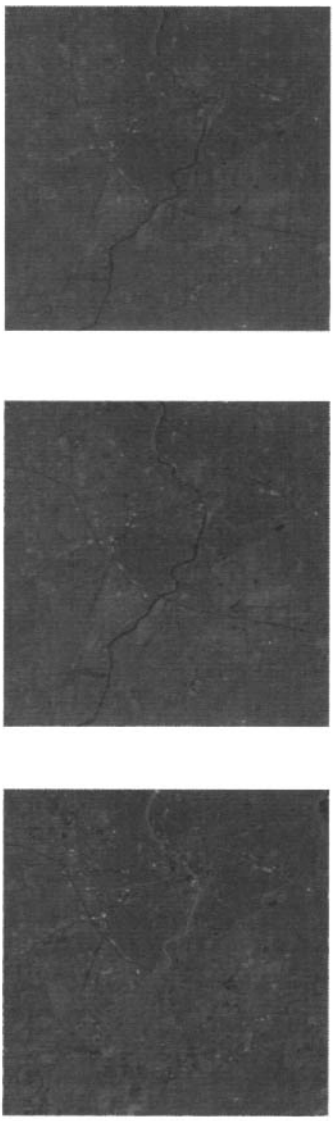


Figure 7.30 Pseudo-invariant feature (PIF) transformation applied to a portion of a Landsat scene. See color plate 7.30.

7.6.1 Inversion to Reflectance

In early work, Gao and Goetz (1990) showed how water vapor values can be spatially mapped by analyzing airborne imaging spectrometer data. This method uses the water vapor absorption lines at 0.94 or 1.14 μm, as well as the region on either side of the absorption lines as observed by the radiance recorded by an imaging spectrometer. If the scattered terms (L_u and L_d) are assumed to be small, the product of target reflectance times total path transmission can be expressed as a radiance ratio ($R_{\Delta\lambda}$), i.e.,

$$R_{\Delta\lambda} = \frac{L_{\Delta\lambda}}{E'_{s\Delta\lambda}\pi^{-1}\cos\sigma} \cong \tau_{1\Delta\lambda}\tau_{2\Delta\lambda}r_{d\Delta\lambda} \tag{7.104}$$

where $L_{\Delta\lambda}$ is the radiance in a narrow bandpass $\Delta\lambda$ observed with a calibrated imaging spectrometer, $E'_{s\Delta\lambda}\pi^{-1}\cos\sigma$ is the radiance in the bandpass for a 100% reflector at the top of the atmosphere (this can be obtained from tabulated solar spectral data, and the earth–sun distance), $\tau_{1\Delta\lambda}\,\tau_{2\Delta\lambda}$ represents the transmission along the total sun-target-sensor path, and $r_{d\Delta\lambda}$ is the reflectance in the narrow bandpass. From spectrometer data of a single pixel, a spectral shape for the total path transmission modified by reflectance can be represented as the radiance ratio from the spectrometer data. The reflectance is assumed to be approximately linear over a narrow spectral region, including the absorption line and its shoulders. If the transmission just outside the absorption line to either side is assumed constant, then the shape of the reflectance "line" in the region of interest can be defined using a straight-line fit between the wings of the absorption line. An estimate of the total path transmission ($\tau_{1\Delta\lambda}\,\tau_{2\Delta\lambda}$) can be produced using an initial estimate for the atmospheric makeup and a high-spectral-resolution atmospheric propagation model. This is modified by the spectral shape of the reflectance and convolved with the spectral response of the spectrometer to produce an initial estimate of the shape of the radiance ratio ($R_{\Delta\lambda}$) as calculated from the spectrometer data. The estimate is then iteratively adjusted by changing the water vapor content in the atmospheric propagation model until an adequate match to the observed spectra is obtained. This process can be repeated for every pixel in the scene. The resultant data can be corrected for path length effects and converted to the amount of water in the vertical atmospheric column over the pixel, based on the atmospheric model chosen. This simplified treatment is valid for nonvegetated regions where no liquid water is present. A more rigorous treatment is required over vegetation and where path-scattered radiance is significant [cf. Gao and Goetz (1990)]. Schlaepfer et al. (1998) introduced the atmospheric precorrected differential absorption (APDA) technique, which specifically included the upwelled radiance in the radiance ratio calculations of Eq. (7.104). Once again, a radiative transfer code is used to predict the behavior in three bands centered on and immediately adjacent to the 940 nm water absorption feature. The predicted behavior is expressed as a ratio of the radiance in the band centered on the feature divided by the radiance that would be interpolated at the absorption feature based on the radiances from the two bands on either wing of the feature (cf. Fig. 7.31). The band ratio predicted by the model is adjusted by changing the

column water vapor amount input to the model until a predicted band ratio and the observed ratio match. That amount of column water vapor is then used in the model to predict the behavior at all other wavelengths of interest. These methods will work best when applied to spectral regions where water vapor is the dominant unknown impacting the absorption and scattering in the RT process. However, in many bands, the number of aerosol molecules also represents a major unknown input to the model limiting the utility of the algorithms mentioned thus far.

To overcome these limitations, Green et al. (1993) suggested a related approach for using imaging spectrometer data to solve for the key unknown inputs to an RT model on a per pixel basis. Because this approach attempts to solve sequentially for most of the key parameters relevant to the reflective band inputs to RT models, we will consider it in more detail. The specifics of the implementation presented here are adapted from an implementation of Green's approach described in Sanders et al. (2001). The problem is simplified by reducing the unknown inputs to the RT process into three key parameters; the pressure-depth, the column water vapor depth, and the aerosol density characterized by the visibility parameter in MODTRAN. The other parameters describing the RT code inputs are assumed known. Most of these other inputs are indeed well known (i.e., time of day, day of year, sensor location, target location, sensor spectral response function, nominal atmospheric profile etc.). One of the big unknowns not fully resolved by this approach is the type of aerosol. This must be estimated from geographic location and prevailing wind conditions. Green's approach assumes that by controlling all the variables known at each point and carefully selecting the spectral region or feature to analyze, the controlling parameters can be found independently from each other.

The first parameter considered is the pressure-depth, which represents the total column of atmosphere between the sun and the sensor (i.e., both legs of the L-shaped path) that will fix the absorption by the well-mixed gases, as well as the Rayleigh molecular scattering. Green suggests using the 760 nm oxygen absorption feature to characterize the pressure depth. Since oxygen is a well-mixed gas, the depth of this narrow feature should be a function of the total optical path. The approach assumes that the target reflectance over several spectral bands (10 nm bandwidth in this case) in an imaging spectrometer can be expressed as

$$r(\lambda) = a_o \lambda + b_o \tag{7.105}$$

where a_o and b_o are unknown parameters that linearly relate reflectance and wavelength on a per pixel basis over the narrow spectral range ($\cong$ 50 nm) around the 760 nm absorption feature. With the other input parameters (e.g., aerosol type, aerosol number density and column water vapor) to an RT code (MODTRAN was used) set to best initial estimates, the remaining free parameter impacting the five-spectral-band region about the 760 nm oxygen feature is the target or ground elevation H_o. Thus, the spectral radiance in the five bands centered on the 760 nm feature can be expressed as

$$\hat{L}_i = f(a_o, b_o, H_o) \tag{7.106}$$

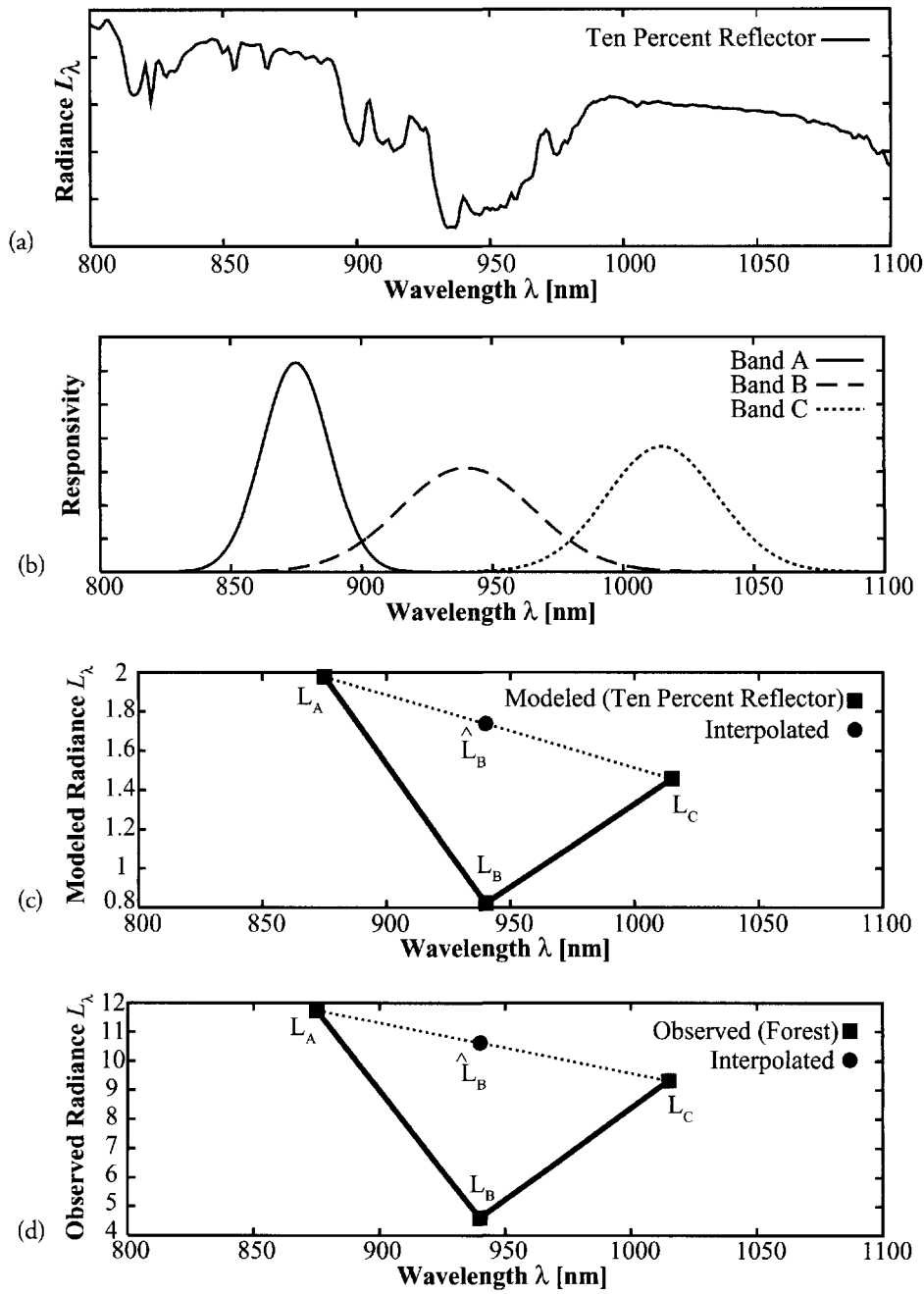


Figure 7.31 Illustration of the use of the radiance observed in three spectral bands near the 940 nm water vapor absorption feature to compute the total column water vapor. (a) The modeled sensor reaching spectral radiance for a given reflectance level and column water vapor amount. (b) Example spectral response functions for bands A, B, and C. (c) Effective spectral radiance expected in bands A, B, and C and the interpolated estimate of the spectral radiance in band B based on interpolation of L_A and L_C to yield $\hat{L}_B$. (d) Observed value of L_A, L_B, and L_C, and the interpolated value of $\hat{L}_B$. Note that the actual magnitude of L_A, L_B, and L_C in (d) and the shape of the interpolation line are a function of the mean magnitude and local slope of the spectral reflectance in the local spectral region. However, the ratio of $\hat{L}_B/L_B$ should be independent of $r(\lambda)$ to first order if $r(\lambda)$ is linear with λ over this region.

where $\hat{L}_i$ is the radiance in the ith spectral band predicted by MODTRAN (including the sensor spectral response) and $f(a_o,b_o,H_o)$ indicates the functional dependence of the predicted value on the input parameters a_o, b_o, and H_o. The actual radiance observed can then be compared to the modeled value and a squared error metric used to define the goodness of fit, i.e.,

$$e_o = \sum (L_i - \hat{L}_i)^2 \tag{7.107}$$

where L_i is the sensed radiance and e_o is the squared error and the sum in over the small number of bands centered on the absorption feature (cf. Fig. 7.32). An iterative solution is then used to solve for the three unknown parameters based on minimizing the squared error. Green suggests using a lookup table (LUT) approach where a three-dimensional lookup table is constructed by running MODTRAN over the expected range of input variables (a_o, b_o, and H_o). The entry in each LUT is the five-band modeled spectral radiance vector. The *nonlinear least-squared spectral fit* (*NLLSSF*) then uses a downhill simplex search method to find the input parameters to the LUT that minimize the squared error for any observed radiance vector [cf. Press (1986)].

In practice, Green et al. (1993) suggest that averaging several pixels around each pixel of interest may reduce the noise and improve the run time for this approach, which yields a map of the apparent surface elevation for use in subsequent steps in the analysis. Experience indicates that this method is not particularly sensitive and that, in general, if topographic relief information is available, it may serve as a more appropriate input to the process [cf. Sanders et al. (2001)].

Given a solution for H_o for each pixel in the image of interest, Green et al. (1993) suggest that we can next solve for the column water vapor appropriate for each pixel using a very similar approach. In this case, the 940 nm water vapor feature is used and the reflectance is once again parameterized in this case according to

$$r(\lambda) = a_1\lambda + b_1 + c_1[r_{H_2O}(\lambda)] \tag{7.108}$$

where a_1, b_1, and c_1 are unknown pixel specific parameters to be solved for and $r_{H_20}(\lambda)$ is a reflectance curve that incorporates the liquid water absorption feature commonly found in vegetation that causes the reflectance of many materials to deviate from approximately linear behavior near the 940 nm atmospheric water vapor feature (cf. Fig. 7.32). The value of c_1 can be thought of as a squelching parameter that suppresses the impact of the liquid water feature on the final reflectance (i.e., as c_i goes to zero, the impact is reduced). A LUT is again constructed, in this case with four variables, a_1, b_1, c_1 and the column water vapor amount input to the model. The entry in the LUT is the radiance vector for the bands around the 940nm feature, and the optimized solution over the four input values is found using the same approach describe above for the pressure-depth. The variable of interest is the column water vapor amount, which can be mapped over the entire image (once again, pixel averaging can be used to reduce noise and decrease run times if required).

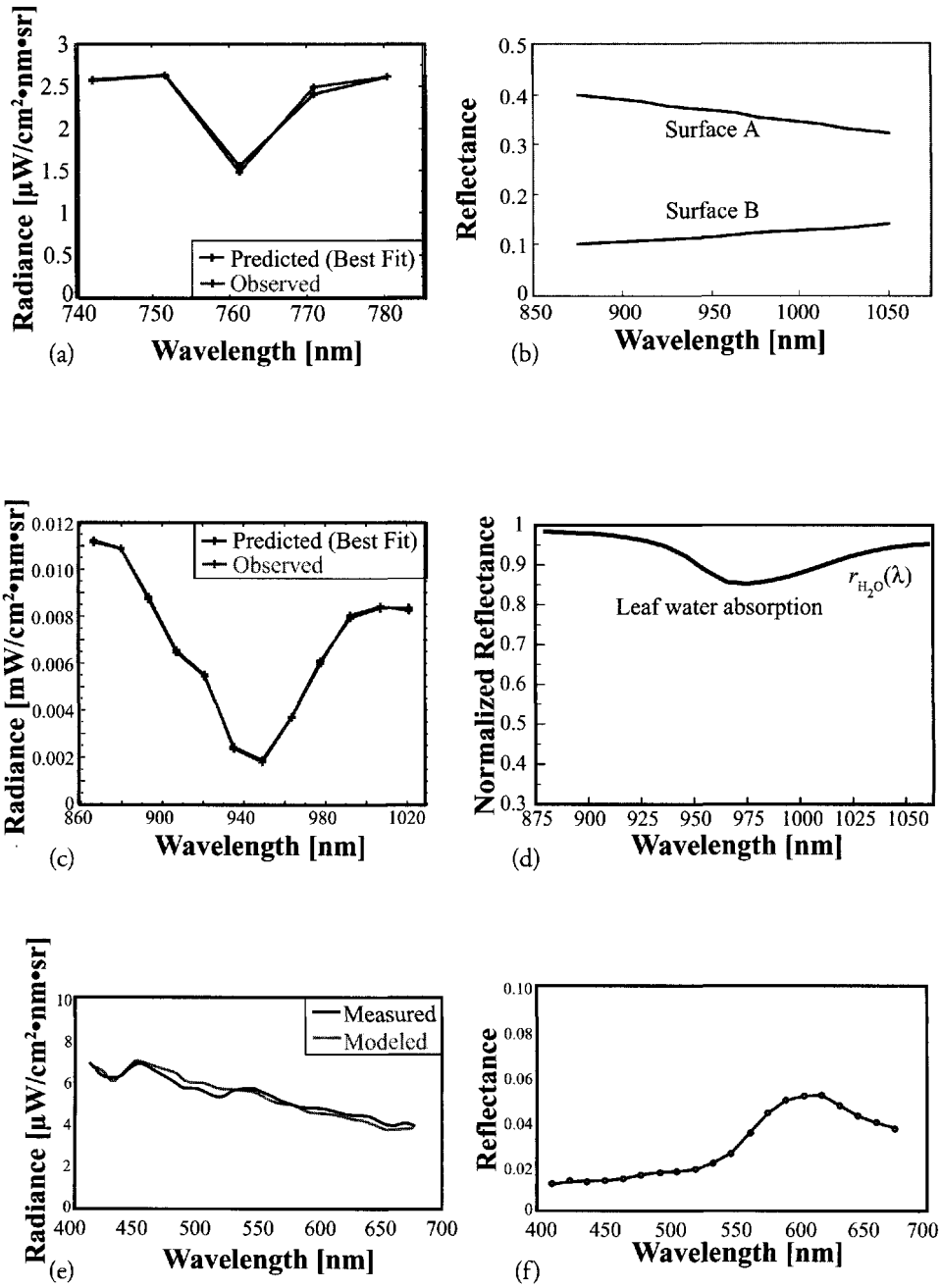


Figure 7.32 Illustration of steps in RT-based atmospheric correction process. (a) Modeled and observed radiance values near the 760 oxygen absorption feature. (b) Example reflectance functions used as inputs to the RT model used to optimize the solution shown in (a). (c) Modeled and observed radiances near the 940 nm water vapor feature. (d) $r_{H_2O}(\lambda)$ curve used with Eq. (7.108) as input to the RT model used to optimize the fit in (c). (e) Modeled and observed spectral radiance in the visible region used to solve for the visibility (aerosol loading) parameter. (f) The $r_{veg}(\lambda)$ reflectance curve used with Eq. (7.109) to optimize the fit in (e). (Data courtesy of NASA JPL.)

The final variable required is aerosol loading in the form of the visibility input to MODTRAN. Having fixed the pressure-depth and water vapor amounts, a third LUT can be generated and searched for optimum parameters. In this case, no specific absorption feature is available, so the broad spectral behavior of aerosol absorption and scattering across the entire visible region is used. The reflectance in this case is described as

$$r(\lambda) = a_2\lambda + b_2 + c_2[r_{veg}(\lambda)] \tag{7.109}$$

where $r_{veg}(\lambda)$ is a reflectance curve that incorporates the blue and red chlorophyll absorption features that give vegetation its green appearance in the visible. The LUT in this case has four inputs; a_2, b_2, c_2 and visibility. The entries are the modeled radiance vectors for spectral bands across the visible spectral region (400-600 nm). The search optimization is the same as above, with the visibility the desired output. Because visibility is not expected to vary rapidly spatially, pixels are often averaged up to hundreds of meters on a side to improve the algorithm performance with the visibility parameter and then interpolated to yield per pixel values for the final step.

This final step is to compute the atmospheric parameters needed to invert radiance to reflectance using MODTRAN on a per pixel basis (i.e., using the values for H_o, column water vapor, and visibility). Figure 7.33 shows three bands of an AVIRIS image of Jasper Ridge and the vertical column water vapor amount derived by this method. Note in particular how the column water vapor is inversely correlated with topography due to the longer path through the atmosphere in the central portion of the image compared to the ridge along the left side of the image. Goetz (1993) points out the importance of observing this type of spatial variation in the atmosphere in terms of its impact on the apparent reflectance observed by Landsat TM sensors. He points out, for example, that in TM band 5 (the TM band most susceptible to water vapor variation), retrieved reflectance values could be in error by as much as 15% of the actual value if variations in water vapor content across a scene are not accounted for.

7.6.2 Spectral Polishing

The spectral atmospheric compensation techniques that employ RT models often have a characteristic residual error resulting from systematic errors in the inversion process. These errors may be due to consistently poor estimates of key variables, limitations in the RT models or most often errors in our knowledge of the sensor's spectral response. When ground truth is available, this residual error can be characterized by taking the ratio of a true reflectance spectrum to a retrieved spectrum for the same target. Ideally, the ratio would be unity in all bands. To the extent there is some random error in the process, one would hope that the average of these ratioed spectra for several samples would yield a spectrum that was unity in all bands. In practice, the residual error characterized in this fashion often has a repeatable systematic structure such that the averaged residual ratio spectrum has spectral character that is very similar to any sample residual ratio spectrum. When

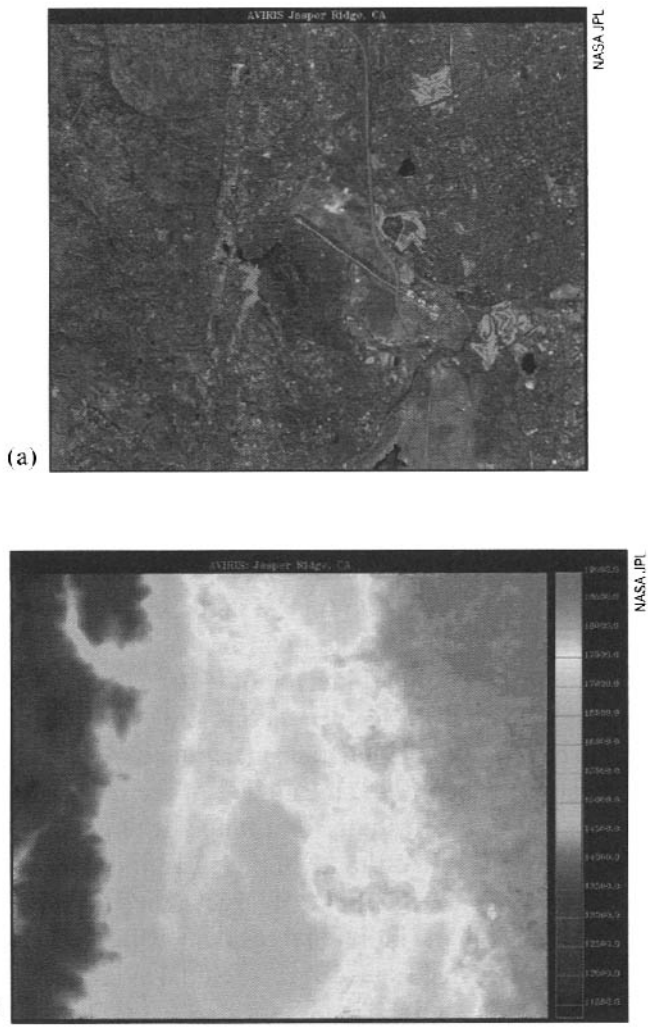


Figure 7.33 (a) AVIRIS image of Jasper Ridge, CA, and (b) a map of column water vapor content derived from the 940 nm absorption line data in the AVIRIS spectra. See color plate 7.33.

many ground truth samples of this sort are available, a simple correction can be achieved by multiplying each retrieved spectrum by the average of the residual ratio spectra, i.e.,

$$\hat{r}_i = r_i C_i \tag{7.110}$$

where $\hat{r}_i$ is the corrected (polished) reflectance in the ith band, r_i is the initial reflectance retrieved for the ith band and C_i the correction factor for the ith band calculated from

$$C_i = \left[\sum_{j=1}^{N} \frac{r_{truthij}}{r_{ij}} \right] / N \tag{7.111}$$

where $r_{truthij}$ is the truth value of reflectance for the jth sample in the ith band, r_{ij} is the corresponding initially retrieved reflectance for the jth sample in the ith band, and N is the number of samples for which good ground truth values are available. This process is known as polishing because it smoothes out the systematic bumps in retrieved spectra. In practice, alternative methods to generate a polishing vector (i.e., a vector composed of C_i values) must be found because of the lack of significant numbers of truth samples. In the extreme case, Eq. (7.111) can be used with a sample size of one if a single target of known reflectance can be identified. To reduce errors, many pixels over the sample may be averaged to identify the retrieved reflectance spectrum and a target with little spectral character chosen to minimize any artifacts due to the target's spectral shape [cf. Green et al. (1996)].

In many cases, no ground truth spectra are available. Yet, we still need to polish the retrieved reflectance spectra to avoid algorithms being triggered by compensation artifacts rather than true spectral signatures. One way to polish the spectra is to generate a spectrally smoothed version of the retrieved spectrum using a cubic spline interpolation to estimate the reflectance in each band based on the spectral value in a band along with that of its neighboring bands [cf. Gao and Davis (1998)]. The degree of smoothing can be controlled by the spline-fitting parameters. The smooth spectrum will have both artifacts and real localized spectral anomalies removed (cf. Fig. 7.34). By taking the ratio of the smoothed to the original reflectance spectra, we can generate an estimate of the correction factor C_i for each band and combine them into an estimate of the residual vector. Residual vectors can be computed in this fashion for a range of targets in a scene and then averaged to generate a composite estimate of the polishing vector to be applied using Eq. (7.110). By selecting a wide range of targets, the averaging process should reduce any spectral features in the individual residual vectors due to localized spectral features in the true reflectance spectrum. The final average residual ratio spectrum should be predominately influenced by spectral features that occurred in all of the samples and that are therefore presumed to be artifacts of the inversion process. When polishing is done using these spectral smoothing techniques, the quality of the results is often a function of the method used to select the pixels to include in the averaging process, and to a lesser extent, the smoothing (e.g., spline) algorithm used. Choi (2002) discusses several approaches for selection of pixels to include in the averaging process.

Our discussion to this point has focused on spectral polishing of reflective data as a final step in the atmospheric compensation process. A related approach described by Borel (2003) has been developed to support atmospheric compensation and temperature emissivity separation of thermal infrared spectral data (cf. Chap. 11).

7.7 SUMMARY OF ATMOSPHERIC COMPENSATION ISSUES

We have presented a number of approaches to atmospheric compensation in this section. They reflect the wide range of tools being used in different spectral bands, for different sensors, and for differing user requirements. The large number of

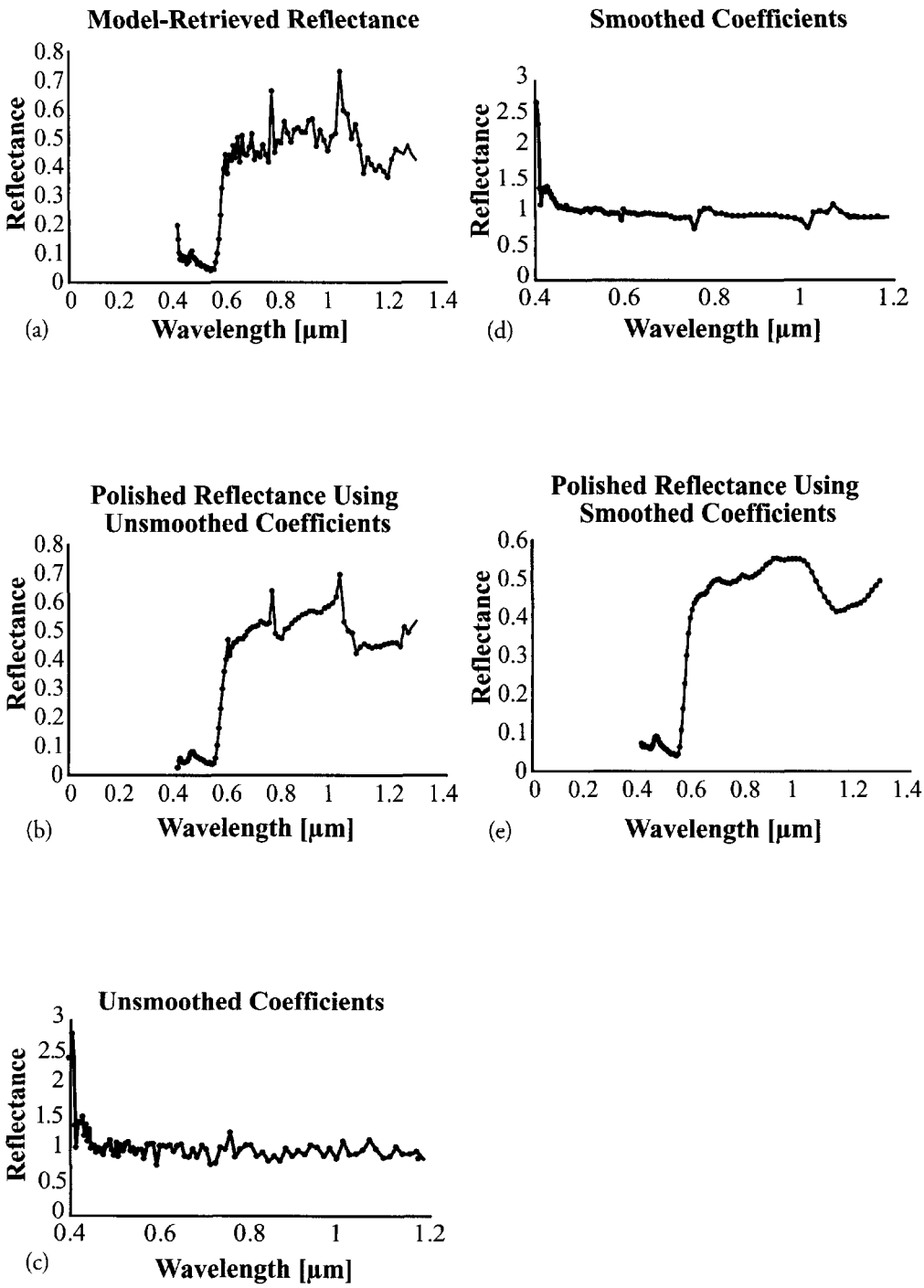


Figure 7.34 Examples of the use of cubic spline fitting to generate residual ratio vectors for use in spectral polishing. (a) Original retrieved reflectance spectra. (b) Smoothed version of (a) resulting from cubic spline smoothing. (c) Ratio of curve (b) to curve (a). (d) Average of many samples of individual pixel residual ratios [i.e. average of many curves shown in (c)], (e) Corrected version of curve shown in (a) using the average correction factor shown in (d).

methods is probably most indicative of the fact that well-accepted operational compensation methods applicable to a broad range of sensors do not yet exist. Much ongoing work is directed at trying to develop more accurate and more robust compensation methods. Some of the most interesting work involves the use of spectrometer data to help characterize the atmosphere (cf. Sec. 7.6). Not all of this work involves imaging systems. For example, infrared microwave sounders are being used from space to retrieve atmospheric temperature profiles and total column absorber numbers of certain gases. If these sounders are flown in formation with imaging satellites, they can provide input to atmospheric propagation models. Li et al. (2005) estimate that the performance specifications for atmospheric retrieval of temperature and water vapor on the next generation of sounders may be sufficient for retrieval of layer temperatures to within 1 K with 1 km vertical resolution and relative humidity to 10 % with 2 km vertical resolution. These types of "hardware" solutions (e.g., flying satellite sensors in formation) are very exciting for the long run, but they are not yet readily available. In the interim, many users are looking to hybrid solutions that combine aspects of the various approaches described in this chapter. A good example of this is discussed by Moran et al. (1993), who show that the in-scene method of dark-object selection for estimating upwelled radiance when coupled with atmospheric propagation models can potentially provide an operational calibration in the reflective region. These results from a limited study under clear atmosphere conditions indicate that reflectance errors of 2 reflectance units (0.02) may be achievable with this type of hybrid method.

A great deal of ongoing work aimed at atmospheric compensation of imaging spectrometer data involves improved methods for taking advantage of advanced RT models in the compensating algorithms. Some of these advanced methods are discussed further in the model-based algorithms included in Chapter 11.

In summary, a number of potential methods for atmospheric compensation exist. However, in general the burden is still on the user to determine the accuracy (e.g., temperature or reflectance errors) required for the specific task at hand and then to determine which of the possible calibration methods can meet the requirements.

7.8 REFERENCES

Adler-Golden, S.M., Levine, R.Y., Berk, A., Bernstein, L.S., Anderson, G.P., & Pukall, B. (1999). Detection of cirrus clouds at 1.13 μm in AVIRIS scenes over land. *Proceedings SPIE*, Vol. 3756, pp. 368-373.

Anding, D.C., & Walker, J.P. (1975). Use of Skylab EREP Data in a Sea Surface Temperature Experiment. Science Applications, Ann Arbor, MI, NASA-CR-144479.

Barsi, J.A., Barker, J.L., & Schott, J.R. (2003). "An atmospheric correction parameter calculator for a single thermal band earth-sensing instrument." Proceedings of the IEEE, IGARSS 2003, Vol. 5, pp. 3014-3016.

Becker, F., Ramanantsizehena, P., & Stoll, M. (1985). Angular variation of the bidirectional reflectance of bare soils in the thermal infrared band. *Appl. Opt.*, Vol. 24, No. 3, pp. 366-375.

Bell, E.F., Eisner, L., Young, J., & Oetjen, R.A. (1960). Spectral radiance of sky and terrain at wavelengths between 1 and 20 microns II, sky measurements. *J. Opt. Soc. Am.*, Vol. 50, No. 12, pp. 1313-1320.

Berk, A., Bernstein, L.S., & Robertson, D.C. (1989). MODTRAN: a moderate resolution model for LOWTRAN 7. GL-TR-89-0122, Spectral Sciences, Boston MA.

Berk, A., Bernstein, L.S., Anderson, G.P., Acharya, P.K., Robertson, D.C., Chetwynd, J.H., & Adler-Golden, S.M. (1998). MODTRAN cloud and multiple scattering upgrades with application to AVIRIS. *Remote Sensing Environ.*, Vol. 65, pp. 367-375.

Borel, C.C. (2003). "ARTEMISS—an algorithm to retrieve temperature and emissivity from hyper-spectral thermal image data." 28th Annual GOMACTech Conference, Hyperspectral Imaging Session, Tampa, FL.

Borel, C.C., Clodius, W.B., Davis, A.B., Smith, B.W., Szymanski, J.J., Theiler, J.P., Villeneuve, P.V., & Weber, P.G. (1999). MTI core science retrieval algorithms. *Proceedings SPIE*, Vol. 3753, pp. 403-415.

Byrnes, A.E., & Schott, J.R. (1986). Correction of thermal imagery for atmospheric effects using aircraft measurement and atmospheric modeling techniques. *App. Opt.*, Vol. 25, No. 15.

Chavez, P.S. (1975). "Atmospheric, solar, and MTF corrections for ERTS digital imagery." Proceedings of the American Society of Photogrammetry, Falls Church, VA., p. 699.

Chedin, H., Scott, N.A., & Berroir, A. (1982). A single channel, double-viewing angle method for sea surface temperature determination from coincident METEOSAT and TIRO-N radiometric measurements. *J. Appl. Meteor. Soc.*, Vol. 21, pp. 613-618.

Choi, H.-G. (2002). "Spectral misregistration correction and simulated for hyperspectral imagery." M.S. Thesis, Center for Imaging Science, Rochester Institute of Technology.

Crippen, R.E. (1986). The regression intersection method of adjusting image data for band ratioing. *Int. J. Remote Sensing*, Vol. 2, pp. 137-155.

Crippen, R.E. (1988). The dangers of underestimating the importance of data adjustments in band ratioing. *Int. J. Remote Sensing*, Vol. 9, No. 4, pp. 762-776.

Fenn, R.W., Clough, S.A., Gallery, W.O., Good, R.E., Kneizys, F.X., Mell, J.D., Rothman, L.S., Shettle, E.P., & Volz, F.E. (1985). Optical properties of the atmosphere. In *Handbook of Geophysics and the Space Environment*, U.S. Air Force Geophysics Laboratory, Bedford, MA, NTIS ADA 167000.

Gao, B., & Davis, C.O. (1998). "A new and fast method for smoothing spectral imaging data." Proceeding of the 19098 Airborne Geoscience Workshop, Jet Propulsion Laboratory, Pasadena, CA JPL Pub. 97-21.

Gao, B.C., & Goetz, A.F.H. (1990). Column atmospheric water vapor and vegetation liquid water retrievals from airborne imaging spectrometer data. *J. Geophys. Res. - Atmos.*, Vol. 95, pp. 3549-3564.

Gao, B.C., Davis, C.O., & Kaufman, Y.J. (1997). Thin cirrus detection and correction of thing cirrus path radiances using near-IR channels near 1.375μm. *Proceedings SPIE*, Vol. 3122, pp. 78-87.

Goetz, A.F.H. (1993). "Effects of water vapor and cirrus clouds on TM-derived apparent reflectance-based on AVIRIS experience." Proceeding of the Workshop on Atmospheric Correction of Landsat Imagery, Torrance, CA, sponsored by the Defense Landsat Program Office, pp. 55-60.

Gonzalez, R.C., & Wintz, P. (1987). *Digital Image Processing*, Addison-Wesley, Reading, MA.

Gordon, H.R., Brown, D.K., Brown, J.W., Evans, O.B., Broenkow, R.H.B., & Broenkow, W.W. (1983). Phytoplankton pigment concentrations in the Middle Atlantic Bight: comparison of ship determinations and CZCS estimates. *Appl. Opt.*, Vol. 22, pp. 20-36.

Green, R.O., Conel, J.E., & Roberts, D.A. (1993). Estimation of aerosol optical depth, pressure elevation, water vapor and calculation of apparent surface reflectance from radiance measured by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) using a radiative transfer code. *Proceedings SPIE*, Vol. 1937, pp. 2-11.

Green, R.O., Roberts, D.A., & Conel, J.E. (1996). "Characterization and compensation of the atmosphere for the inversion of AVIRIS calibrated radiance to apparent surface reflectance." Proceedings of the AVIRIS Geoscience Workshop, Jet Propulsion Laboratory, Pasadena, CA, JPL Pub. 96-4.

Gubareff, G.G., et al. (1960). *Thermal Infrared Radiation Properties Survey: A Review of the Literature*, 2nd ed. Honeywell Research Center, Minneapolis-Honeywell Regulator Company, Minneapolis, MN.

Hall, F.G., Strebel, D.E., Nickeson, J.E., & Goetz, S.J. (1991). Radiometric rectification: toward a common radiometric response among multidate, multisensor images. *Remote Sensing Environ.*, Vol. 35, pp. 11-27.

Hanel, R.A., Conrath, B.J., Jennings, D.E., & Samuelson, R.E. (1992). *Exploration of the Solar System by Infrared Remote Sensing*. Cambridge University Press, NY.

Heilman, J.L., & Moore, D.G. (1980). Thermography for estimating near surface soil moisture under developing crop canopies. *J. Appl. Meteorol.*, Vol. 9, pp. 324-328.

Herman, B.M., Caudill, T.R., Flittner, D.E., Thome, K.J., & Ben-David, A. (1995). Comparison of the Gauss-Seidel spherical polarized radiative transfer code with other radiative transfer codes. *J. App. Opt.*, Vol. 34, No. 21, pp4563-4572.

Isaacs, R.G., Wang, W.-C., Warsham, R.D., & Goldenberg, S. (1987) Multiple scattering LOWTRAN and FASCODE models. *J. App. Opt.*, Vol. 26, No. 7. pp. 1272-1281.

Jensen, J.R., ed. (1983). Urban/suburban land use analysis. In Colwell, R.N., ed. *Manual of Remote Sensing*, Vol. 2, 2nd ed., American Society of Photogrammetry, Falls Church, VA, pp. 1571-1666.

Johnson, R.W., Bahn, G.S., & Thomas, J.P. (1981). Synoptic thermal and oceanographic parameter distributions in the New York bight apex. *Photogrammetric Engin. and Remote Sensing*, Vol. 48, No. 11, pp. 1593-1598.

Kneizys, F.X., Shettle, E. P., Gallery, W.O., Chetwynd, J.H., Jr., Abreu, L.W., Selby, J.E.A., Clough, S.A., & Fenn, R.W. (1983). "Atmospheric transmittance/radiance: computer code LOWTRAN 6." Optical Physics Division, Air Force Geophysics Laboratory, Air Forces Systems Command, Hanscom AFB, MA.

Kneizys, F.X., Shettle, E.P., Abreu, L.W., Chetwynd, J.H., Anderson, G.P., Gallery, W.O., Selby, J.E.A., & Clough, S.A. (1988). "Users guide to LOWTRAN 7," AFGL-TR-88-0177, Environmental Research Papers, No. 1010, Optical/ Infrared Technology Division, Air Force Geophysics Laboratory, Hanscom AFB, MA.

Legeckis, R. (1978). A survey of worldwide sea surface temperature fronts detected by environmental satellites. *J.of Geophysi. Res.*, Vol. 83 (C9), pp. 4501-4522.

Li, J., Wang, F., Schimt, T., Menzel, W.P., & Gurka, J. (2005). Vertical Resolution Study on the GOES-R Hyperspectral Enviromental Suite(HES). *Proceedings SPIE,* Vol. 5685, pp. 57-65.

Lowe, D.S. (1978). "Effects of emissivity on airborne observation of roof temperature." Proceedings of Thermosense 1, sponsored by the American Society of Photogrammetry.

Marsh, S.E., & Lyon, R.J.P. (1980). Quantitative relationship of near-surface spectra to Landsat radiometric data. *Remote Sensing Environ.*, Vol. 10, No. 4, pp. 241-261.

Martonchik, J.W., Diner, D.J., Crean, K.A., & Bull, M.A. (2002). Regional aerosol retrieval results from MISR. *IEEE Trans. Geoscie. Remote Sensing*, Vol. 40, pp. 1520-1531.

McMillan, L.M. (1975). Estimation of sea surface temperatures from two infrared window measurements with different absorptions. *J. Geophysical Res.*, Vol. 80, pp. 5113-5117.

Mericsko, R. (1993). "Enhancements to atmospheric-correction techniques for multiple thermal images." Proceedings of the Workshop on Atmospheric Correction of Landsat Imagery, sponsored by Defense Landsat Office, Torrance, CA.

Moran, M.S., Jackson, R.D., Slater, P.N., & Teillet, P.M. (1993). "Simplified procedures for retrieval of surface reflectance factors from Landsat TM sensor output." Proceeding of the Workshop on Atmospheric Correction of Landsat Imagery, Torrance, CA, sponsored by the Defense Landsat Program Office, pp. 85-90.

Piech, K.R., & Walker, J.E. (1974). Interpretation of soils. *Photogrammetric Engin. Remote Sensing*, Vol. 40, pp. 87-94.

Piech, K.R., Schott, J.R., & Stewart, K.M. (1978). The blue-to-green reflectance ratio and lake water quality. *Photogrammetric Engin. Remote Sensing*, Vol. 44, No. 10, pp. 1303-1319.

Piech, K.R. (1980). Material identification using broad band visible data. *Proceedings SPIE,* Vol. 238.

Prabhakara, C., Dalu, G., & Kunde, V.G. (1974). Estimation of sea surface temperature from remote sensing in the 11 to 13 μm window region. *J. Geophys., Res.*, Vol. 79, No. 33, pp. 5039-5044.

Press, W.H. (1986). *The Art of Scientific Computing: Numerical Recipes*. Cambridge University Press, Cambridge.

Price, J.C. (1984). Land surface temperature measurements from the split window channels of the NOAA 7 advanced very high resolution radiometer. *J. Geophys. Res.*, Vol. 89, No. D5, pp. 7231-7237.

Rothman, L.S., Barbe, A., Benner, C.D., Brown, L.R., Camy-Peyret, C., Carleer, M.R., Chance, K., Clerbaux, C., Dana, V., Devi, V.M., Fayt, A., Flaud, J.M., Gamache, R.R., Goldman, A., Jacquemart, D., Jucks, K.W., Lafferty, W.J., Mandin, J.Y., Massie, S.T., Nemtchinov, V., Newnham, D.A., Perrin, A., Rinsland, C.P., Schroeder, J., Smith, K.M., Smith, M.A.H., Tang, K., Toth, R.A., Vander Auwera, J., Varanasi, P., & Yoshino, K. (2003). The HITRAN molecular spectroscopic database: edition of 2000 including updates through 2001. *J. Quanti. Spectrosc. & Radiative Transfer*, Vol. 82, pp. 5-44.

Rouse, J.W., Haas, R.H., Schell, J.A., & Deering, D.W. (1973). "Monitoring vegetation systems in the Great Plains with Third ERTS." ERTS Symposium, NASA No. SP-351, pp. 309-317.

Salisbury, J.W., & D'Arian, D.M. (1992). Emissivity of terrestrial material in the 8-14 μm atmospheric window. *Remote Sensing Environ.*, Vol. 42, pp. 83-106.

Salvaggio, C. (1993). "Radiometric scene normalization utilizing statistically invariant features." Proceeding of the Workshop on Atmospheric Correction of Landsat Imagery, Torrance, CA, sponsored by the Defense Landsat Program Office, pp. 155-160.

Sanders, L.C., Schott, J.R., & Raqueño, R.V. (2001). A VNIR/SWIR atmospheric correction algorithm for hyperspectral imagery with adjacency effect. *Remote Sensing Environ.*, Vol. 78, No. 3, pp. 252-263.

Saunders, P.M. (1967). Aerial measurement of sea surface temperature in the infrared. *J. Geophys. Res.*, Vol. 72, pp. 4109-4115.

Scarpace, F.L., Madding, R.P., & Green, T., III (1974). "Scanning thermal plumes." Ninth International Symposium on Remote Sensing of Environment.

Schlaepfer, D., Borel, C.C., Keller, J., & Itten, K. (1998). Atmospheric pre-corrected differential absorption technique to retrieve columnar water vapor. *Remote Sensing Environ.*, Vol. 65, No. 3, pp. 353-366.

Schott, J.R. (1979). Temperature measurement of cooling water discharged from power plants. *Photogrammetric Engin. Remote Sensing*, Vol. 45, No. 6, pp. 753-761.

Schott, J.R. (1986). Incorporation of angular emissivity effects in long wave infrared image models. *Proceedings SPIE*, XII, Vol. 685, pp. 44-52.

Schott, J.R. (1993). Methods for estimation of and correction for atmospheric effects on remotely sensed data. *SPIE*, Vol. 1968, No. 51, pp. 448-482.

Schott, J.R., & Schimminger, E.W. (1981). "Data use investigations for applications explorer mission A (heat capacity mapping mission)." Calspan Report No. 6175-M-1, NASA Accession No. E81-10079.

Schott, J.R., & Wilkinson, E.P. (1982). Quantitative methods in aerial thermography. *Opt. Eng.*, Vol. 21, No. 5. pp. 864-867

Schott, J.R., Biegel, J.D., & Volchok, W.J. (1985). "Comparison of methods for removal of atmospheric effects from remotely sensed color images." Presented at SPSE's 25th Fall Symposium – Imaging '85, Arlington, VA.

Schott, J.R., Salvaggio, C., & Volchok, W.J. (1988). Radiometric scene normalization using pseudoinvariant features. *Remote Sensing Environ.*, Vol. 26, No. 1, pp. 1-16.

Shettle, E.P., & Fenn, R.W. (1979). "Models for the aerosol of the lower atmosphere and the effect of humidity variations on their optical properties." AFGL-TR 790214, Opt. Phys. Div., Air Force Geophysics Laboratory, Hanscom Air Force Base, MA.

Slater, P.N. (1980). *Remote Sensing: Optics and Optical Systems*. Addison-Wesley, Reading, MA.

Slater, P.N., Biggar, S.F., Holm, R.G., Jackson, R.D., Mao, Y., Moran, M.S., Palmer, J., & Yuan, B. (1987). Reflectance- and radiance-based methods for the in-flight absolute calibration of multispectral sensors. *Remote Sensing Environ.*, Vol. 22, pp. 11-37.

Smith, H.J.P., Dube, D.J., Gardner, M.E., Clough, S.A., Kneizys, F.X., & Rothman, L.S. (1978). "FASCODE- Fast Atmospheric Signature Code (Spectral Transmittance and Radiance)." Air Force Geophysics Laboratory Technical Report AFGL-TR-78-0081, Hanscom Air Force Base, MA.

Stamnes, K., Tsay, S.-C., Wiscombe, W., & Jayaweera, K. (1988). Numerically stable algorithm for discrete-ordinate-method radiative transfer in multiple scattering and emitting layered media. *J. App. Opt.*, Vol. 27, pp. 2502-2509.

Suits, G.H. (1972). The calculation of the directional reflectance of a vegetative canopy. *Remote Sensing Environ.*, Vol. 2, pp. 117-125.

Switzer, P., Kowalik, W.S., & Lyon, R.J.P. (1981). Estimation of atmospheric path-radiance by the covariance matrix method. *Photogrammetric Engin. Remote Sensing,* Vol. 47, No. 10, pp. 1469-1476.

Thomas, J.R. & Gausman, H.W. (1977). Leaf reflectance vs. leaf chlorophyll and carotenoid concentration for light crops. *Agronomy J.*, Vol. 69, pp. 799-802.

Thomas, J.R., Myers, V.I., Heilman, M.D., & Wiegand, C.O. (1966). "Factors affecting light reflectance of cotton." Proceedings 4th Symposium on Remote Sensing of Environment, Ann Arbor, MI, pp. 305-312.

Vermote, E.F., Tanre, D., Deuze, J.L., Herman, M., & Marcette, J.-J. (1997). Second simulation of the satellite signal in the solar spectrum, 6S: an overview. *IEEE Trans Geoscie. Remote Sensing*, Vol. 35, No. 3, pp. 675-687.

Walton, F.F., McClain, E.P., & Sapper, J.F. (1990). "Recent changes in satellite-based multi-channel sea surface temperature algorithms." MTS90, Marine Technology Society, Washington, DC.

Yuan, D., & Elvidge, C.D. (1993). "Application of relative radiometric rectification procedure to Landsat data from use in change detection," Proceedings of the Workshop on Atmospheric Correction of Landsat Imagery, Torrance, CA, sponsored by the Defense Landsat Program Office, pp. 162-166.

This page intentionally left blank

CHAPTER 8
DIGITAL IMAGE PROCESSING PRINCIPLES

The field of *digital image processing* (*DIP*) continues to grow at a very rapid rate. This growth is spurred by improvements in the speed of computers and the consequent reduction in the cost of computing power. When combined with technological advances in detectors and electronics (particularly in analog-to-digital converters), the advance in technology now puts digital images at every user's fingertips. The treatment presented here is a very small sampling of digital image technology aimed at providing the user with a common terminology and a limited exposure to some DIP tools applicable to most monochrome (single band) or color images that are frequently used in remote sensing. A more thorough general treatment of DIP can be found in Gonzalez and Woods (2002), Rosenfeld and Kak (1982), and Pratt (1991), to name only a few. We will explore most of the general image processing tools in the form of examples. The diversity of applications for these tools in remote sensing is almost endless, but generally represents simple logical extensions of the examples presented here adapted to a user's specific needs. The only necessary mathematical background is elementary algebra, other than in Section 8.4, which is based on the principles of basic linear systems that are well treated elsewhere [cf. Gaskill, (1978)]. Consequently, we will not derive the algorithmic expressions but only examine the operational form or, in some cases, the resultant images.

Chapters 9–11 extend these basic DIP techniques and address in greater detail some image processing and analysis techniques more specifically targeted at remotely sensed imagery. The methods and techniques used in multispectral imaging and the analysis of spectrometer data are considered in Chapters 9 and 10,

respectively. Chapter 11 focuses on image analysis methods incorporating physics-based models. The reader should recognize that these divisions are somewhat arbitrary and that many of the algorithms are appropriate in multiple chapters.

For simplicity, in these chapters we assume that all images are in digital form, meaning that they have been sampled into picture elements (pixels) with the brightness of each pixel represented by a numerical value (usually an integer) described as the *gray level* or *digital count* (*DC*) level (cf. Fig. 8.1). A digital image is commonly characterized by the number of pixels in the x and y directions, the number of spectral bands, and the number of bits required to store the range of gray levels for each pixel in each band. For example, a digital image with 512 pixels along the x dimension and 1024 pixels in the y direction, three spectral bands, and 256 gray levels per band would be called a 512 × 1024 × 3-band × 8-bit image. Because many images use 8 bits (256 gray levels) of dynamic range for display, this is often abbreviated to 512 x 1024 x 24 bits, where the division of the data into three 8-bit bands is assumed. Unless otherwise specified, we will assume that multiband images are co-registered such that the information in the (i,j)th pixel location in band 1 corresponds to the same point on the ground as the (i,j)th pixel location in bands 2, 3, etc.

In most cases, the images from EO systems are produced such that each pixel corresponds to one GIFOV of the sensor at nadir. In the case of film systems, the images can be digitized at a range of "resolutions." If the full resolution of the film image is to be preserved, *sampling theory* dictates that two samples be taken per just-resolvable line pair. So a film system that resolves 50 line pairs per millimeter in each direction will require

$$S_x = \frac{50 \text{ line pairs}}{\text{mm}} \cdot \frac{2 \text{ samples}}{\text{line pair}} = 100 \text{ samples/mm} \tag{8.1}$$

For 9-inch (nominal) color film from a mapping camera (actual format size, 230 mm), the number of pixels at this resolution is

$$N = S_x X S_y Y = (100 \text{ samples/mm} \cdot 230\text{mm})^2 = 0.53 \cdot 10^9 \text{ samples} \tag{8.2}$$

where $S_x = S_y$ are the sample rates in the x and y directions, and X and Y are the film format dimensions in the x and y directions. For three 8-bit spectral bands, this represents a total number of bits of

$$V_b = 3_{\text{bands}} \frac{8 \text{ bits}}{\text{band}} \cdot N = 12.7 \cdot 10^9 \text{bits} \tag{8.3}$$

which is more commonly expressed as bytes of data storage:

$$V = \frac{V_b \text{bits}}{8 \text{ bits/byte}} = 1.6 \cdot 10^9 \text{ bytes} \tag{8.4}$$

This is an enormous amount of computer storage and reflects the high resolution and high information density of photographic film. Often film data are digitized at less than full resolution if resolution is not a critical issue in the digital image analysis task. The original film data are still used directly for obtaining spatial

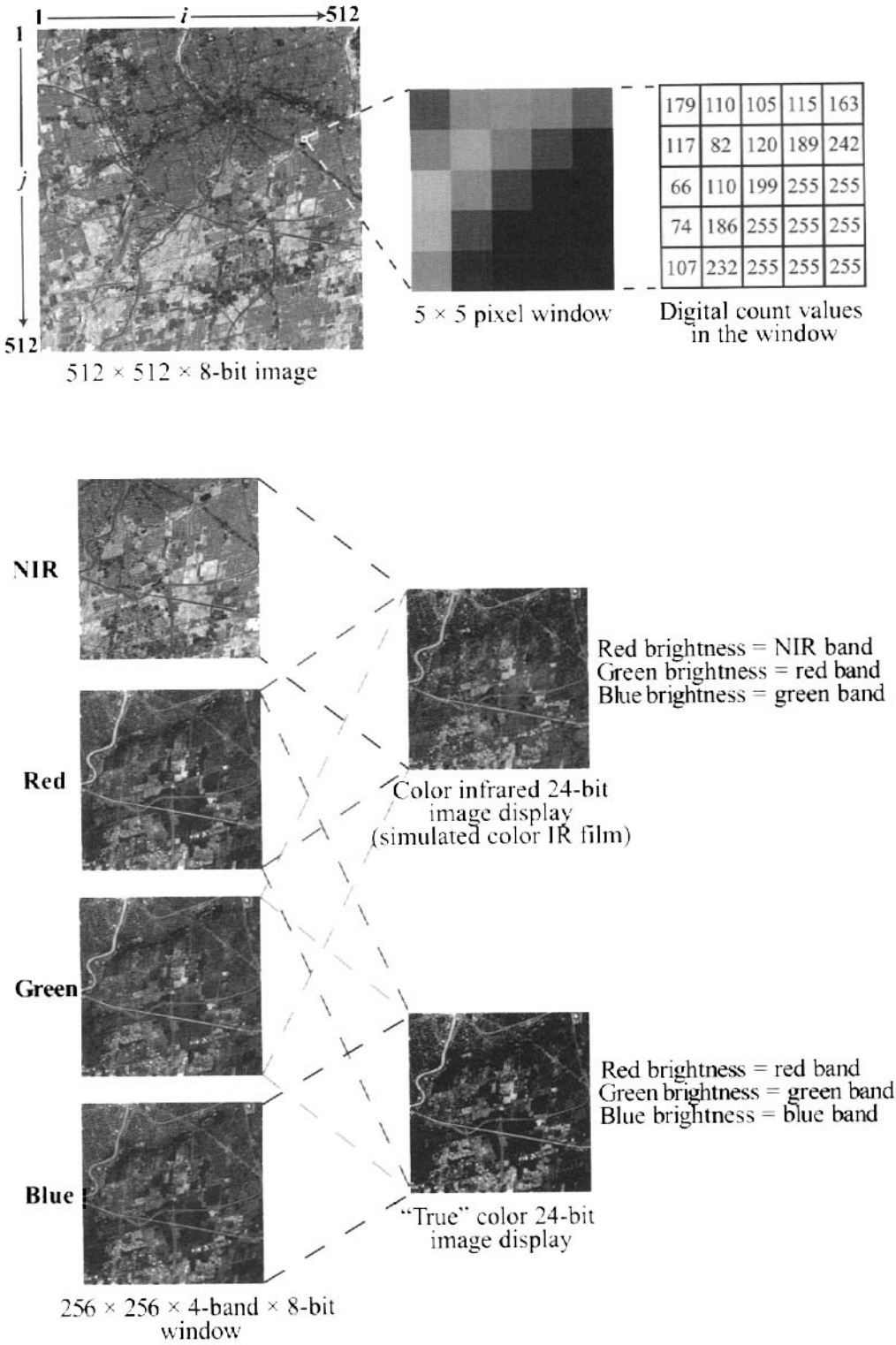


Figure 8.1 Digital image concepts. See color plate 8.1.

detail. EO images can also become very large with a 6000 x 6000 pixel, 7-band, 8-bit Landsat TM image requiring 250 megabytes of data storage. Today's general-purpose workstation computers can handle storage and processing of images of this size. However, processing a full image can still be prohibitively slow. Often a portion of an image (e.g., a 512 × 512 window) or a subsampled version of the image (e.g., every 10th pixel in x and y) is analyzed first to define all processing steps, and then the entire image is analyzed using the "script" developed interactively on the image subset. Image subsets of 512 × 512 to 2048 × 2048 are most commonly used for interactive image processing because soft copy displays (computer monitors) are usually restricted to these sizes.

The speed of today's high-speed computer workstations is such that all the algorithms presented here can be run on fair-size images (e.g., 1024 × 1024) in a matter of minutes. When required, parallel image processing tools are available to expedite processing of large images or implementation of more complex algorithms.

This chapter presents some of the DIP tools commonly used to enhance or exaggerate the appearance of images to facilitate visual image analysis, as well as methods used to preprocess images to improve machine processing approaches. The machine processing approaches considered in later chapters emphasize statistical pattern recognition of spectral patterns. The whole field of spatial pattern recognition is of keen interest to many aspects of remote sensing, but is beyond the scope of this treatment. The interested reader should consider Tou and Gonzalez (1974), Rosenfeld and Kak (1982), and Fu (1974).

8.1 POINT PROCESSING

The simplest and most widely used class of image processing operations is called *point processing*, which is a synonym for gray-level manipulation. It involves adjusting the DC level of pixels without concern for the values of neighboring pixels (i.e., the operation is applied to each individual point rather than to a neighborhood). The operation can be implemented by passing the image through a *lookup table* (*LUT*), which is a mapping of gray values from input to output. The process can be most easily seen in a graphical plot as shown in Figure 8.2; it is represented mathematically by:

$$g(i,j) = \mathrm{LUT}[f(i,j)] \tag{8.5}$$

where $f(i,j)$ and $g(i,j)$ are the DC values of the input image (f) and output image (g) at each location (i,j) and LUT is the lookup table function.

Point processing operations are often viewed in terms of their impact on the histogram of the image. The *histogram* of an image (f) [designated HIS(f)] is simply a one-dimensional vector whose components are the number of pixels in the image having DC = 0, DC = 1,..., DC = DC_{max}. Note that if the components of the histogram vector are divided by the total number of pixels in the image (N), the normalized histogram (HIS′) becomes an estimate of the discrete gray level *probability distribution* for the image, i.e.,

$$\mathrm{HIS}'(f) = \frac{\mathrm{HIS}(f)}{N} \tag{8.6}$$

We recognize that the value of each element in the vector HIS′ is the probability p(DC) of a pixel having the DC (brightness level) corresponding to that entry (bin). We often plot the normalized histograms of the input and output images along with the LUT as depicted in Figure 8.2 to clarify the impact of the LUT on the image. In most cases, we will also plot the identity LUT for reference. The identity operation maps each input DC level back to that same DC level in the output image.

The input image in Figure 8.2 is from an 8-bit system, with fixed gain and offset to ensure that bright objects with high sun angles and dark objects at low sun angles are all accommodated in the same 8-bit dynamic range. As a result, the

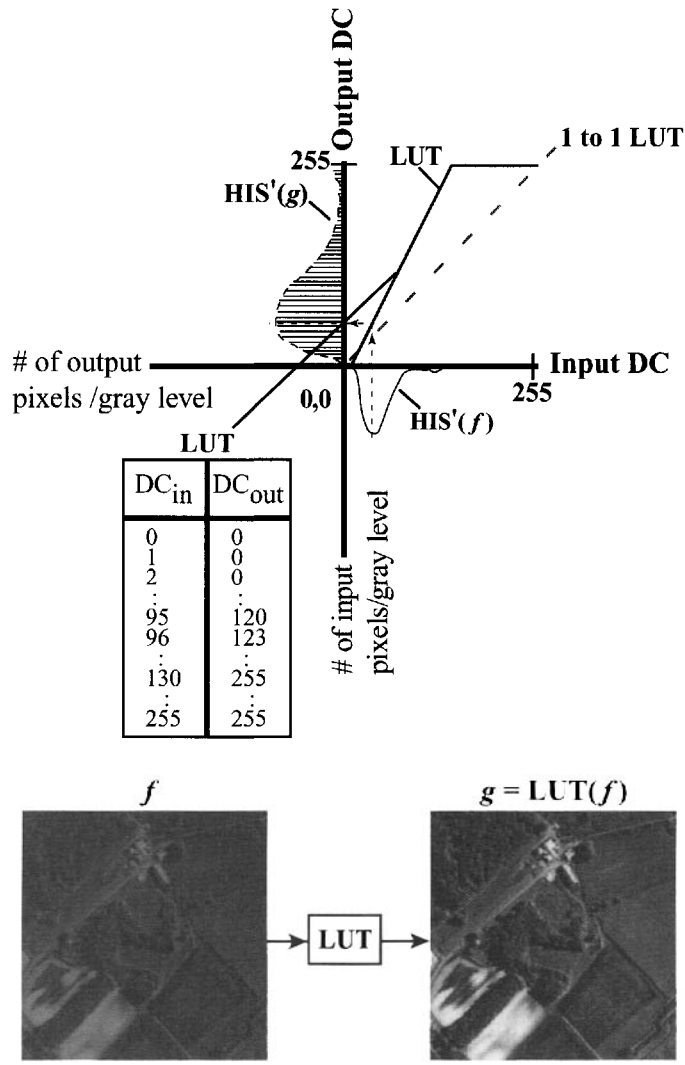


Figure 8.2 Lookup table concepts.

quantization must be somewhat coarse, and any given image will generally occupy only a limited portion of the available dynamic range. The resulting images often have low contrast when initially displayed. By passing the image through a linear contrast enhancing LUT, as depicted in Figure 8.2, the full 8-bit dynamic range of the display can be utilized and the apparent image contrast improved. Note that this process adds no new information or true contrast to the scene, but only makes the existing contrast more apparent to the observer. Examples of the use of lookup table operations include enhancing dark regions of an image (e.g., shadows) as shown in Figure 8.3, and isolation of certain DC levels (e.g., for isotherm mapping) as shown in Figure 8.4.

Another commonly used LUT operation is *histogram equalization*. This operation is designed to normalize contrast by spreading the output image data approximately equally across the available dynamic range. It is based on the *cumulative distribution function* (*CDF*) of the input image histogram, which is simply a one-dimensional vector where each element is that fraction of the pixels in the image with DC values less than or equal to the number of the element (note the index starts at zero), i.e.,

$$\mathrm{CDF}_n = \sum_{i=0}^{n} \mathrm{HIS}'_i(f) \tag{8.7}$$

where HIS'_i is the entry in the normalized histogram for a digital count equal to i. For example, in an 8-bit image, the zeroth element would be the fraction of pixels in the image with DC = 0, the third element would be the fraction of the image with DC values between 0 and 3 inclusive, and the 255th element would be 1 since it would be the fraction of the pixels with DC values between 0 and 255. Gonzalez and Wintz (1987) demonstrate that by multiplying the CDF vector by DC_{max}, converting the resultant values in the vector to integers, and using the integer vector as a LUT, an input histogram can be converted to an approximately equalized histogram (the approximation is due to the discrete nature of the integer data). This process, demonstrated in Figure 8.5, is commonly used as an initial preprocessor for image display. Although its effect is somewhat harsh for pictorial images, it is often acceptable for remotely sensed images. Note that the lookup table for histogram equalization is generally nonlinear. Because nonlinear processes have unpredictable effects on the spatial frequency content of the image, histogram equalization must be used with care in applications where the spatial frequency spectra of multiple images are compared.

Because the CDF is a monotonically increasing function, the lookup table can be inverted to a good approximation. The process is shown in Figure 8.5 to convert the equalized image back to the original image. The inversion is only approximate where "multiple binning" has taken place, i.e., at those bins where two gray levels in the input were mapped to a single output level. When reversed, the pixels that come from multiple bins are assigned to a single gray level (or may be randomly assigned to the two gray levels), resulting in subtle changes in the image. We review this here not so much for the current discussion but to caution the reader that many gray-level manipulations are nonreversible and can result in loss of information due to multiple binning. Our interest in using the scaled CDF LUT

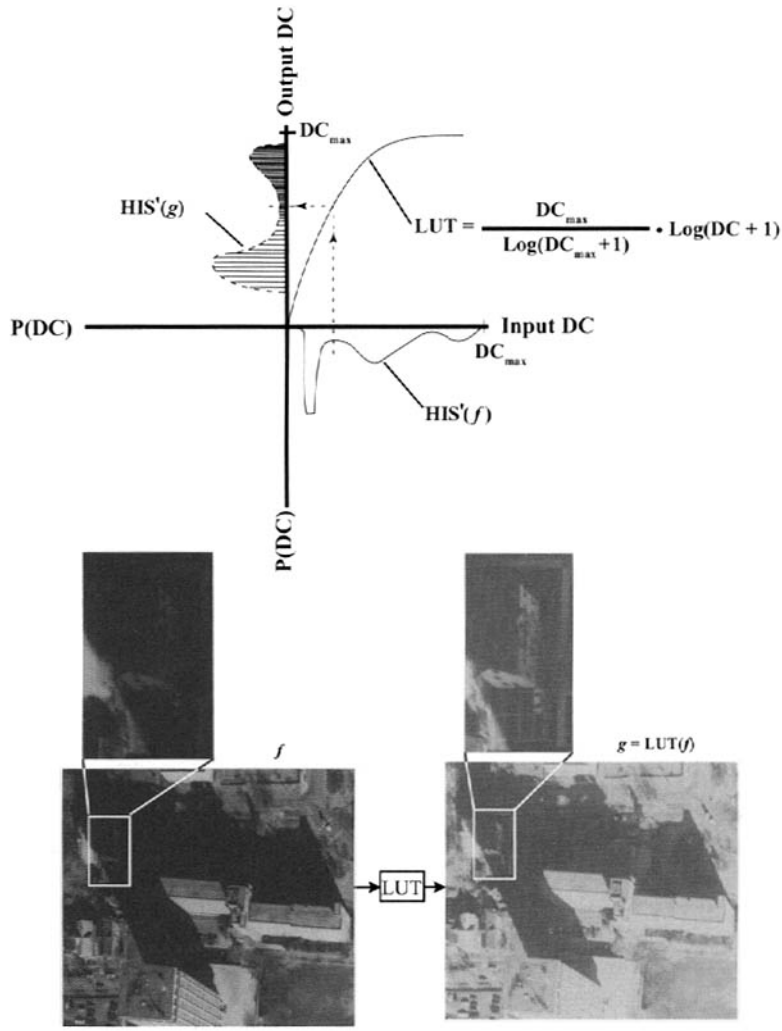


Figure 8.3 Logarithmic LUT used to enhance shadows in an air photo.

in reverse is not to reverse the input image but to force one image to appear approximately the same as a second.

Consider two images of an identical scene taken with different sensor gain and bias or under different illumination levels. If the brightness distribution function (histogram) of the second image can be processed to match the distribution of the first, then the apparent differences between the images should be removed. Gonzalez and Woods (2002) show that in the discrete case this probability matching or specification can be approximated using a process called *histogram specification* (Figure 8.6), where the histograms from two images of the same scene are computed along with their CDFs. We assume these images have nearly identical reflectance distributions, and we want to force the lower contrast ("day-2") image to have a histogram that matches the higher contrast (day-1) image in order to approximately remove illumination and sensor effects. This is accomplished by conceptually passing the low-contrast (day-2) input image through a LUT composed

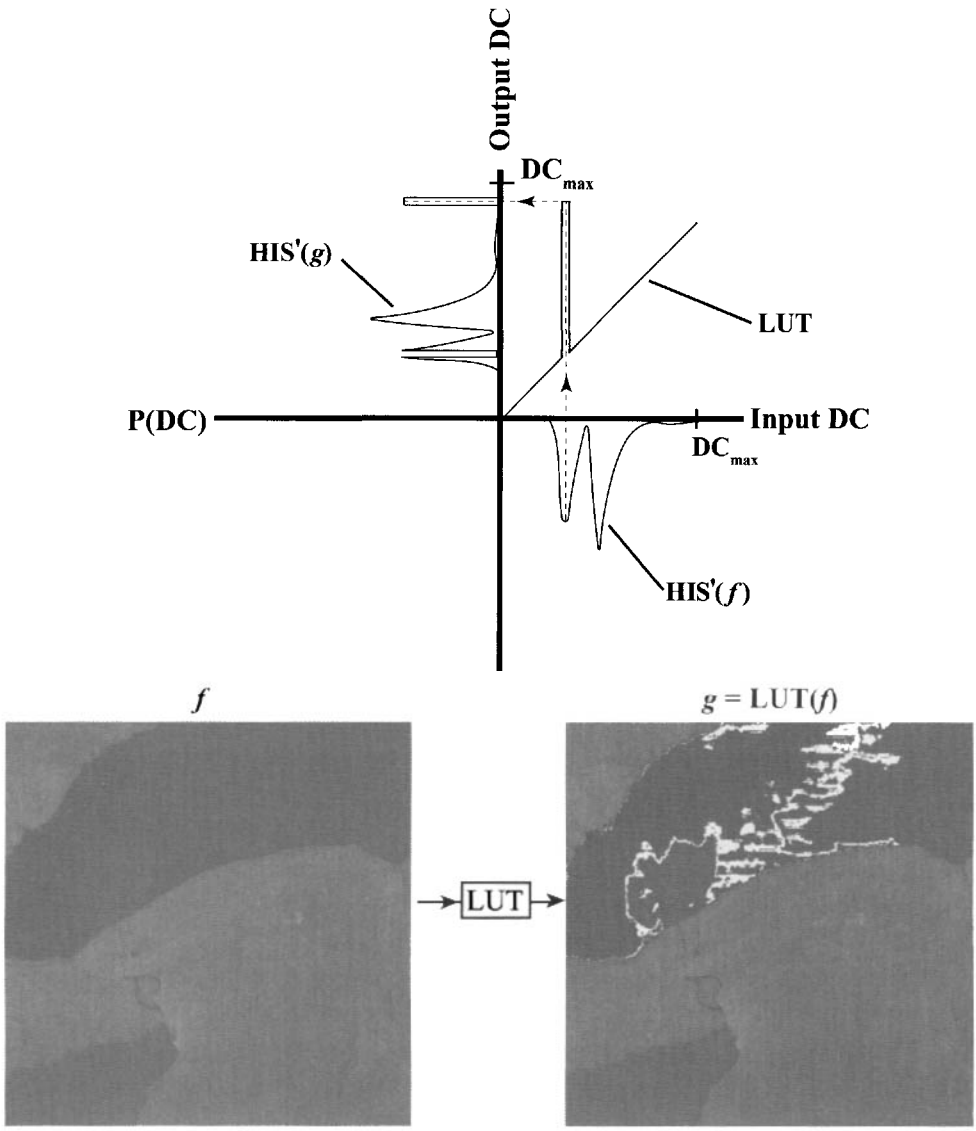
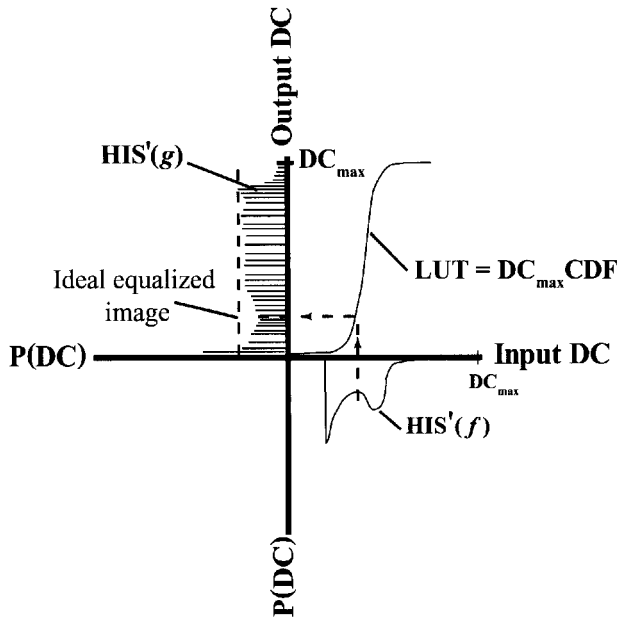


Figure 8.4 LUT used to isolate selected digital counts for use in generating isopleth maps. The input image is a Landsat TM thermal band of Lake Ontario.

of its scaled CDF, resulting in an equalized histogram as illustrated in Figure 8.5. This equalized image is then passed backward through the scaled CDF of the high-contrast ("day-1") image to force its histogram to match that of the high-contrast image. In practice, this is accomplished in a single step by applying a LUT that combines the forward and reverse CDF LUTs. As seen in Figure 8.6, this process removes gross contrast differences between the images. However, in cases where the assumptions of identical reflectance distributions are seriously violated, false changes can be introduced. In this case, the water in the day-2 input image that was fairly clear has been forced to approximately match the turbidity present on day 1.



$$g(i,j) = \text{LUT}[f(i,j)] = \text{DC}_{\max}\text{CDF}[f(i,j)]$$

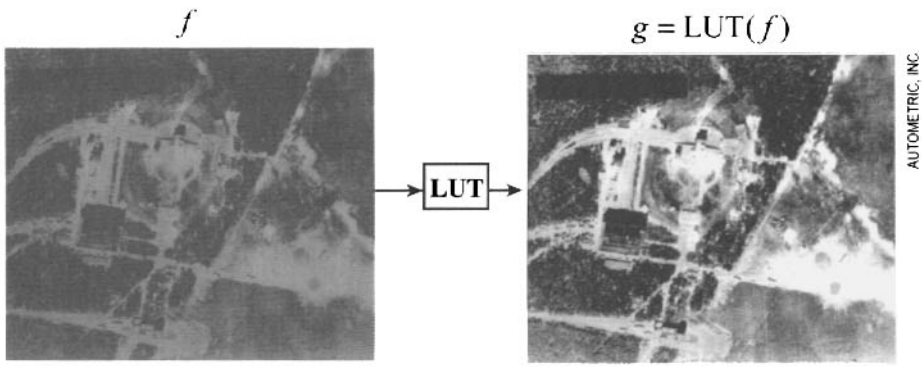


Figure 8.5 Histogram equalizing process. Image is a digitized U.K. reconnaissance photo of the V-2 rocket test facility at Peenemunde acquired July 4, 1944.

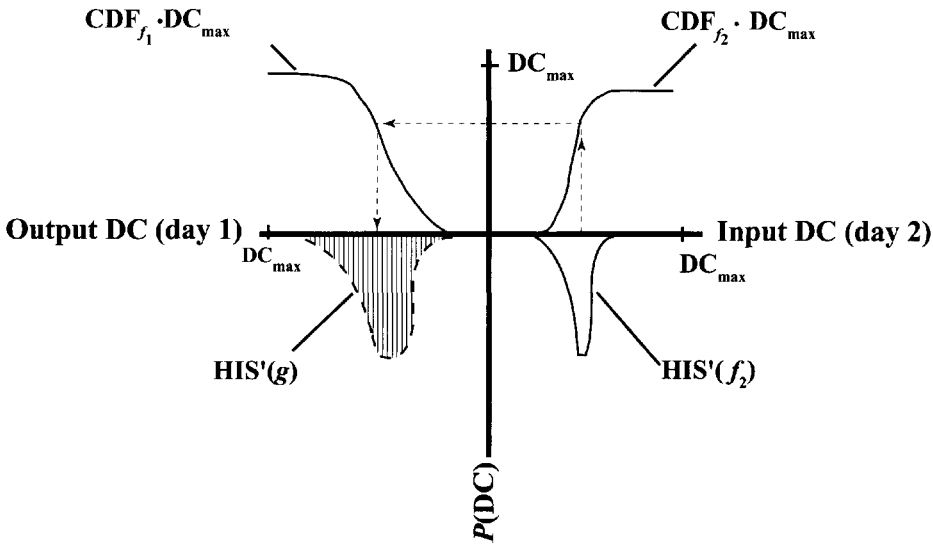
Thus, as discussed in Section 7.5.2, care should be taken when using this approach as a quantitative normalization process.

We have implicitly assumed that the point operations presented here are applied uniformly over the entire image. They can also be applied adaptively by constructing LUTs for local image regions. For example, an algorithm can be designed that increases contrast only in low-contrast regions of the scene. The local contrast is computed in the neighborhood of each pixel. If the contrast is less than some user-defined threshold, the contrast is stretched about the mean of the local window. This results in an output image whose gray levels are unchanged, on average, but with enhanced contrast of localized areas of low contrast such as shadows

and highlights. Such adaptive methods are not truly point operations, since the nature of the transform depends on the neighborhood around the point in question.

8.2 NEIGHBORHOOD OPERATIONS-KERNEL ALGEBRA

The next class of DIP algorithms generates the gray value of the output pixel as an explicit function of some number of neighboring pixels. In most cases, neighborhood operations are designed to alter the sharpness (local edge contrast) of an image. Most of the common neighborhood operations of interest can be implemented using what is referred to as a *convolution kernel* (cf. Fig. 8.7). The *kernel*



The histogram specification process can be expressed as

$$g(i,j) = \mathrm{DC}_{\max} \bullet \mathrm{CDF}_{f1}^{-1}\left[\mathrm{DC}_{\max} \bullet \mathrm{CDF}_{f_2} f_2(i,j)\right]$$

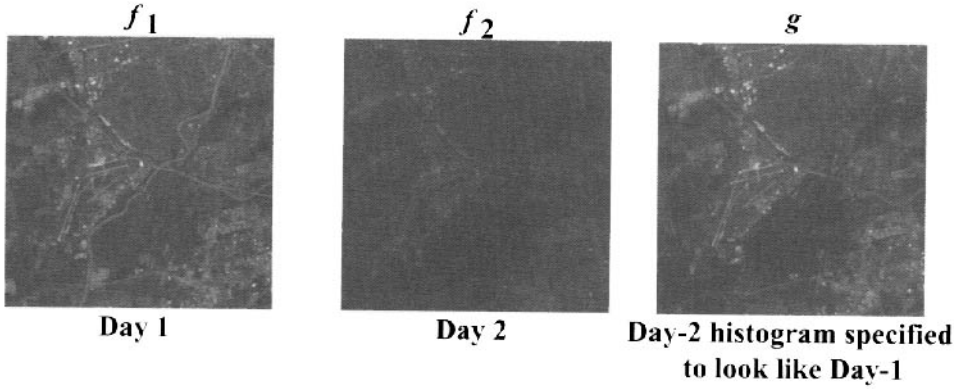


Figure 8.6 Histogram specification process.

can be thought of as a small window of weighting values whose size defines the size of the kernel (most often square and with an odd number of pixels on each axis). The kernel operation involves locating the center of the kernel over each input pixel location (i,j) and multiplying each weight in the kernel by the DC value in the input image under the kernel. The sum of the products is the gray value of the output pixel $g(i,j)$. The kernel is said to be operating on the pixel under the center element of the kernel (hence the common use of odd-size kernels). The kernel is stepped sequentially over each pixel in the input image. Note that the edge pixels (i.e., where a portion of the kernel falls outside the boundary of the image) are generally either left undefined or filled by applying one of a variety of methods to maintain the overall image statistics. Schowengerdt (1983) shows that this process is mathematically the equivalent of the discrete *correlation* except at the edges:

$$g(i,j) = \sum_{m=i-w/2}^{i+w/2} \sum_{n=j-w/2}^{j+w/2} f(m,n)h(m-i,n-j) = f(i,j) \bigstar h(i,j) \tag{8.8}$$

where m and n are dummy variables of summation, h is the kernel of size w x w, $w/2$ is a truncated integer value, and "★" is the symbol used to denote correlation. The kernel is indexed over the range $-w/2$ to $w/2$, with the center location defined as (0,0). If the values in the kernel are "flipped" about the x and y axis (equivalent to rotating the kernel by 180°), then the operation is the mathematical equivalent of discrete *convolution*:

$$g(i,j) = \sum_{m=i-w/2}^{i+w/2} \sum_{n=j-w/2}^{j+w/2} f(m,n)h(i-m,j-n) = f(i,j) * h(i,j) \tag{8.9}$$

where $*$ denotes the convolution operation. For example, a 3 x 3 element correlation would be implemented mathematically as:

$$g(i,j) = \sum_{m=i-1}^{i+1} \sum_{n=j-1}^{j+1} f(m,n)h(m-i,n-j) \tag{8.10}$$

The convolution operation so fundamental to linear system theory can also be implemented using Eq. (8.10) by flipping the kernel values about x and y.

Let's examine the effect of various simple kernels to improve our intuitive understanding of kernel operations. Consider these simple 3 x 3 element kernels implemented using Eq. (8.8):

$$h_1(i,j) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad h_2(i,j) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & k & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$h_3(i,j) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \quad h_4(i,j) = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

Kernal h_1 is the identity operation and simply replicates the input image. Kernel h_2 produces an image with gray values amplified by the factor of k (i.e., it produces a contrast-enhanced image for $k > 1$ or a contrast-reduced image for $k < 1$). Kernel h_3 replaces each pixel by its neighbor immediately to the right, producing an output

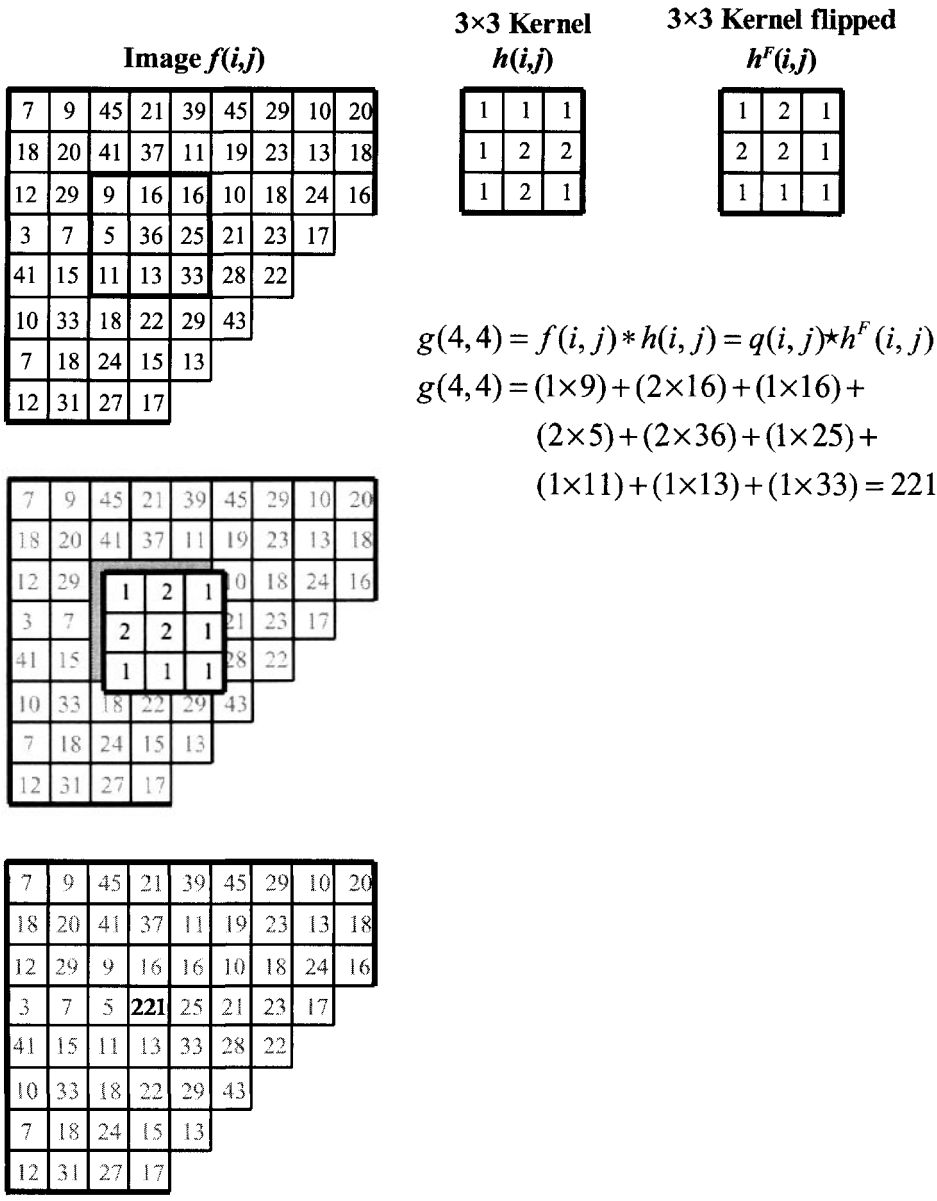


Figure 8.7 Correlation (★) and convolution (*) operations with a kernel operator.

image that is shifted one pixel to the left. Note that if the convolution operation is performed using Eq. (8.9), the image would be translated one pixel to the right. Kernel h_4 generates output pixels that are the sums of the input pixel and its eight nearest neighbors; it is a "*local integrator.*" The resultant image will be nine times brighter than the input image and blurred slightly. Integrating kernels are often scaled by dividing by the sum of the weights to ensure that the dynamic range of the output image is approximately equal that of the input. The scaled version of h_4 is

$$h_5 = \frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} \end{bmatrix} = \frac{1}{9} h_4(i,j) \tag{8.11}$$

An output pixel will be the mean value of the corresponding input pixel and its eight nearest neighbors. The local averager "blurs" the image. In linear systems, it is referred to as a *low-pass filter* or a low-pass kernel, since it has little impact on sinusoidal components with small spatial frequencies (regions with slowly varying gray value) and filters out ("blurs") sinusoidal components with large spatial frequencies (required to form edges) (cf. Sec. 8.4 for additional treatment of filtering operations).

Kernel operations can also be used to locate or exaggerate edges by local differencing operations. Consider the following operators:

$$h_6 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \quad h_7 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \end{bmatrix} \quad h_8 = \begin{bmatrix} 0 & +1 & 0 \\ +1 & -4 & +1 \\ 0 & +1 & 0 \end{bmatrix}$$

Kernel h_6 subtracts each input pixel from its neighbor on the right. The resultant image will have zero values if the two adjacent pixels have the same gray value, large positive values if the pixel to the right is significantly brighter, and large negative values if the pixel to the right is significantly smaller than the center pixel. This operation is equivalent to the discrete first derivative of the image with respect to x:

$$\frac{df}{dx} = \lim_{\Delta x \to 0} \frac{f(x+\Delta x, y) - f(x,y)}{\Delta x} = \frac{f(i+1,j) - f(i,j)}{1} = f(i,j) \bigstar h_6(i,j) \tag{8.12}$$

where the smallest nonzero value of $\Delta x = 1$, the spatial interval between pixels. Bipolar images such as those resulting from the first-derivative operator are usually "biased up" (and the gray scale may be compressed) so that a gray value of zero is depicted by a mid-gray in the image. Darker or brighter pixels have negative or positive output values, respectively. The extreme gray values of images produced by kernel h_6 are produced at the locations of edges in the image, and the outputs are unaffected by any constant gray-value component of the input. This is no surprise; the derivative of any constant value must be zero.

Kernel h_7 evaluates the first derivative in the y direction. Its extreme output values occur at the locations of horizontal changes in gray value (horizontal "edges"), and the operator is again insensitive to constant values. Kernel h_8 is the Laplacian operator and is the discrete analogue of the sum of the second partial derivatives with respect to x and y, i.e.,

$$g = \nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} = [f(i+1,j) - f(i,j)] - [f(i,j) - f(i-1,j)]$$
$$+[f(i,j+1) - f(i,j)] - [f(i,j) - f(i,j-1)] = f(i+1,j) + f(i-1,j)$$
$$+f(i,j+1) + f(i,j-1) - 4f(i,j) = f(i,j) \bigstar h_8(i,j) \tag{8.13}$$

This kernel can be thought of as subtracting the original image from a blurred version of itself formed from averaging its four nearest neighbors. The resulting image highlights edges and isolated bright or dark pixels. The negative-valued weights require that it be biased up for display in the same manner used for first-derivative images. The Laplacian is a form of *high-pass filter* because low-frequency information is attenuated and the average constant value is "blocked" (filtered out); high-frequency components necessary to create edges are preserved and enhanced. Figure 8.8 shows the effects of a variety of kernels on the same input image. Convolution is widely used in image processing because it is the operation required to implement any linear spatial filter, including modeling of most optical processes (e.g., atmospheric and sensor degradation and blur due to detector sampling), as well as electronic processes [e.g., preamp frequency response effects (cf. Secs. 8.4 and 13.1 for more on filtering and modeling of images].

Nonlinear image processing operations are also of interest to remote sensing. In particular, the evaluations of local medians and modes are often useful. These operations also are calculated over a local neighborhood defined by a window centered about the pixel of interest. The window is stepped from pixel to pixel, and the desired statistic is calculated for the neighborhood in the same fashion as the correlation and convolution kernel operations. The pixels to be considered for the operation are often labeled by unit weights in the defining window; pixels not to be used are labeled by zeros. An example of the output resulting from a mode filter is shown in Figure 8.9. This operation is commonly used to remove random class values at class boundaries due to mixed pixel effects in a map of classes (cf. Chap. 9). The values scaled by the unit weights in the window are sorted to determine the digital count value(s) that occur(s) most often (mode operation). When the input digital counts correspond to classes rather than gray value, this operation tends to replace isolated pixels or strings of pixels with the most common class in the neighborhood, while smoothing the boundaries between classes. Note that this process can have detrimental effects, as well. For example, if the sizes of objects in a critical class are small (e.g., about one or two GIFOVs), then this operation may remove these targets from the final class map.

The median filter is a nonlinear filter commonly used to remove "salt and pepper" noise in images. It removes pixels whose values are very different than the majority in the neighborhood (i.e., with very high and low values that can be introduced by certain types of electronic noise). The median filter is one type of "*rank-order filter*." A more general form for rank-order filters is described by Hardie and Boncelet (1993).

The majority of the image processing tools presented in the last two sections can improve the appearance of an image, making it easier for an analyst to visually extract information. Many can be used as preprocessing algorithms for machine analysis (e.g., noise reduction and edge delineation). Some of the neighborhood operations that are particularly useful as a preprocessor for machine-based statistical pattern recognition algorithms (treated in Chap. 9) are texture measures.

8.3 STRUCTURE OR TEXTURE MEASURES

Structure metrics are a class of neighborhood operations designed to characterize the variability (*texture*) in the neighborhood around a pixel. In general, they are calculated over a window of selectable size that moves over the image in the same fashion as a kernel operation. The output pixel is computed from some measure of the texture in the window. A large variety of texture metrics exists. Some are very direction specific, some require fairly large windows and are aimed at characterizing low-frequency texture, and others work best with small windows and are designed to differentiate high-frequency texture patterns. One of the simplest structure metrics is the range of digital counts in the neighborhood. For example, using a 3×3 window centered on $f(i,j)$, the output $g(i,j)$ would be the range $(DC_{max} - DC_{min})$ evaluated for $f(i,j)$ and its eight nearest neighbors. The output image would be bright in regions containing fine structure (large spatial frequencies) and dark in regions containing only slowly varying small-amplitude components. A similar texture metric is the local standard deviation, where the output image pixel value $g(i,j)$ is simply the standard deviation of the pixels in the specified neighborhood centered on $f(i,j)$ in the input image. Some of these texture metrics include the co-occurrence metrics described by Haralick et al. (1973), the run-length metrics discussed by Galloway (1975), and texture spectra described by Wang and He (1990). It is also common to use texture metrics isolated in the frequency domain as described by Stromberg and Farr (1986) (cf. Sec. 8.4 for more on frequency-domain metrics). Figure 8.10 represents a sample set of images formed from texture metrics. The gray values in these images can be used in algorithms designed to classify material types on the basis of texture. For example, in a single-band image containing only trees and water, the gray levels may overlap between the two classes. However, at certain resolutions, the local standard deviation in a 3 x 3 window of water pixels would be very small compared to that computed in a neighborhood containing forest pixels. In this case, a simple threshold of the standard deviation image can dramatically improve classification accuracy. In most cases, we must consider many texture metrics when trying to segment different material classes or to merge texture metrics with multispectral digital count values to provide greater differences between classes.

One of the most widely used sources of texture metrics is the *gray-level co-occurrence* matrix, which describes the probability that two specific gray values or digital counts occur at two relative locations in a processing or sampling window. The processing window is designed to characterize the pixel in the center of the window based on its gray level and that of its neighbors within the window. Like

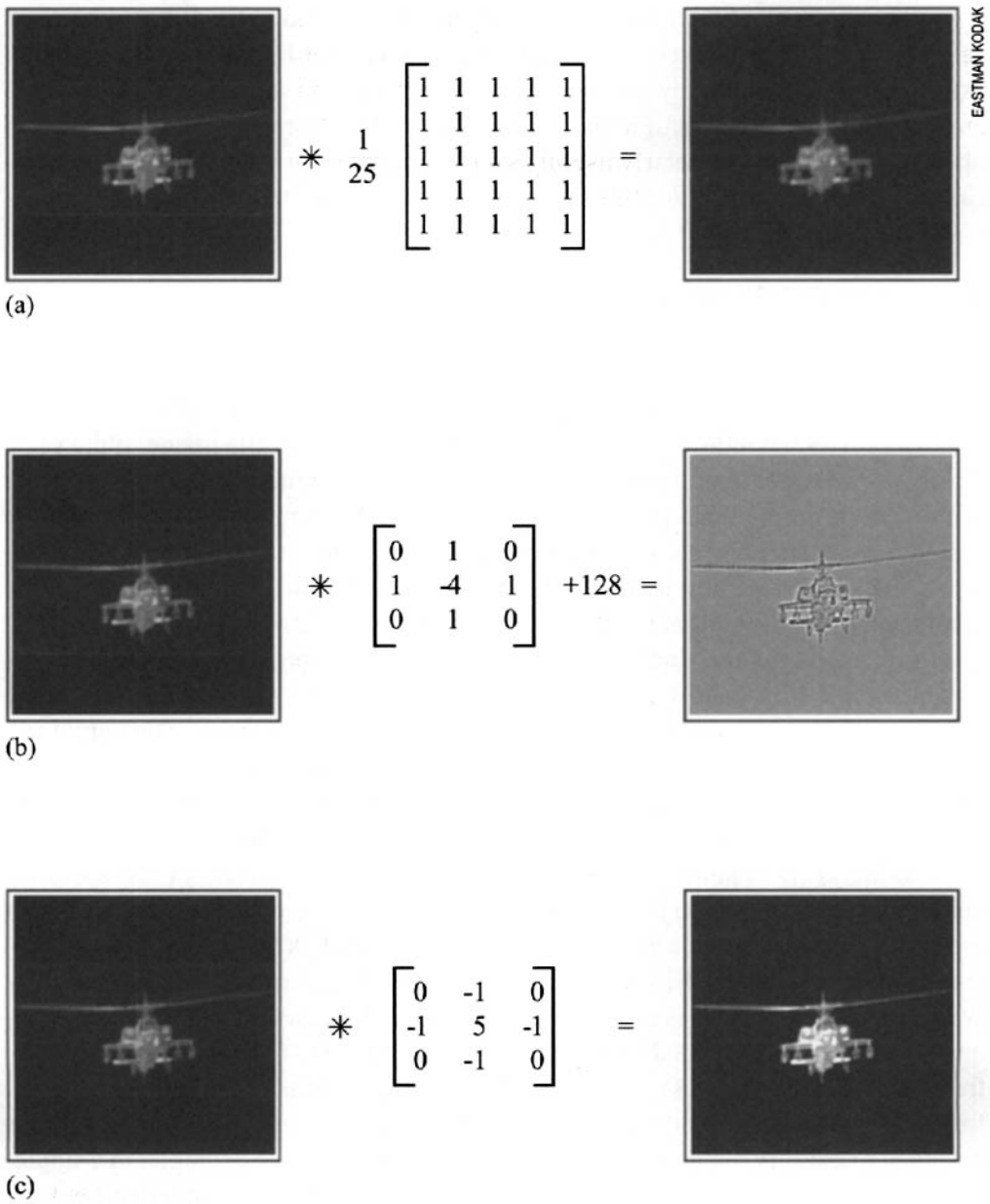
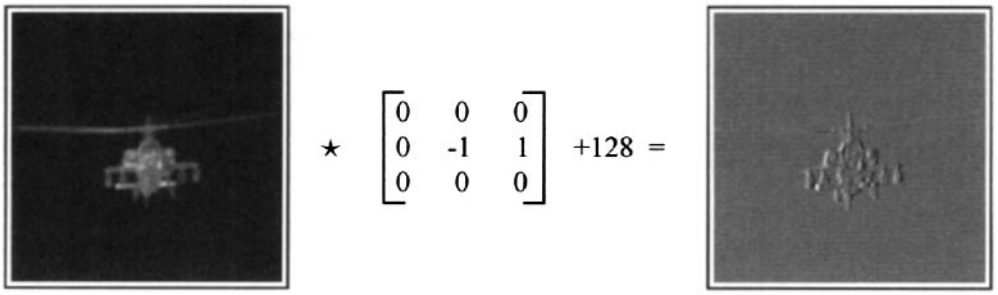
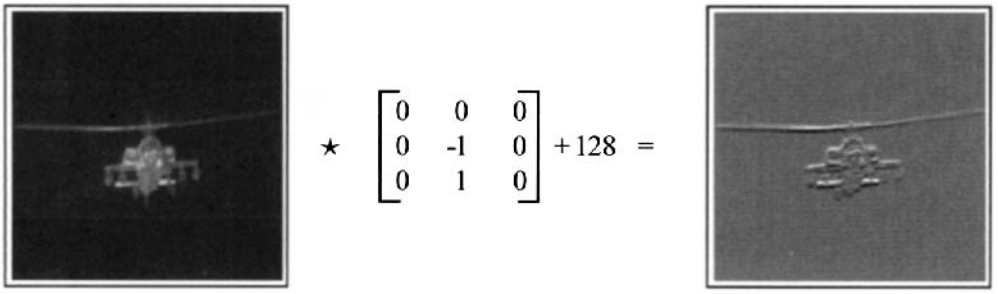


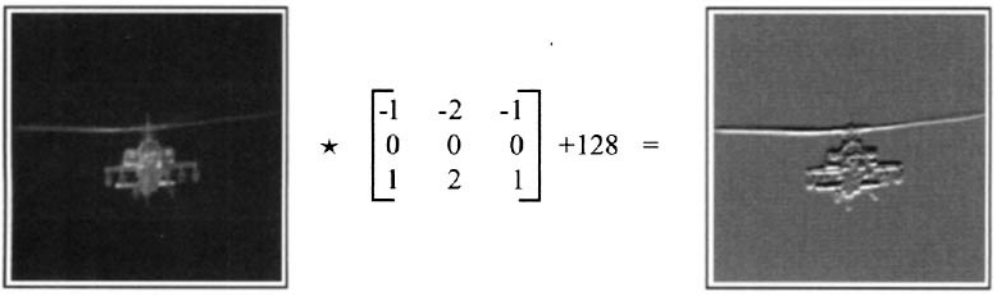
Figure 8.8 Examples of neighborhood operations employing a kernel operator. Image is an MWIR image taken with a 2-D array sensor. (a) 5 × 5 blur kernel (low-pass filter) used for noise reduction. (b) Laplacian kernel (high-pass filter) to highlight edges. (c) Image minus its Laplacian used to sharpen edges (high-frequency boost). (d) First-derivative image with respect to x. (e) First-derivative image with respect to y. (f) Sobel horizontal-edge detector.



(d)



(e)



(f)

Figure 8.8 continued.

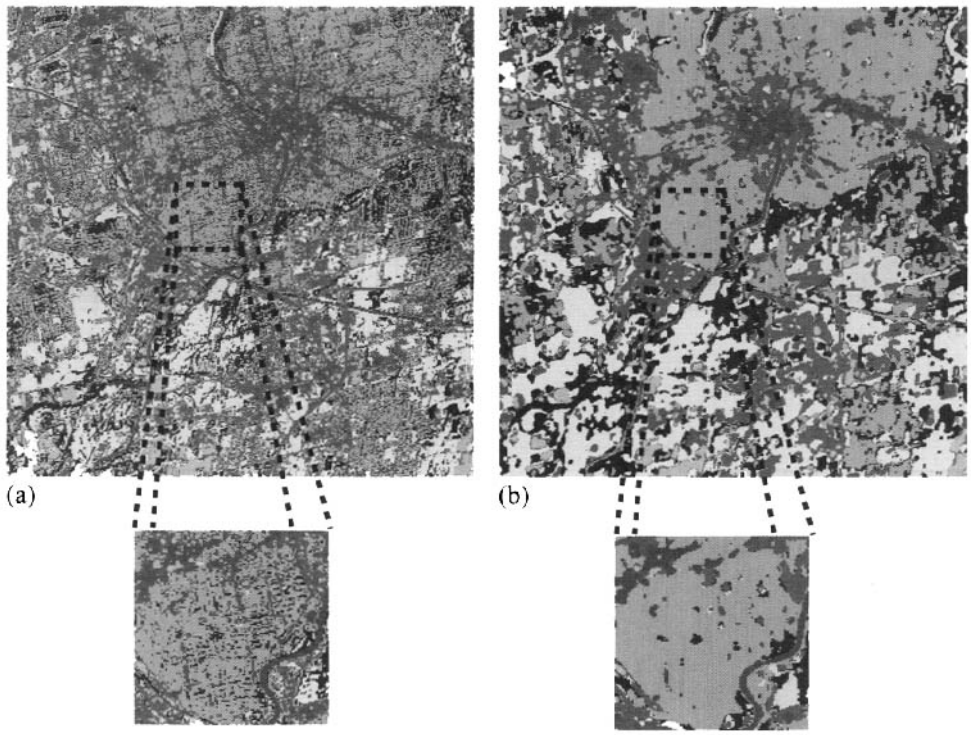


Figure 8.9 Mode filter operation and application to a class map image. (a) Original class map of a portion of a Landsat image of Rochester, NY. (b) Class map after application of a mode filter.

the kernel operators described previously, the window is moved over the image operating sequentially on each pixel. Consider, for example, the 5 x 5 neighborhood shown below for an image with gray levels (0–3):

0	1	3	3	2
2	1	3	3	2
1	0	0	0	1
2	3	1	3	2
3	3	3	3	2

The size of the co-occurrence matrix is determined by the number of gray values in the image, 4 x 4 in this example. The co-occurrence matrix for a pixel and its neighbor to the lower right is determined by the frequency of occurrence of each pair of gray values. We first fill the i,jth location in an accumulation matrix with the number of times the digital counts $(i–1, j–1)$ occur in the prescribed orientation [i.e., with digital count $(i–1)$ down and right of digital count $(j–1)$], yielding

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 2 & 0 & 0 & 1 \\ 1 & 0 & 0 & 2 \\ 1 & 3 & 1 & 2 \end{bmatrix} \tag{8.14}$$

from which we see that A(4,2) tells us that we have a digital count of three occurring diagonally below and right of a digital count of one three times. To find the final co-occurrence matrix in probability form, we divide matrix **A** by the total population of occurrences (16 in our case) to form the gray-level co-occurrence (GLC) matrix **C**, i.e.,

$$\mathbf{C} = \frac{1}{16}\mathbf{A} \tag{8.15}$$

For remotely sensed images with 8–12 bits of information per pixel, the GLC matrices can become very large and sparsely populated (i.e., many possible combinations of pixels occur infrequently if at all). To avoid this problem, it is common to first equalize the histogram of the entire image and then requantize the image to fewer gray values. Furthermore, the orientation of the pixels in the joint probability may not be important, so sequences and orientations may be combined. For example, a 3,1 diagonally down one sample and right one sample may be treated the same at a 1,3 diagonally down one and right one. GLC matrices can capture some of the spatial texture information within a window. By adjusting the size and relative orientations used in the computation of the matrix, we can potentially increase the amount of information. However, as we create a co-occurrence matrix (or perhaps several) for each pixel by moving the sampling window over the image, we quickly overwhelm the capability of an analyst or an algorithm to cope with the volume of data. Haralick et al. (1973) suggested several ways to combine the values in a co-occurrence matrix into texture metrics. Each of these texture metrics can then be placed in the image location corresponding to the center of the sample window. An image can then be formed representing that metric by moving the window over the entire image. While many of the texture metrics are highly correlated, many investigators have found that some subset of the metrics derived from the GLC matrices are useful for analyzing spatial texture. Some examples of texture metrics suggested by Haralick et al. (1973) that can be used to generate texture feature images are the energy (T_1), contrast (T_2) and entropy (T_3) given by

$$T_1 = \sum_{i=1}^{n}\sum_{j=1}^{n}[C(i,j)]^2 \tag{8.16}$$

$$T_2 = \sum_{i=1}^{n}\sum_{j=1}^{n}(i-j)^2 C(i,j) \tag{8.17}$$

$$T_3 = -\sum_{i=1}^{n}\sum_{j=1}^{n} C(i,j)Log[C(i,j)] \tag{8.18}$$

where $C(i,j)$ is the (i,j)th entry in the GLC matrix ($\boldsymbol{C}$) and n is the number of gray levels (i.e., $\boldsymbol{C}$ is of dimension $n \times n$). Examples are shown in Figure 8.10.

While texture metrics may aid in visual exploitation, their main function is to be used as inputs to subsequent pattern recognition algorithms. The statistical pattern recognition tools discussed in Chapter 9 describe procedures for image classification using spectral and/or textural metrics. Because of the wide variety of possible textural metrics, it can be difficult to isolate those that are truly useful. Rosenblum (1990) describes a method for selecting texture features similar to the method described by Swain (1978) to select spectral bands useful for classifying land cover. These band optimization methods (discussed in greater detail in Sec. 13.3) rely on maximizing the statistical separability between classes characterized using the Gaussian maximum likelihood theory developed in Chapter 9.

8.4 GLOBAL OPERATIONS

Many convolution operations used in low-pass or bandpass filtering require rather large kernels to ensure that at least one cycle of the sinusoidal component is included in the neighborhood. At some point, the number of computer operations becomes large compared to an equivalent approach of performing the filtering operation in the frequency domain [cf. Gonzalez and Woods (2002) for a more complete treatment of global operations]. This involves transforming the entire image from the spatial domain into the frequency domain using the discrete Fourier transform:

$$F(u,v) = \mathcal{F}[f(n,m)] = \frac{1}{N}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1} f(n,m)e^{-2\pi i\frac{un+vm}{N}} \tag{8.19}$$

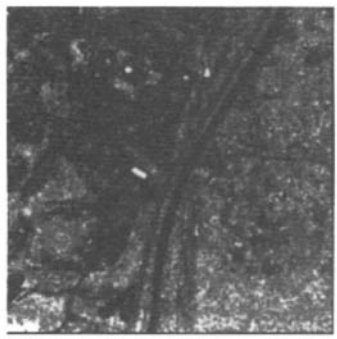
where u,v are the *spatial frequency* indices in the transformed image oriented in the x and y direction, respectively; u/N, v/N are the spatial frequencies; $\mathcal{F}$ is the Fourier transform operation; N is the side dimension of the image in pixels, $i = (-1)^{1/2}$; and F is the complex frequency domain representation of the image. This transform represents all the information in the input image in terms of the sinusoidal spatial frequencies present in the image. It is a global operation because each output value $F(u,v)$ is a function of all the input values $f(n,m)$. The output of the transform is an $N \times N$ array of complex numbers that can be visualized by displaying the data as a pair of images, one being proportional to the real part of the Fourier transform and the other to the imaginary part, i.e.,

$$F(u,v) = \mathrm{Re}[F(u,v)] + i\mathit{Im}[F(u,v)] \tag{8.20}$$

Alternatively, and more commonly, the data can be represented as the magnitude and phase derived from the Euler representation of complex numbers, i.e.,

$$F(u,v) = |F(u,v)|e^{i\phi(u,v)} \tag{8.21}$$

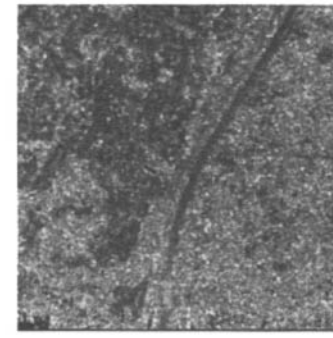
where



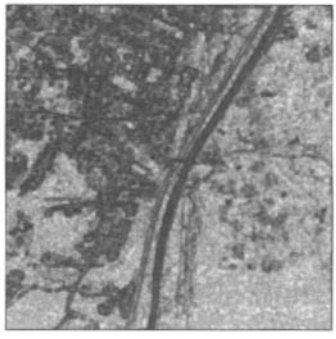
(a)

(b)

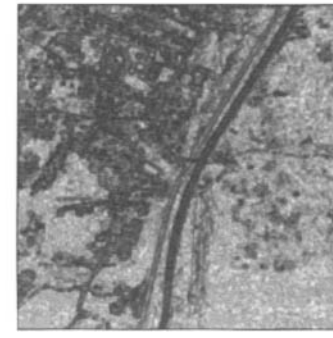
(c)



(d)



(e)



(f)

Figure 8.10 A digitized air photo (a) and several images representing a range of texture metrics or features. Note how different metrics accentuate different patterns or structures in the image.

$$|F(u,v)| = \left[\mathrm{Re}(F(u,v))^2 + \mathrm{Im}(F(u,v))^2\right]^{\frac{1}{2}} \tag{8.22}$$

is the magnitude of the Fourier transform (also referred to as the *Fourier spectrum*), $|F(u,v)|^2$ is referred to as the *power spectrum*, and

$$\phi(u,v) = \tan^{-1}\left[\frac{\mathrm{Im}[F(u,v)]}{\mathrm{Re}[F(u,v)]}\right], \pi \geq \phi(u,v) \geq -\pi \tag{8.23}$$

is the phase. The power (squared magnitude) of the Fourier transform at any spatial frequency coordinate (u,v) is proportional to the amount of energy in the entire input image that oscillated at (or contained) that frequency. From this perspective, the image is perceived to be made up of sinusoidal functions whose magnitude and relative phase determine their impact on the final image. The spatial frequency interval in the output array can be expressed as

$$\Delta u = \Delta v = \frac{1}{N\Delta x} = \frac{1}{N\Delta y}, \text{ if } \Delta x = \Delta y \tag{8.24}$$

where Δx and Δy are the sample intervals (i.e., distance between pixel centers assumed equal in this case) in x and y in the spatial domain representation of the image. We should recognize that the Fourier transform is a cyclic function with period N, so that the maximum frequency represented is the *Nyquist frequency*:

$$\frac{N}{2}\Delta u = \frac{N}{2}\Delta v = \frac{1}{2\Delta x} \tag{8.25}$$

For example, the smallest nonzero spatial frequency evaluated for a 512 × 512 subimage of a Landsat TM scene with a 30-m GIFOV, would be

$$u = 1\Delta u = \frac{1}{512 \cdot 30 \text{ m}} = \frac{1 \text{ cycle}}{15{,}360 \text{ m}} \text{ or } 65 \cdot 10^{-6} \text{ cycles/m} \tag{8.26}$$

i.e., just larger than zero frequency. This corresponds to a sinusoidal component that oscillates through one cycle across the 512 pixels of the image. The highest frequency sampled is defined by the Nyquist limit of the sampling process and is

$$u = \frac{N}{2}\Delta u = \frac{512}{2} \cdot \Delta u = \frac{1 \text{ cycle}}{60 \text{ m}} \text{ or } 1.67 \cdot 10^{-2} \frac{\text{cycles}}{\text{m}} \tag{8.27}$$

which corresponds to a sinusoid that oscillates through one cycle in two pixels.

It is important to recognize that the Fourier transform is a reversible process, so we can convert the image information from the frequency domain to the spatial domain. In other words, the inverse Fourier transform "synthesizes" the image from its frequency components:

$$f(n,m) = \mathcal{F}^{-1}[F(u,v)] = \frac{1}{N}\sum_{u=0}^{N-1}\sum_{v=0}^{N-1} F(u,v)e^{+2\pi i\frac{un+vm}{N}} \tag{8.28}$$

This means that the space-domain version of a processed Fourier transform may be evaluated, thus allowing us to take advantage of computational efficiency in the frequency domain. Regrettably, the mathematical underpinning of Fourier analysis and the intricacies of the links between optical processes and their mathematical representation in terms of frequency domain operations are beyond the scope of this book. The reader is urged to consider Bracewell (1986) or Papoulis (1962) for a detailed treatment of the Fourier transform, Gaskill (1978) or Goodman (1996) for treatment of the links between optical systems and frequency domain operations (linear systems), and Gonzalez and Woods (2002) for applications to digital image processing.

For the sake of brevity, our treatment of frequency-domain operations will be largely pictorial in an attempt to provide the reader with an intuitive understanding. The terminology used, particularly for special functions, is adapted from Gaskill (1978). Many of the subtleties of the process (e.g., aliasing) are usually neglected by such a treatment, and the reader unfamiliar with linear systems is encouraged to delve deeper into the subject before relying heavily on frequency-domain operations. Figure 8.11 shows an image and the various representations of the components of its Fourier transform, the so-called "*spectrum*" of the space-domain image. Note that the origin of coordinates in the frequency domain is placed in the center, as is customary in the treatment of optical systems. Also note that the periodic nature of the function is such that if the input image is real valued (which is always the case for digital images of optical intensity), then the magnitude (and thus the squared magnitude, or power) of the Fourier transform will be symmetric about the origin in the zero-centered representation. In many remote sensing images, most of the visually interpretable data in the Fourier transform are contained in the magnitude. To save space, we will usually show only the power spectrum of the Fourier transform in the rest of this section. To help in visualizing the information in the Fourier transform, subsections of an image with particular characteristics can be treated as subimages and their transforms computed. This is shown for several windows in Figures 8.12 and 8.13. In Figure 8.13, we can see how the Fourier transforms of images of different land-cover types appear significantly different. Stromberg and Farr (1986) suggested that these differences in the Fourier spectrum could be used to aid in classifying features with different texture. Ehrhard et al. (1993) demonstrated that the outputs of a set of bandpass filters can be combined into a texture metric to successfully classify background land-cover types using conventional statistical classifiers (cf. Chap. 9 for statistical classifiers).

Several important links exist between operations in the spatial domain and corresponding operations in the frequency domain. The one of greatest concern is the relationship of the various filters in the two domains. The *convolution* theorem from linear systems expresses these relationships in the "filter theorem" and "modulation theorem":

$$f(n,m) * g(n,m) = f^{-1}[F(u,v) \cdot G(u,v)] \tag{8.29}$$

$$f(n,m) \cdot g(n,m) = f^{-1}[F(u,v) * G(u,v)] \tag{8.30}$$

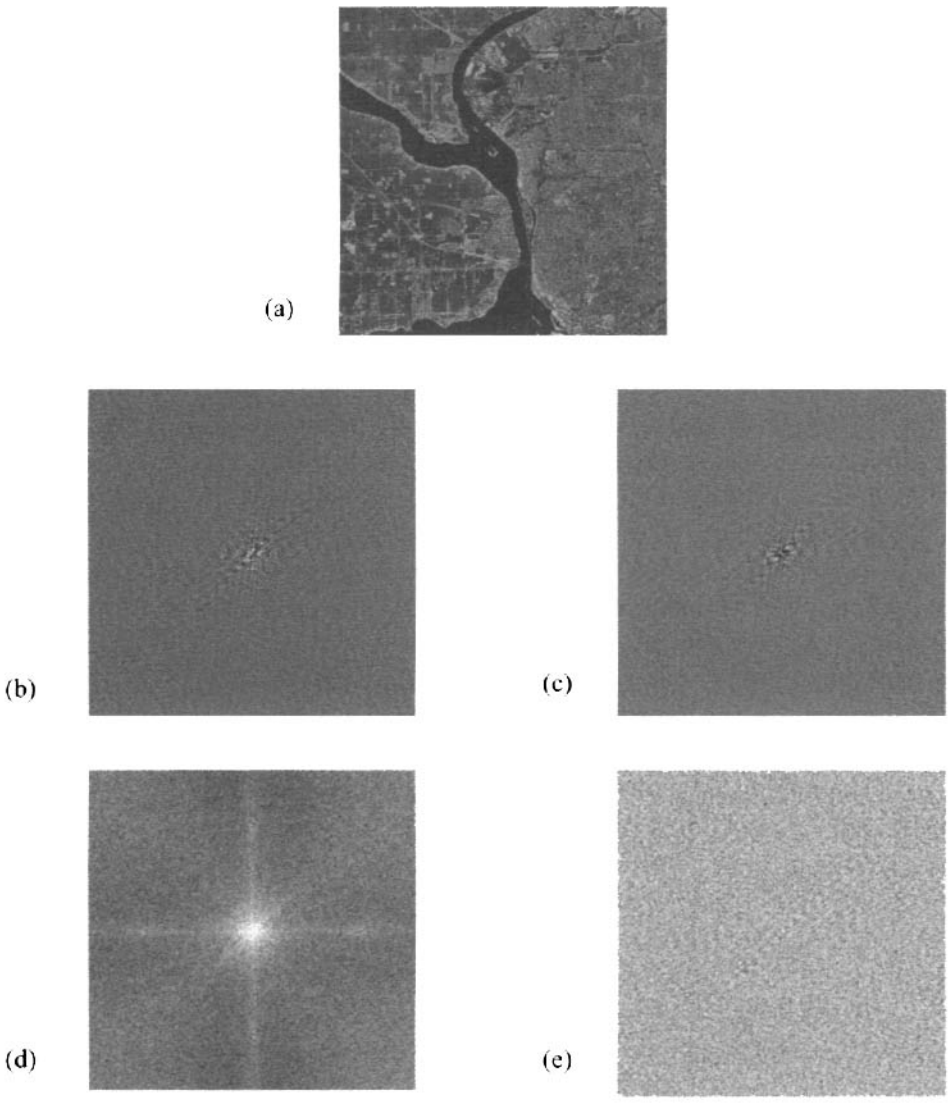


Figure 8.11 (a) Fourier transform of a digitized air photo of Buffalo, NY, and Niagara River, (b) real and (c) imaginary part of the Fourier transform, and (d) magnitude and (e) phase of the Fourier transform. Note that these images have been scaled for display purposes.

where $f(n,m)$ and $g(n,m)$ are image arrays and $F(u,v)$ and $G(u,v)$ are their Fourier transforms. This theorem states that the convolution operation in the spatial domain is equivalent to multiplication in the frequency domain and, conversely, that multiplication in the spatial domain is the equivalent of convolution in the frequency domain. Recall from Chapter 1 that we are assuming that the image chain can be reasonably approximated as a linear shift-invariant system.

Processing an image by changing the relative amount of energy in various spatial frequencies is most easily envisioned in the frequency domain. To facilitate this, we first introduce the concept of radial frequency to simplify our characteriza-

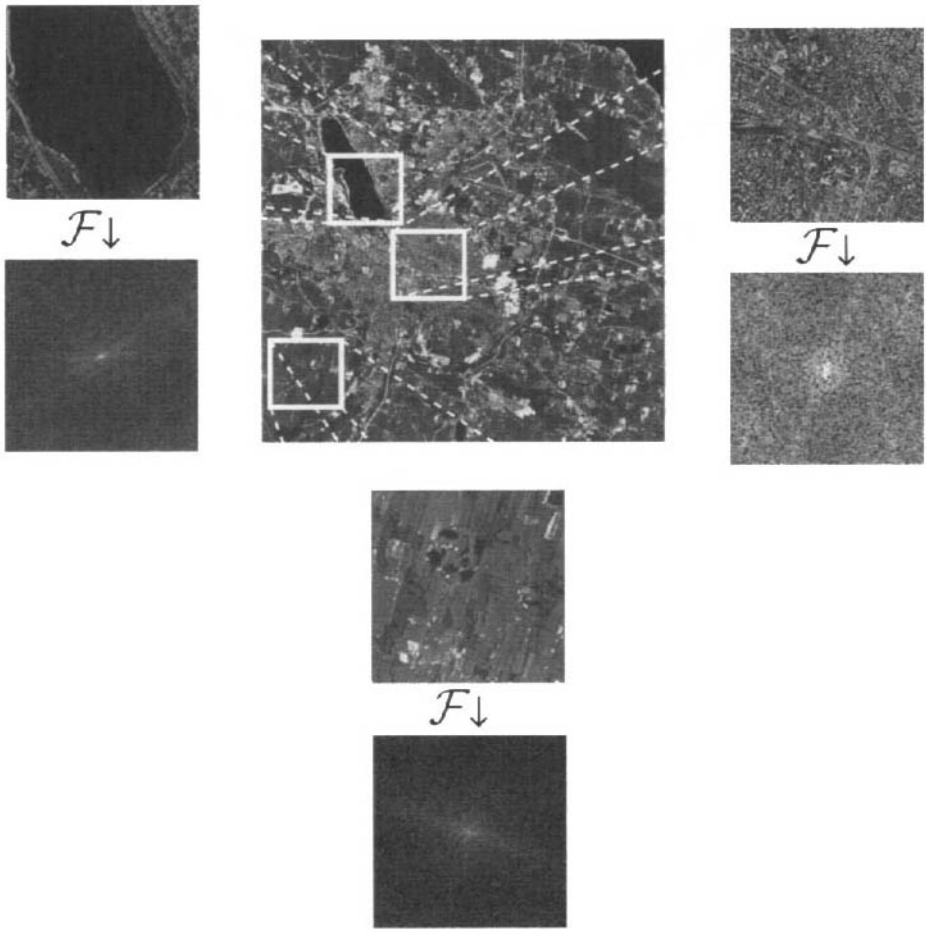


Figure 8.12 Magnitude representation of the Fourier transform of several windows in an image of Washington, DC.

tion of circularly symmetric filters. Let p be the radial spatial frequency evaluated from the Cartesian frequencies u and v via the Pythagorean theorem:

$$p(u,v) = (u^2 + v^2)^{\frac{1}{2}} \tag{8.31}$$

For circularly symmetrical filters, the relative amount of filtration for each radial frequency can be specified and the two-dimensional filter generated by rotation about the origin, as shown in Figure 8.14. The value of the filter $H(u,v)$ at any point defines how much of the energy at that spatial frequency will be present in the output image, where the output image is defined as

$$g(n,m) = \mathcal{F}^{-1}[F(u,v) \cdot H(u,v)] = \mathcal{F}^{-1}[G(u,v)] \tag{8.32}$$

For example, where the filter $H(u,v)$ is zero, so is the output amplitude $G(u,v)$. Values of $H(u,v)$ between 0 and 1 will reduce ("filter") the amount of energy, and values greater than 1 will increase the energy and exaggerate the frequencies affected. Figure 8.15 shows the impact of several frequency-domain filters on remotely sensed images.

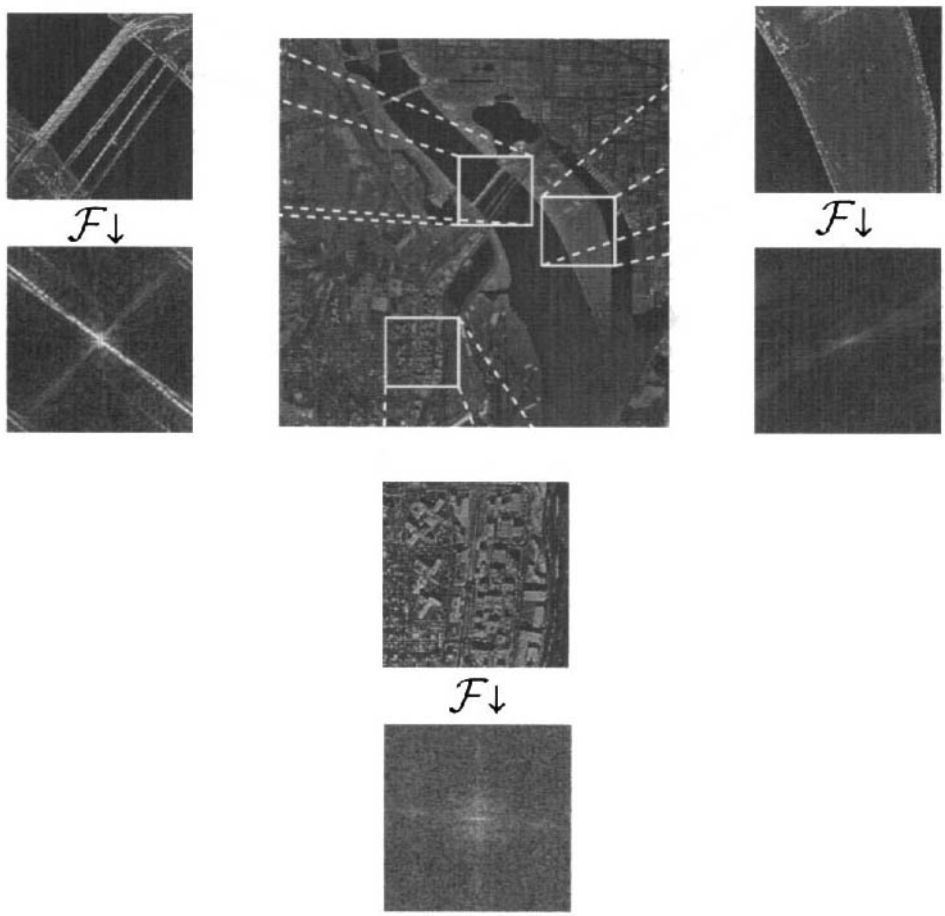


Figure 8.13 Magnitude representation of the Fourier transform of several windows in an image of Syracuse, NY.

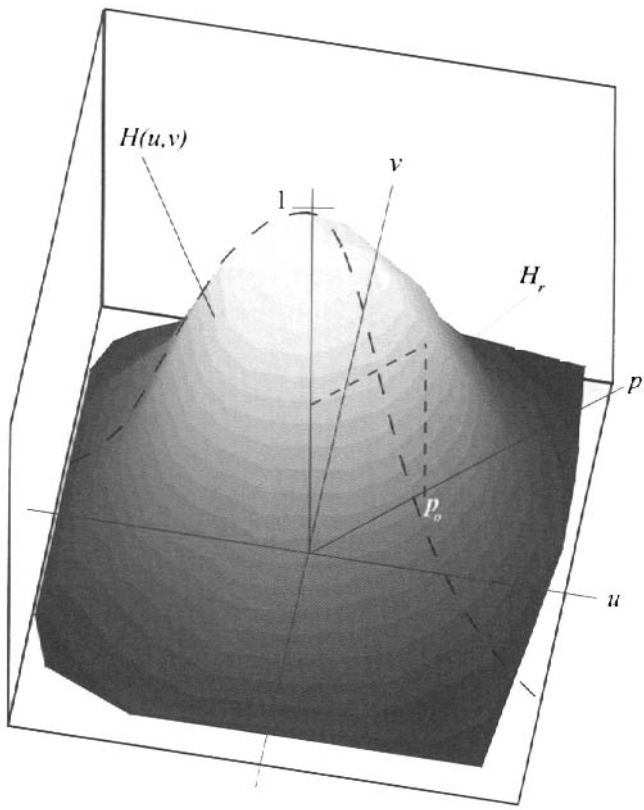
Clearly, filter design is conceptually simpler in the frequency domain. The number of elements in the kernel used to implement the filter in the spatial domain typically controls whether it is more appropriate to apply the filter by multiplication in the frequency domain (then back-transforming the filtered image) or by inverse-transforming the filter and convolving the image with the filter in the spatial domain. The ideal low-pass filter [Fig. 8.15(c)] eliminates all frequencies above the user-defined cutoff frequency (p_0). The resultant image is blurred and has some ringing artifacts due to the abrupt cutoff of the filter. The Butterworth low-pass filter [Fig. 8.15(d)] has no visible ringing due to its gentler transition and has less blurring effects even with the same cutoff since some high-frequency data are allowed to pass. The frequency-domain expression for the Butterworth low-pass filter is

$$H(u,v) = \frac{1}{1+\left(\dfrac{p(u,v)}{p_0}\right)^{2n}} \tag{8.33}$$

where $p_o=\sqrt{(u_0^2+v_0^2)}$ is a user-defined cutoff frequency (where the value of the filter goes to one-half the maximum) and n is the order of the filter and controls the steepness of the cutoff. The Butterworth high-pass filter shown in Figure 8.15(e) can be expressed as

$$H(u,v) = \frac{1}{1+\left(\dfrac{p_0}{p(u,v)}\right)^{2n}} \tag{8.34}$$

and has the same role of mitigating the ringing effects common in ideal high-pass filters. In many cases involving image enhancement, we are interested in a filter that amplifies ("boosts") the higher frequencies to compensate for the blurring in the imaging process that has reduced the high-frequency content in the image. Fig-



$$H(u,v) = \frac{1}{1+\left(\dfrac{u^2+v^2}{p_o}\right)^2}$$

$$H_r = \frac{1}{1+\left(\dfrac{p}{p_o}\right)^2}$$

Figure 8.14 Use of radial frequency concept to simplify definition of circularly symmetric filters.

ure 8.15(f) shows the general shape of a boost "filter" and its sharpening impact on the image. Regrettably, most amplifying filters inevitably exaggerate high-frequency noise in the image, such that amplifying beyond a certain point degrades the appearance of the image.

8.5 IMAGE RESTORATION

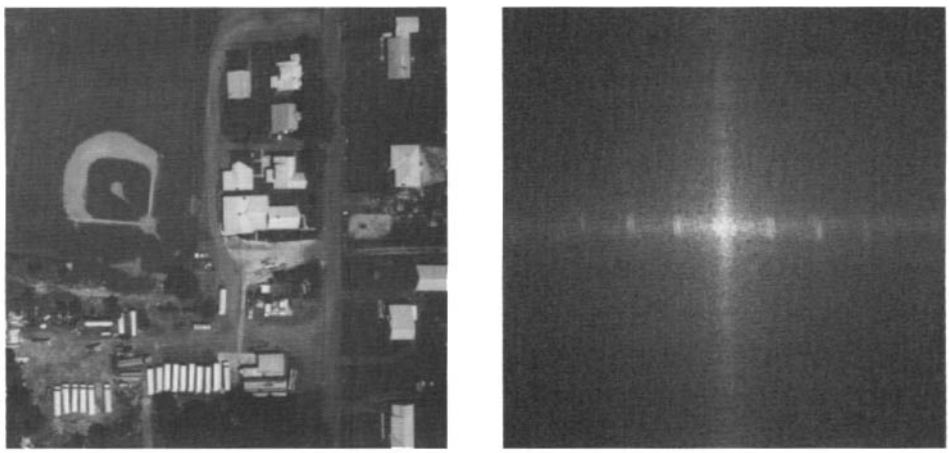
The ideal filter to restore an image degraded due to linear shift-invariant blurring in the absence of noise can be found from the expression for the appropriate blurring kernel:

$$g(x,y) = f(x,y) * h(x,y) \tag{8.35}$$

where $g(x,y)$ is the observed image, $f(x,y)$ is the "true" or undegraded image, and $h(x,y)$ is the blur kernel or *impulse response*. In imaging systems, it is often referred to as the *point spread function* (*PSF*). This is the image we would acquire if we imaged a point source. Mathematically, the undegraded image can be represented as a *delta function* (cf. Fig. 8.18), i.e.,

$$h(x,y) = \delta(x - x_0, y - y_0) * h(x,y) \tag{8.36}$$

where δ is the Dirac delta function defined to be zero everywhere except at (x_0,y_0), where it has unit volume. Using the convolution theorem of Eq. (8.29), we can transform Eq. (8.35) into the equivalent frequency domain expression by taking the Fourier transform of both sides to yield;



(a) (b)

Figure 8.15 Application of several circularly symmetric filters to a remotely sensed image: (a) image, (b) magnitude of the Fourier transform, (c) idealized low-pass filter, (d) Butterworth low-pass filter, (e) Butterworth high-pass filter, and (f) high-frequency boost filter.

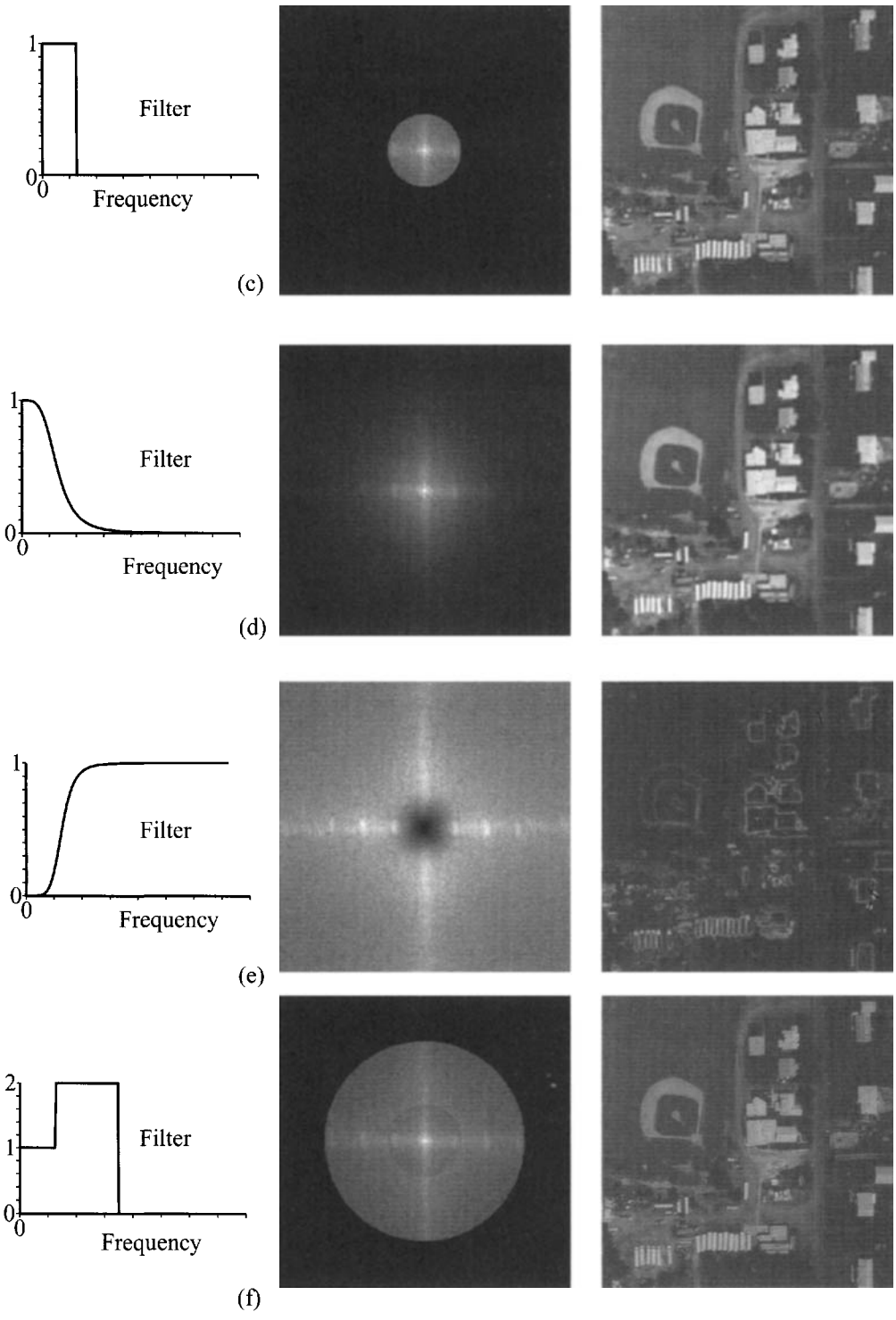


Figure 8.15 continued.

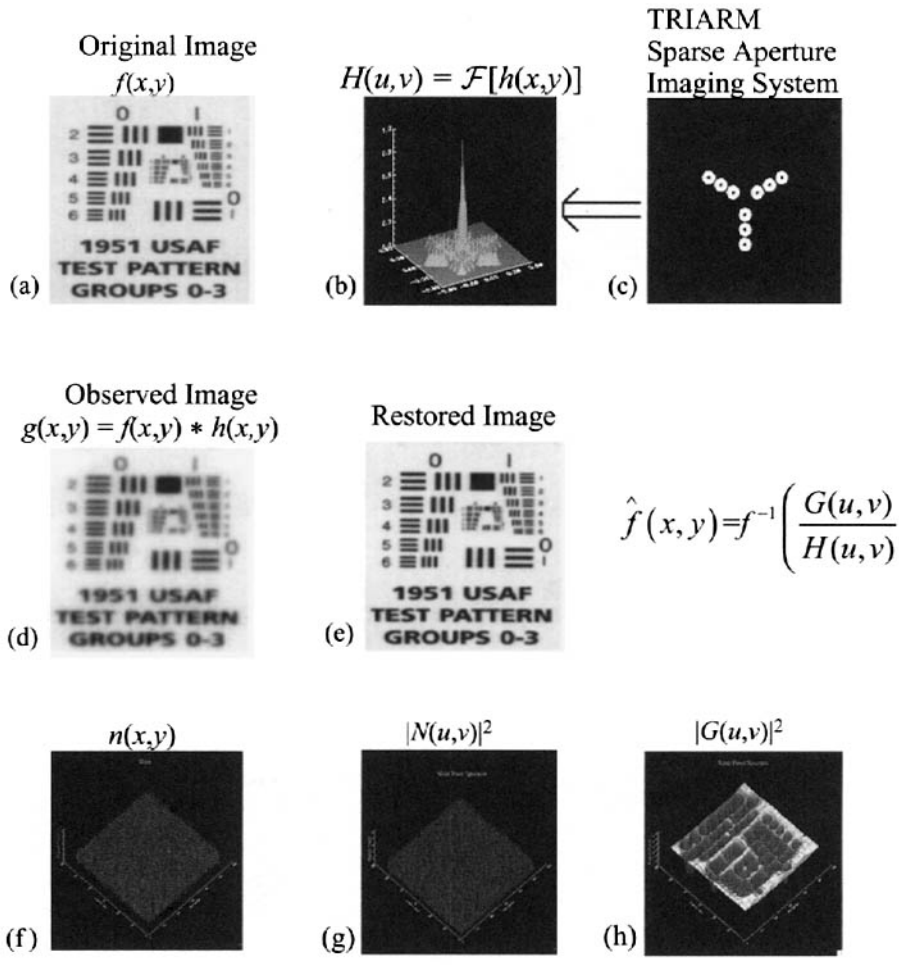


Figure 8.16 Illustration of image restoration concepts. (a) Original undegraded image (b) Frequency domain representation of the impulse response resulting from imaging with a telescope with a partially filled aperture illustrated in (c). (d) Simulation of the observed image. (e) Restoration of the blurred image shown in (d) using only the impulse response in (b), showing that full restoration is achieved. (i) The same image as shown in (d) with the noise shown in (f) added. (j) The restored image showing the failure of the ideal inverse filter in the presence of noise. (k) The results of an iterative Weiner-Helstrom filter produced using the impulse response shown in (b) along with the noise power spectrum shown in (g) and the observed image power spectrum shown in (h).

$$G(u,v) = F(u,v) \cdot H(u,v) \tag{8.37}$$

where $H(u,v)$ is the Fourier transform of the PSF, the so-called *transfer function*. The reciprocal of the transfer function $1/H(u,v)$ is the ideal *inverse filter*; it is applied to the spectrum of the observed image to yield

$$\frac{1}{H(u,v)} \cdot G(u,v) = \frac{F(u,v) \cdot H(u,v)}{H(u,v)} = \hat{F}(u,v) \tag{8.38}$$

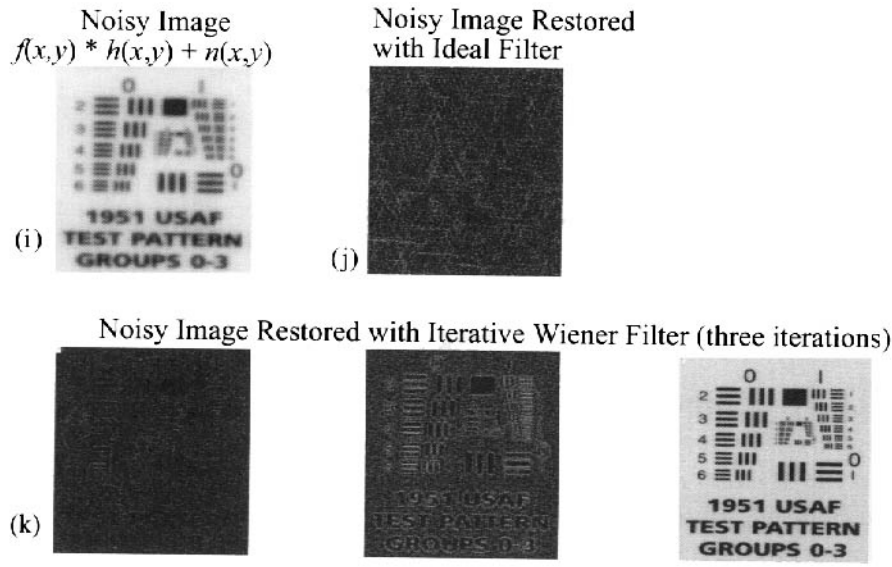


Figure 8.16 continued.

The Fourier transform of this expression ($\hat{F}$) yields the (restored) image $\hat{f}(x,y)$. Note that this can only be performed if $H(u,v) \neq 0$ at all frequencies (u,v). If we know the PSF and if the input image is noise free, we can fully restore the original image as illustrated in Figure 8.16. In the real world, we seldom have access to anything resembling a noise-free image, and a more appropriate expression for the observed image is

$$g(x,y) = f(x,y) * h(x,y) + n(x,y) \tag{8.39}$$

or, equivalently,

$$G(u,v) = F(u,v) \cdot H(u,v) + N(u,v) \tag{8.40}$$

where $n(x,y)$ is the additive noise image and $N(u,v)$ is its Fourier transform. In many cases, we can treat the noise as approximately additive (cf. the noise discussion in Sec. 5.4). In the presence of such noise, at frequencies where $H(u,v)$ is small and $N(u,v)$ is not (this commonly occurs at high frequencies), the ideal inverse filter $[1/H(u,v)]$ of Eq. (8.38) exaggerates the noise and fails, as shown in Figure 8.16(j). The solution to this problem was demonstrated by Wiener (1949) for the analog case where the impulse response approximates a delta function, and by Helstrom (1967) for the more general case. The solution modifies the inverse filter to account for the relative magnitude of the noise in the frequency domain compared to the power in the original image in the frequency domain. The resulting operator, referred to as a *Wiener-Helstrom filter*, is

$$W(u,v) = \frac{1}{H(u,v)} \cdot \frac{|H(u,v)|^2}{|H(u,v)|^2 + \frac{[N(u,v)]^2}{|F(u,v)|^2}} \tag{8.41}$$

As discussed by Helstrom (1967), this expression is optimal in a least-squares sense; i.e., it minimizes the squared error:

$$\varepsilon = \sum_x \sum_y [(f, x)y - \hat{f}(x,y)]^2 \tag{8.42}$$

where $\hat{f}(x,y)$ is the restored image resulting from a linear operation on the observed image $[g(x,y)]$. Inspection of Eq. (8.41) shows that the Wiener-Helstrom filter reduces to the inverse filter in the absence of noise (i.e., $N(u,v)=0$). More important, we can think of the ratio of the noise power spectrum over the image power spectrum as a modulator on the inverse filter that reduces the impact of the filter at frequencies where the ratio is large (i.e., where noise dominates).

Clearly, there is a problem with implementing the Wiener-Helstrom filter in that it requires complete knowledge of the impulse response $[h(x,y)]$, the noise power spectrum [N(u,v)], and the original image power spectrum [F(u,v), or at least the ratio of the noise to image power spectrum]. In practice, we often do know the instrument's impulse response or point spread function and can estimate the noise power spectrum from a sample of the noise when observing a dark field or a uniform scene. However, we seldom know the power spectrum of the original image. Hillery and Chin (1991) suggest that the observed image power spectrum may be a good initial estimate and use $|G(u,v)|^2$ in Eq. (8.41) as an initial estimate of $|F(u,v)|^2$. Our first estimate of the restored image in the frequency domain is then

$$^1\hat{F}(u,v) = G(u,v) \cdot W(u,v) \tag{8.43}$$

where

$$W(u,v) = \frac{1}{H(u,v)} \cdot \frac{|H(u,v)|^2}{|H(u,v)|^2 + \frac{|N(u,v)|^2}{|G(u,v)|^2}} \tag{8.44}$$

Clearly, after this first iteration the original image power spectrum can be better estimated as $|^1F(u,v)|^2$ and a new estimate of the image generated using an updated Wiener-Helstrom filter. Figure 8.16 shows an example of using this iterative Wiener-Helstrom filter to restore an image corrupted with additive noise.

In some cases, the dominant noise source in an image is structured or periodic noise at a limited range of spatial frequencies. Periodic noise in the image can often be filtered out so that it does not detract from the original image or from a boosted version in cases where the noise is subtle. An example of this process is shown in Figure 8.17. In this case, neither the original image [Fig. 8.17(a)] nor a

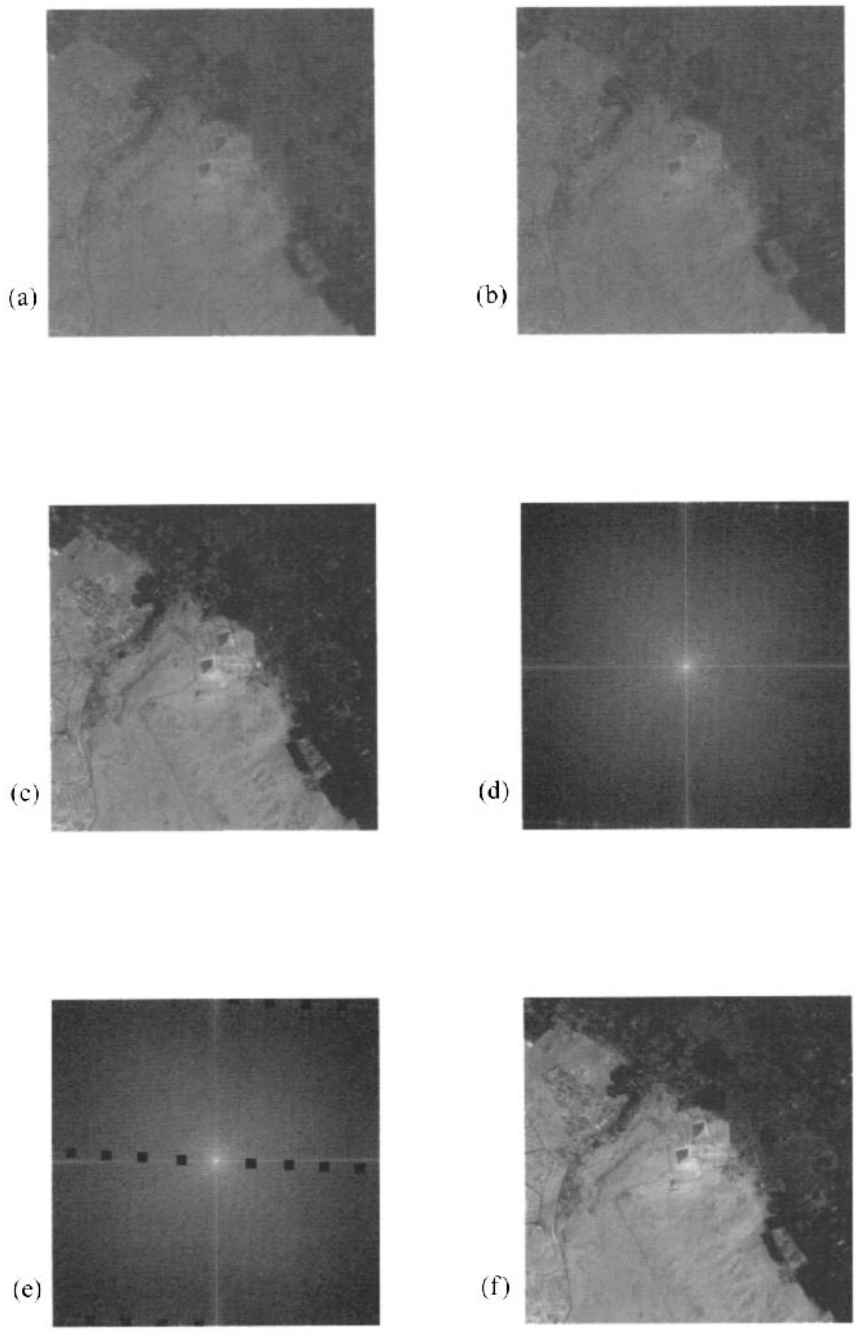


Figure 8.17 Filtering of noise in a SPOT image of the great pyramids: (a) original image, (b) contrast enhanced image, (c) contrast-enhanced version of (a) after high-frequency boost, (d) power spectrum of (a), (e) power spectrum showing where frequencies associated with periodic noise have been filtered, and (f) contrast-enhanced and high-frequency–boosted version of (a) after periodic noise filter was applied in the frequency domain.

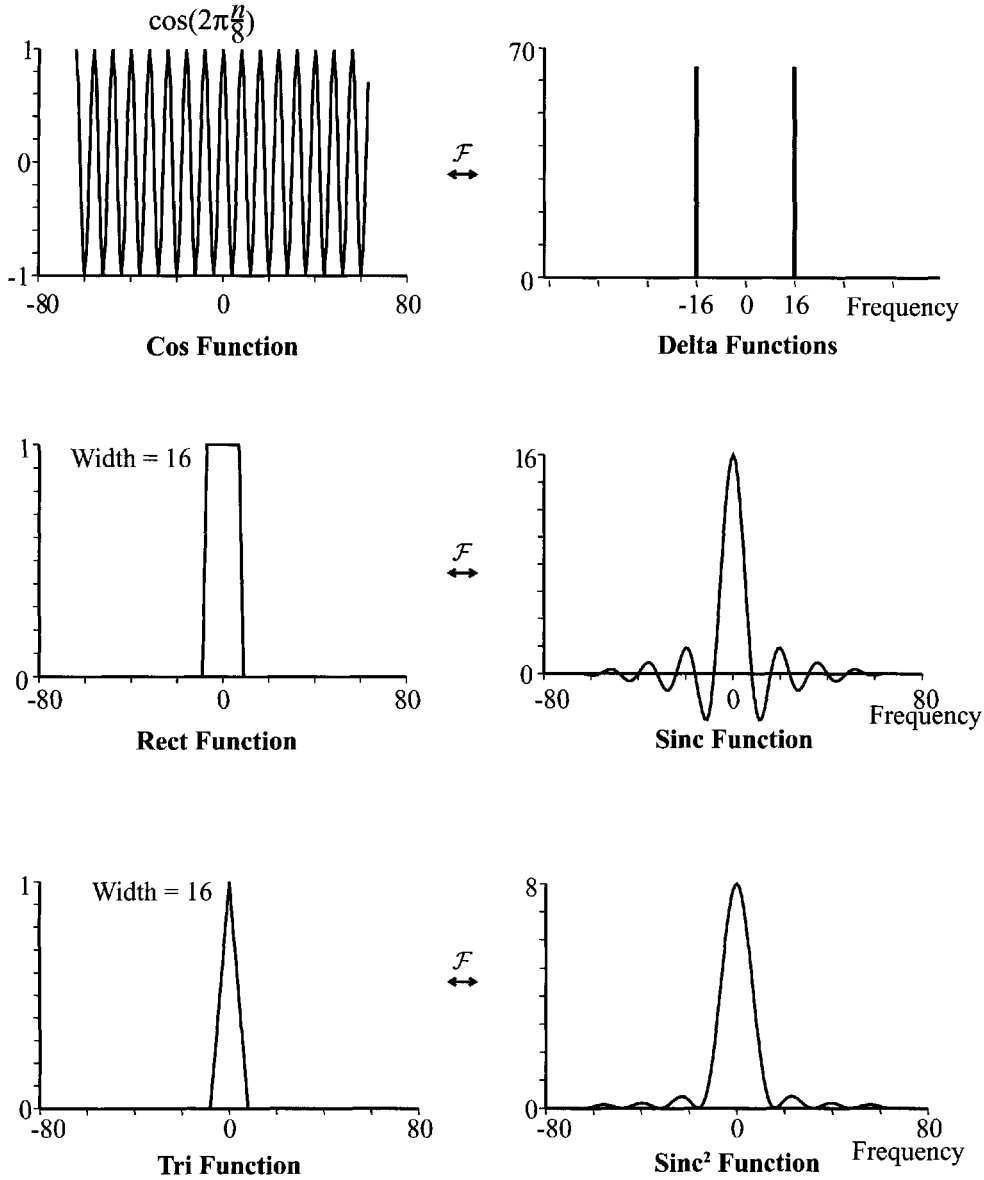


Figure 8.18 Fourier transform pairs.

contrast-enhanced version [Fig. 8.17(b)] shows any appreciable noise. However, when the original image is filtered with a boost filter to sharpen the edges and then the contrast is enhanced [Fig. 8.17(c)], periodic noise patterns become visible. These subtle periodic noise patterns are common in most EO imagery as a result of slight variations in detector-to-detector calibration, periodic noise sources in the processing electronics, or periodic noise in the readout and digitizing electronics. By transforming the original image into the frequency domain and then enhancing the contrast of the image of the magnitude of the Fourier transform, it is possible to identify the elevated energy content at certain frequencies that represent the unwanted noise [Fig. 8.17(d)]. By making a filter whose response tapers rather quickly to zero at these frequencies, it is possible to filter out the noise and leave the rest

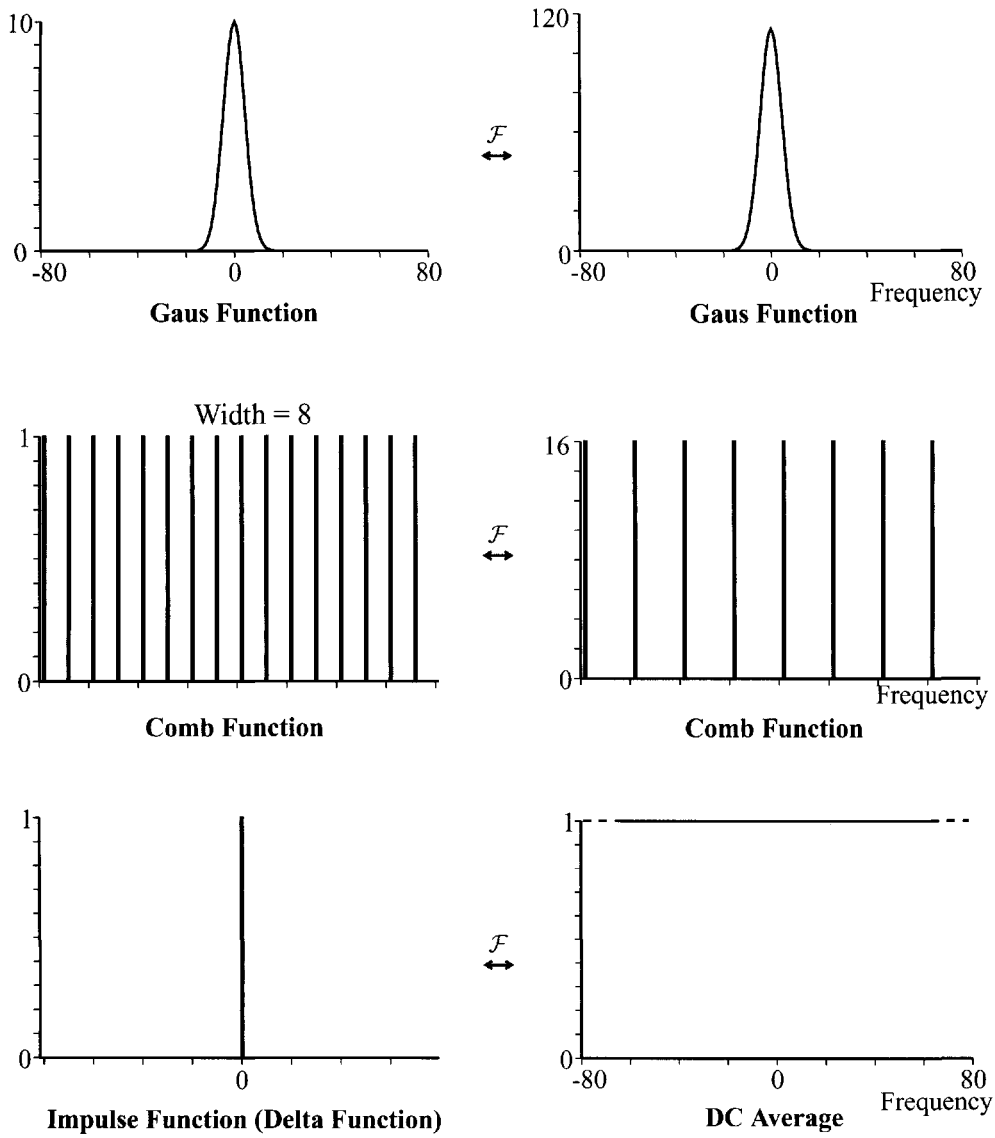


Figure 8.18 continued. Fourier transform pairs.

of the image largely unaffected [Fig. 8.17(e)]. When the boost filter is applied after the noise filter and the resulting image contrast stretched (after being transformed back to the spatial domain), the noise effects are removed [Fig. 8.17(f)].

The frequency domain operations (or their spatial domain equivalents (cf. Fig. 8.18) discussed in this section can be used to perform numerous image enhancement operations, including blurring, sharpening, noise filtering, edge definition, patterned noise isolation, and scene segmentation based on texture when coupled with other algorithms. In addition, one of their most useful functions is in modeling the performance characteristics of imaging systems relative to spatial image fidelity and in defining filters for reducing the apparent impact of spatial degradations along the imaging chain (cf. Sec. 13.1). This section on global operations and frequency-domain filters has just scratched the surface of the use of linear

systems theory in image processing. A much more in-depth treatment of many of the topics introduced here can be found in Gaskill (1978) and Gonzalez and Woods (2002).

This chapter has introduced the basics of image processing with an emphasis on processing of monochrome or single-band images. In Chapter 9, we will focus on processing of multispectral images with an emphasis on processing algorithms developed for analyzing remotely sensed data. Recognize that all algorithms introduced in this chapter can be applied to multiband images on a band-by-band basis. However, be cautioned that many of the multi- and hyperspectral algorithms rely on measurements of either the relative or absolute spectral radiances. The image processing algorithms presented in this chapter are often targeted at visual exploitation and will change the gray levels (apparent spectral radiance values) and, as a result, may adversely impact spectral analysis algorithms. As a result, most of these algorithms should be employed in parallel with or after the spectral algorithms.

8.6 REFERENCES

Bracewell, R.O. (1986). *The Fourier Transform and Its Applications*, 2nd ed., rev., McGraw-Hill, NY.

Ehrhard, D.G., Easton, R.L., Jr., Schott, J.R., & Comeau, M.J. (1993). Frequency-domain texture features for classifying SAR images. *Proceedings SPIE,* Vol. 1960, No. 3, pp. 21–32.

Fu, K.S. (1974). *Syntactic Methods in Pattern Recognition*. Academic, NY.

Galloway, M.M. (1975). Texture classification using gray level run lengths. *Comp. Graphics Image Proc.*, Vol. 4, pp. 172–179.

Gaskill, J.D. (1978). *Linear Systems, Fourier Transforms and Optics*. Wiley, NY.

Gonzalez, R.C., & Wintz, P. (1987). *Digital Image Processing*. 2nd. ed., Addison-Wesley, Reading, MA.

Gonzalez, R.C., & Woods, R.E. (2002). *Digital Image Processing*, 2nd. ed., Prentice-Hall, Upper Saddle River, NJ.

Goodman, J.W. (1996). *Introduction to Fourier Optics*, 2nd ed., McGraw-Hill, NY.

Haralick, R.M., Shanmugam, K., & Dinstein, I. (1973). Textural features for image classification. *IEEE Trans. Systems, Man, Cybernetics,* Vol. SMC-3, No. 6, pp. 610-621.

Hardie, R.C., & Boncelet, C.G. (1993). LUM filters: a class of rank-order-based filters for smoothing and sharpening. *IEEE Trans. Signal Processing*, Vol. 41, No. 3, pp. 1061–1076.

Helstrom, C.W. (1967). Image restoration by the method of least squares, *J. Opt. Soc. Am.*, Vol. 57, pp. 297-303.

Hillery, A.X., & Chin, R.T. (1991). Iterative Wiener filters for image restoration. *IEEE Trans. Signal Processing*, Vol. 39, No. 8, pp. 1892-1899.

Papoulis, A. (1962). *The Fourier Integral and Its Applications*. McGraw-Hill, NY.

Pratt, W.K. (1991). *Digital Image Processing*, 2nd ed., Wiley, NY.

Rosenblum, W. (1990). "Optimal Selection of Textural and Spectral Features for Scene Segmentation." Masters thesis, Center for Imaging Science, Rochester Institute of Technology, Rochester, NY.

Rosenfeld, A., & Kak, A.C. (1982). *Digital Picture Processing*. 2nd ed., Academic Press, Orlando, FL.

Schowengerdt, R.A. (1983). *Techniques for Image Processing and Classification in Remote Sensing*. Academic Press, Orlando, FL.

Stromberg, W.D., & Farr, T.G. (1986). "A Fourier-based textural feature extraction procedure." *IEEE Trans. Geosci. Remote Sensing*, Vol. GE-24, No. 5, pp. 722–731.

Swain, P.H. (1978). "Fundamentals of pattern recognition in remote sensing." In P.H. Swain and S.M. Davis, eds. *Remote Sensing: The Quantitative Approach*. McGraw-Hill, New York, NY.

Tou, J.T., & Gonzalez, R.C. (1974). *Pattern Recognition Principles*. Addison-Wesley, Reading, MA.

Wang, L., & He, D. (1990). A new statistical approach for texture analysis. *Photogrammetric Engin. Remote Sensing*, Vol. 56, No. 1, pp. 61-66.

Wiener, N. (1949). *Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications*, Wiley, NY.

This page intentionally left blank

CHAPTER 9
MULTISPECTRAL REMOTE SENSING ALGORITHMS: LAND COVER CLASSIFICATION

This chapter introduces a number of algorithms that are targeted at analysis of multispectral data. As part of that process, we introduce the basic vector-matrix notation and operations that will be used extensively in Chapters 9, 10, and 11. In fact, the division of the material into these three chapters is somewhat arbitrary. This chapter deals primarily with techniques that have been developed or used initially on multispectral data. However, many of them are also applicable to imaging spectrometer data and could be included in Chapter 10. Chapter 10 addresses a range of algorithms that have been applied most extensively to imaging spectrometer data. Many of these imaging spectrometer algorithms can also be applied to multispectral data if a sufficient number of bands exist. Chapter 11, the third chapter in this grouping, addresses a subset of algorithms that were also developed primarily for analyzing imaging spectrometer data. This subset consists of algorithms that draw heavily on physics-based models to support or constrain the algorithms.

Most of the algorithms and analysis techniques presented here were motivated by the ready access to many-band digital data that was ushered in by the Landsat program in the early 1970s. The data from Landsat, SPOT, AVHRR, and airborne multispectral line scanners made it possible for the first time to generate maps of the Earth's land cover at local to global scales. For several decades, much of the effort in the remote sensing field was focused on trying to understand the extent to which we could map land cover and then developing, implementing, and testing land cover classification procedures. This chapter emphasizes the basic land cover classification algorithms and data processing techniques that are the result of that effort. The success of the remote sensing community in this task and the

relatively operational use of land cover mapping have helped to enable the field of geographic information systems (GIS) discussed in Chapter 12. The reader interested in a more extensive treatment of multispectral image analysis should consult Schowengerdt (1997) and Richards (1999).

9.1 REVIEW OF MATRIX METHODS

This section is intended to provide a very cursory review of the linear algebra and matrix methods needed for Chapters 9, 10, and 11. In teaching the material in these chapters, I have often found it necessary to remind the students about the material presented in this section. I have therefore,included it here for those readers who need a reminder on some of these topics. A more complete discussion can be found in Johnson and Wichern (1998), Wylie and Barret (1995), and Press et al (1992). For those of you too caught up in the remote sensing plot to be bothered with this math lesson, you can skip this section—I will refer you back to the critical sections as needed. Furthermore, recognize that this review is intended to yield conceptual insight. In most cases, much more computationally efficient methods exist for performing the matrix algebra computations described here.

9.1.1 Vector Algebra

Let x be an $\ell \times 1$ dimension vector (using row, column designation to define the size of a vector or matrix, i.e., $\ell \times 1$ is read ℓ rows by 1 column). The *vector* $\mathbf{x}$ can then be expressed as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_\ell \end{bmatrix} \tag{9.1}$$

where $x_1, x_2, \cdots x_\ell$ are the scalar elements of the vector $\mathbf{x}$. For our studies, the vector $\mathbf{x}$ will often represent a pixel comprised of the scalar digital count, radiance, or reflectance values in each band. We will tend to use a nomenclature system where lowercase boldface letters ($\mathbf{x}$) are used to designate vectors and uppercase boldface letters indicate an $\ell \times$ m matrix ($\mathbf{X}$).

Multiplication of a vector ($\mathbf{x}$) by a scalar (c) yields

$$c\mathbf{x} = \begin{bmatrix} cx_1 \\ cx_2 \\ \vdots \\ cx_\ell \end{bmatrix} \tag{9.2}$$

Addition of two ($\ell \times 1$) vectors $\mathbf{x}$ and $\mathbf{y}$ yields an $\ell \times 1$ vector according to

$$\mathbf{z} = \mathbf{x} + \mathbf{y} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_\ell \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_\ell \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_\ell \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_\ell + y_\ell \end{bmatrix} \tag{9.3}$$

The combination of all scalar multiples and sums of vectors is called a *vector space.*

9.1.1.1 Linear Independence and Basis Vectors

A set of vectors are said to be *linearly dependent* if for k vectors

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots a_k\mathbf{x}_k = 0 \tag{9.4}$$

where the a_k values are not all zero; otherwise, the set is linearly independent. In more familiar form, if we can express a vector $\mathbf{x}_i$ in the following form:

$$\mathbf{x}_i = c_1\mathbf{x}_1 \tag{9.5a}$$

or

$$\mathbf{x}_i = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 \tag{9.5b}$$

then $\mathbf{x}_i$ is linearly dependent on $\mathbf{x}_1$ [Eq. (9.5a)] or $\mathbf{x}_1$ and $\mathbf{x}_2$ [Eq. (9.5b)]. Any linear independent set of vectors is a *basis set* for the vector space because any vector $\mathbf{x}$ can be expressed as a linear combination of the basis vectors; i.e., for three dimensions

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + a_3\mathbf{x}_3 = \mathbf{x} \tag{9.6}$$

However, note that if the basis vectors are of dimension k > 3, then the vector space defined by the three basis vectors is really a three-dimensional hyperplane; i.e., it spans a three-dimensional *subspace* defined by a_1, a_2, and a_3. This can be seen in Figure 9.1, where we have two independent three-dimensional vectors and see that the two vectors span a two-dimensional plane (i.e. a linear combination of them can be constructed to yield any point in the two-dimensional plane).

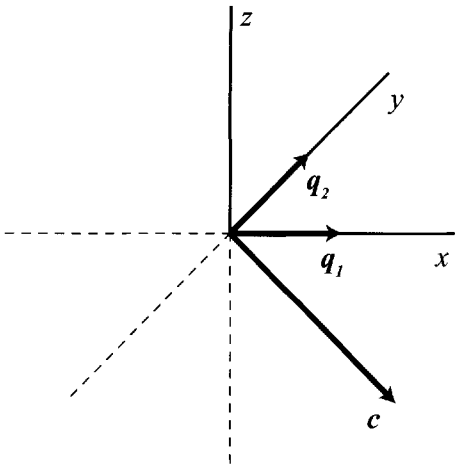
9.1.1.2 Orthogonality and Generation of Orthogonal Basis Vectors

The *transpose* of an $\ell \times 1$ column vector is a $1 \times \ell$ row vector, and vice versa, where the entry in the first row becomes the entry in the first column, etc.; i.e., the *transpose* of the vector $\mathbf{x}$ is

$$\mathbf{x}^T = [x_1, x_2 \cdots x_\ell] \tag{9.7}$$

The *inner* or *dot product* of two vectors is a scalar found by

$$\mathbf{x}^T\mathbf{y} = [x_1y_1 + x_2y_2 + \cdots x_\ell y_\ell] \tag{9.8}$$



$$q_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \qquad q_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

Any vector c in the (x,y) plane can be expressed as

$$c = d_1 q_1 + d_2 q_2$$

Figure 9.1 Illustration of how linear independent vectors define a subspace spanned by linear combinations of the vectors. In this case, the two-dimensional x, y plane is spanned by linear combinations of the vector q_1 and q_2.

and the dot product of a vector $\mathbf{x}$ with a unit vector ($\mathbf{y}_u$) is the scalar magnitude of the projection of $\mathbf{x}$ onto the direction defined by $\mathbf{y}_u$ where a unit vector is a vector of unit magnitude, i.e.,

$$\|\mathbf{y}_u\| = \left(\mathbf{y}_u^T \mathbf{y}_u\right)^{\frac{1}{2}} = 1 \tag{9.9}$$

Any vector can be converted to a unit vector by dividing by its *magnitude* or *length* $\|y\|$

$$\mathbf{y}_u = \frac{\mathbf{y}}{\|\mathbf{y}\|} = \frac{\mathbf{y}}{\left(\mathbf{y}^T \mathbf{y}\right)^{\frac{1}{2}}} \tag{9.10}$$

Two vectors are *orthogonal* (perpendicular) if their inner or dot product is zero, i.e.,

$$\mathbf{x}^T \mathbf{y} = x_1 y_1 + x_2 y_2 + \cdots x_\ell y_\ell = 0 \tag{9.11}$$

Given a set of k linear independent vectors, we can construct a set of k orthogonal vectors that span the same space using the Gram-Schmidt orthogonalization process. Let $\mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_k$ be the linear independent vectors; then, the orthogonal vectors are sequentially constructed as follows:

$$\mathbf{u}_1 = \mathbf{x}_1 \tag{9.12}$$

$$\mathbf{u}_2 = \mathbf{x}_2 - \frac{\mathbf{x}_2^T \mathbf{u}_1}{\mathbf{u}_1^T \mathbf{u}_1} \mathbf{u}_1 \tag{9.13}$$

$$\mathbf{u}_k = \mathbf{x}_k - \frac{\mathbf{x}_k^T \mathbf{u}_1}{\mathbf{u}_1^T \mathbf{u}_1} \mathbf{u}_1 - \cdots - \frac{\mathbf{x}_k^T \mathbf{u}_{k-1}}{\mathbf{u}_{k-1}^T \mathbf{u}_{k-1}} \mathbf{u}_{k-1} \tag{9.14}$$

For convenience, we often scale the Gram-Schmidt vectors to unit magnitude by dividing by their length according to

$$z_i = \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|} \tag{9.15}$$

Note this would lead to $\mathbf{u}_2 = \mathbf{x}_2 - [\mathbf{x}_2^T \mathbf{z}_1]\mathbf{z}_1$. This process is illustrated in Figure 9.2.

9.1.2 Matrix Algebra

A ($m \times n$) *matrix* **M** is an array of scalars with m rows and n columns. The *transpose* of an $m \times n$ matrix is an $n \times m$ matrix where the first row of the transposed matrix is the first column of the original, second row is the second column, etc. If we let

$$\mathbf{M} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix} \tag{9.16}$$

then the transpose of **M** is

$$\mathbf{M}^T = \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \end{bmatrix} \tag{9.17}$$

The product of two matrices **A** and **B** of dimension ($m \times n$) and ($n \times p$) respectively is a $m \times p$ matrix where each element (i,j) is the dot product of the ith row of **A** with the jth column of **B**, e.g.:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & \cdots \\ \vdots & \ddots \end{bmatrix} \tag{9.18}$$

The *determinant* of a scalar is the scalar value. The determinant of a 2×2 matrix is given by

$$|\mathbf{A}| = a_{11}a_{22} - a_{12}a_{21} \tag{9.19}$$

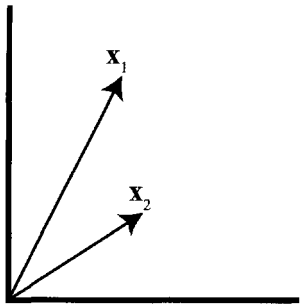
In general, for a $k \times k$ matrix,

$$|\mathbf{A}| = \sum_{j=1}^{k} a_{1j} |\mathbf{A}_{1j}| (-1)^{i+j} \tag{9.20}$$

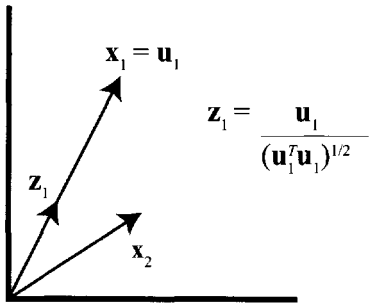
where $|\mathbf{A}_{1j}|$ is the determinant of the reduced matrix obtained by eliminating the first row and the jth column.

The *rank* of a matrix is the maximum number of linearly independent rows or columns (they will be equal; i.e., row rank = column rank). A square matrix is *nonsingular* if its rank equals the number of rows or columns; otherwise, it is *singular*. An alternate definition is that a matrix (**A**) will be nonsingular if

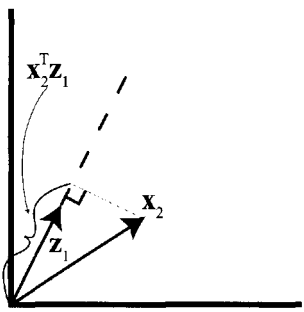
$$\mathbf{Ax} = 0 \tag{9.21}$$



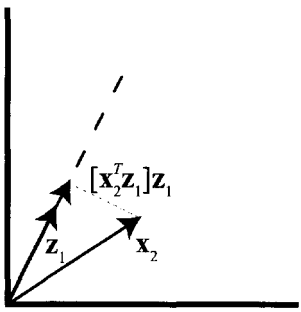
(a) Given a set of linear independent vectors $\mathbf{x}_1$, $\mathbf{x}_2$.



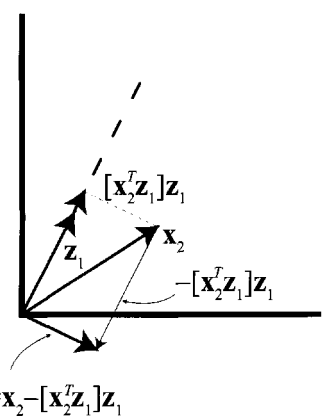
(b) Choose $\mathbf{x}_1$ to be the first Gram-Schmidt vector and normalize to unit length.



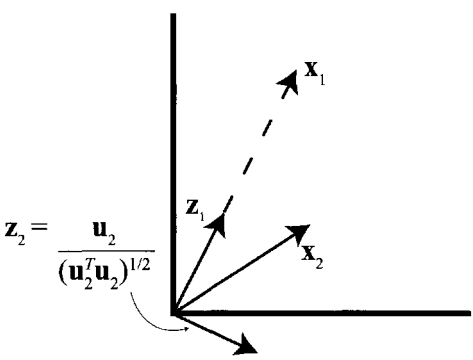
(c) Project $\mathbf{x}_2$ onto $\mathbf{z}_1$.



(d) Multiplication by $\mathbf{z}_1$ yields a vector of magnitude $[\mathbf{x}_2^T\mathbf{z}_1]$ in the $\mathbf{z}_1$ direction.



(e) Subtract the new vector from $\mathbf{x}_2$ to yield $\mathbf{u}_2$.



(f) Normalize $\mathbf{u}_2$ to unit length.

Figure 9.2 Illustration of the Gram-Schmidt orthoganalization process for a two-dimensional case.

is true only if $\mathbf{x} = 0$. Note that the converse would require $\mathbf{A}$ to be composed of some linearly dependent vectors and, therefore, of less than *full rank.*

If $\mathbf{A}$ is a nonsingular square matrix, then its *inverse* $\mathbf{A}^{-1}$ exists and has the property

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{I} = \mathbf{A}^{-1}\mathbf{A} \tag{9.22}$$

where $\mathbf{I}$ is the *identity* matrix with ones along the primary diagonal and zeros elsewhere. Note that

$$\mathbf{A}\mathbf{I} = \mathbf{A} \tag{9.23}$$

The inverse of a 2×2 matrix $\mathbf{A}$ is

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|}\begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix} \tag{9.24}$$

Note that $|\mathbf{A}| \neq 0$ must be true for the inverse to exist. In general, if $\mathbf{B}$ is the inverse of $\mathbf{A}$, then the row i, column j entry in $\mathbf{B}$ is given by

$$b_{ij} = \frac{1}{|\mathbf{A}|}|\mathbf{A}_{ji}|(-1)^{1+j} \tag{9.25}$$

where $\mathbf{A}_{ji}$ is the matrix obtained by deleting the jth row and ith column of $\mathbf{A}$. Note that matrix $\mathbf{A}$ is nonsingular if $|\mathbf{A}| \neq 0$ or equivalently if $\mathbf{A}$ has an inverse.

The *trace* of a square matrix is the sum of the values along the primary diagonal; i.e., if $\mathbf{A}$ is of dimension $k \times k$,

$$\text{Trace}(\mathbf{A}) = \sum_{i=1}^{k} a_{ii} \tag{9.26}$$

A square matrix $\mathbf{A}$ is said to be orthogonal if its rows (or columns) form an *orthonormal basis* set. This means its row (or columns) are orthogonal to each other and of unit length. Thus, the product of an orthogonal matrix with itself is the identity matrix:

$$\begin{aligned} &\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I} \\ &\text{so } \mathbf{A}^{-1} = \mathbf{A}^{\mathrm{T}} \end{aligned} \tag{9.27}$$

9.1.3 Eigenvectors and Singular Value Decomposition (SVD)

The *eigenvectors* and *eigenvalues* of a matrix are also referred to as the *characteristic vectors* and *characteristic roots*. A $(k \times k)$ matrix A has a $(k \times 1)$ characteristic vector $\mathbf{e}$ and a corresponding (1×1) characteristic root λ if

$$\mathbf{A}\mathbf{e} = \lambda\mathbf{e} \tag{9.28}$$

or, equivalently,

$$\mathbf{A}\mathbf{e} = \lambda\mathbf{I}\mathbf{e} \tag{9.29}$$

Rearranging yields

$$\mathbf{Ae} - \lambda\mathbf{Ie} = [\mathbf{A} - \lambda\mathbf{I}]\mathbf{e} = 0 \tag{9.30}$$

For this to be true in the nontrivial case (i.e., $\mathbf{e} \neq 0$), the columns of $[\mathbf{A}\text{-}\mathbf{I}\lambda]$ must be linearly dependent, i.e.

$$col_1[\mathbf{A} - \mathbf{I}\lambda]\mathbf{e}_1 + col_2[\mathbf{A} - \mathbf{I}\lambda]\mathbf{e}_2 + \cdots col_k[\mathbf{A} - \mathbf{I}\lambda]\mathbf{e}_k = 0 \tag{9.31}$$

where col_i is the ith column vector. This implies that

$$|\mathbf{A} - \mathbf{I}\lambda| = 0 \tag{9.32}$$

Equation (9.32) can be expanded into a polynominal expression in λ, for example in the (2×2) case, we have

$$\begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = 0 \tag{9.33}$$

yielding

$$(a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} = 0 \tag{9.34}$$

which yields two roots for λ. In general, a $(k \times k)$ matrix will yield a kth- order polynominal and k roots or eigenvalues.

For example, let

$$\mathbf{A} = \begin{bmatrix} 2 & 0 \\ 2 & 3 \end{bmatrix} \tag{9.35}$$

then

$$|\mathbf{A} - \mathbf{I}\lambda| = 0 = \begin{vmatrix} 2 - \lambda & 0 \\ 2 & 3 - \lambda \end{vmatrix} \tag{9.36}$$

or

$$(2 - \lambda)(3 - \lambda) = 0 \tag{9.37}$$

yielding $\lambda_2 = 2, \lambda_1 = 3$ where we have adopted the convention of labeling the largest eigenvalue 1, next largest 2, and so on. To solve for the eigenvectors, we simply substitute each λ_i into an equation of the form of Eq. (9.28), i.e.,

$$\mathbf{Ae}_1 = \lambda_1\mathbf{e}_1 = \begin{bmatrix} 2 & 0 \\ 2 & 3 \end{bmatrix}\begin{bmatrix} e_{11} \\ e_{12} \end{bmatrix} = \begin{bmatrix} \lambda_1 e_{11} \\ \lambda_1 e_{12} \end{bmatrix} = \begin{bmatrix} 3e_{11} \\ 3e_{12} \end{bmatrix} \tag{9.38}$$

$$\mathbf{Ae}_2 = \lambda_2\mathbf{e}_2 = \begin{bmatrix} 2 & 0 \\ 2 & 3 \end{bmatrix}\begin{bmatrix} e_{21} \\ e_{22} \end{bmatrix} = \lambda_2\begin{bmatrix} e_{21} \\ e_{22} \end{bmatrix} = \begin{bmatrix} 2e_{21} \\ 2e_{22} \end{bmatrix} \tag{9.39}$$

yielding for Eq. (9.38)

$$2e_{11} = 3e_{11} \tag{9.40}$$

$$2e_{11} + 3e_{12} = 3e_{12} \tag{9.41}$$

e_{11} must equal 0 and e_{12} can take on any value, so we choose 1 for convenience, yielding

$$\lambda_1 = 3, e_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{9.42}$$

For Eq. (9.39) we obtain

$$2e_{21} = 2e_{21} \tag{9.43}$$

$$2e_{21} + 3e_{22} = 2e_{22} \tag{9.44}$$

e_{21} can take on any value, so we choose 1 for convenience; then,

$$2(1) + 3e_{22} = 2e_{22} \tag{9.45}$$

or

$$e_{22} = -2 \tag{9.46}$$

yielding

$$\lambda_2 = 2, \mathbf{e}_2 = \begin{bmatrix} 1 \\ -2 \end{bmatrix} \tag{9.47}$$

In general, we are interested in the direction of the eigenvectors, not their magnitude, so we normalize the values, defining

$$\mathbf{e}_i = \frac{\mathbf{e}_i}{\left(\mathbf{e}_i^T \mathbf{e}_i\right)^{\frac{1}{2}}} \tag{9.48}$$

In general, for matrices larger than (3×3), the polynominal expression approach for solving for eigenvalues becomes painfully prohibitive, and we will use more efficient computational methods that are beyond the scope of the current discussion [cf. Press et al. (1992)].

For a real valued *symmetric matrix* where each column is equal to the corresponding row such that

$$\mathbf{A}^{\mathrm{T}} = \mathbf{A} \tag{9.49}$$

the eigenvectors are not only real, but also orthogonal, and since we have normalized them, they form an orthonormal basis set. As we will see in the next section, the spectral covariance matrix for a data set is real and symmetric and will become an important descriptor for much of our spectral analysis. The eigenvectors formed from such a matrix will form an orthogonal basis set that spans the spectral space of the original data.

A (k×k) symmetric matrix can be reconstructed from its eigenvectors and eigenvalues using

$$\mathbf{A} = \sum_{i=1}^{k} \lambda_i \mathbf{e}_i \mathbf{e}_i^{\mathrm{T}} \tag{9.50}$$

This shows us the particular importance of the magnitude of the eigenvalues in the *decomposition* (or reconstruction) of the matrix **A**. For example, as the λ_i values become very small, they will add essentially no value, and the matrix can be reconstructed to a good approximation without the corresponding eigenvectors.

The *singular value decomposition* (*SVD*) can be used to decompose a nonsquare matrix **A** into a sum of components expressed as

$$\mathbf{A} = \mathbf{U}\Lambda\mathbf{V}^{\mathrm{T}} \tag{9.51}$$

where **A** is an $m \times k$ matrix of real numbers, **U** is a $m \times m$ orthogonal matrix, **V** is a $k \times k$ orthogonal matrix, and Λ is an $m \times k$ matrix with r singular values $\lambda_i > 0$ for r less than or equal the smaller of m or k and zeros elsewhere. The non zero values of λ_i are called *singular values,* and their number r is equal to the rank of **A**. If we can solve for the singular values, we can express **A** as

$$\mathbf{A} = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}} = \mathbf{U}_r \Lambda_r \mathbf{V}_r^{\mathrm{T}} \tag{9.52}$$

where $\mathbf{U}_r = [\mathbf{u}_1, \mathbf{u}_2, \cdots \mathbf{u}_r]$ is the matrix made up of the $(m \times 1)$ orthogonal unit vectors in **U** corresponding to the non zero λ_i values, $\mathbf{V}_r$ is the matrix made up of the $k \times 1$ orthogonal unit vectors in **V** corresponding to the non zero λ_i values and Λ_r is the diagonal matrix of non zero λ_i singular values. **U** is made up of the eigenvectors $\mathbf{u}_i$ of $\mathbf{A}\mathbf{A}^T$ with corresponding eigenvalues λ_i^2 (note only r will be non zero), **V** is made up of the eigenvectors $\mathbf{v}_i$ of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ with the identical eigenvalues λ_i^2 (note the singular values are the square roots of the eigenvalues of the $\mathbf{A}\mathbf{A}^T$ or $\mathbf{A}^T\mathbf{A}$).

Similar to the eigenvector analysis described above, the SVD analysis can be used to generate a set of basis vectors suitable for spanning a spectral subspace. Consider a matrix **A** operating on a vector **x** to yield a vector **b** according to

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{9.53}$$

This can be thought of as a linear mapping (or projection) of the vector space **x** onto the vector space **b**. If **A** is singular, then there is a subspace of **x** called the null subspace that will map to zero, i.e.,

$$\mathbf{A}\mathbf{x} = 0 \tag{9.54}$$

There is also a subspace of **b** that can be spanned (reached) by **A** (i.e., some **x** will be mapped to **b**). This subspace is called the range of **A** and will have dimension r equal to the rank of **A** (i.e., r = the number of nonzero singular values of **A**). The SVD process generates orthonormal basis vectors that span both the range and the null space of **A**. The columns of **U** whose corresponding singular values are nonzero, make up an orthonormal set of basis vectors spanning the range of **A** (i.e., that portion that will map x to b). This means that linear combinations of vectors making up **U** will span the vector space. Note that the singular values are often

slightly nonzero due to numerical approximations, and a method to set small values to zero is often required.

In the rest of this chapter and those that follow, we will draw on the mathematical tools reviewed here to build spectral image analysis tools.

9.2 IMAGE CLASSIFICATION

One of the most common requirements in remote sensing is the need to segment or classify an image into land cover, material, or object classes. In some cases, this is an end in itself; in others, it is one step in a more involved process (cf. Sec. 9.4 on hierarchical classification). This section focuses on classification methods based on the spectral makeup of each pixel. In most cases, the classification is performed with a man-in-the-loop on uncalibrated imagery, so we will concentrate this section on techniques used in these cases. Unsupervised classification methods are discussed briefly in Section 9.2.3, and the use of calibrated or normalized data is discussed in Section 9.2.6. A complete derivation of the multivariate statistics behind the Gaussian maximum likelihood classifier emphasized in Section 9.2.2 is beyond the scope of this treatment [cf. Morrison (1967)]. To provide a stronger understanding of the multivariate process, we will explore a rigorous development of univariate classification methods in Section 9.2.1 and then extend that treatment, largely by analogy, to the multivariate (multispectral) case in Section 9.2.2.

9.2.1 Supervised Classification of a Single-Band Image

In this section we will assume that a single-spectral-band digital image is to be segmented into classes by material or land cover type. We are going to use what is usually referred to as a *supervised*, or *man-in-the-loop,* process where the user identifies a sample of pixels of each type or class, and digital processing algorithms are used to assign all similar pixels to one of the classes. To simplify the process, assume that the image is composed of k classes. For clarity in the diagrams, we will set k equal to 3. We can think of this process as teaching or "training" the computer algorithms what each class "looks" like, and then using some attributes of the class to classify other pixels into the class they most "look" like. These single-band approaches are seldom used in practice because of the strong overlap between classes in the single-band case. Nevertheless, that treatment can add significant insight to the multivariate approaches to follow.

9.2.1.1 Univariate Minimum Distance to the Mean

The simplest way to characterize what a group of pixels "looks" like in a panchromatic or single-band image is by their mean digital count. Figure 9.3 shows the results of selecting several pixels from each class and computing the mean gray value of each class using

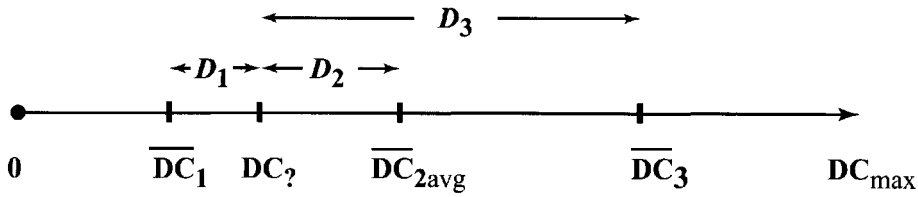


Figure 9.3 Use of the minimum distance to the mean classifier in a three-class single-band image classification problem. Using the minimum distance to the mean classifier, $DC_?$ is assigned to class 1 since D_1 is the smallest.

$$\overline{DC_i} = \frac{\sum_{q=1}^{N} DC_{iq}}{N} \tag{9.55}$$

where $\overline{DC_i}$ is the mean of the ith class, DC_{iq} is the digital count for the qth sample in the ith class, and N is the total number of pixels in the sample set for that class. We have now "trained" the computer algorithm to recognize what each class "looks" like as defined by the mean digital count of the samples. Any other pixel $DC_?$ must then be compared to the means to see which one it most "looks" like. In the simplest case, this can be defined by selecting the class whose mean digital count ($\overline{DC_i}$) is the minimum distance (D) from the pixel to be classified $DC_?$ as defined by

$$D_i = \left[\left(\overline{DC_i} - DC_? \right)^2 \right]^{\frac{1}{2}} \tag{9.56}$$

where D_i is the distance of the pixel from the mean of the ith class. The pixel is assigned to the class (i) for which the D_i value is a minimum (hence the name of the method: *minimum-distance-to-the-mean* or MDM). In the example shown in Figure 9.3, the unknown pixel would be assigned to class 1. In this way, every pixel in the scene can be classified. It is also possible to use threshold values so that pixels too far from any class mean are assigned to a class of unidentified pixels.

9.2.1.2 Simple Statistical Measures of Distance

A limitation of the MDM method is illustrated in Figure 9.4. In this case, we see not only the mean of the sampled values but also a normalized histogram [$p(DC|i)$] of the sampled values for each class and a measure of the variation in the sampled values in the form of the standard deviation from the class mean (σ_i). In this case, we see that pixels belonging to class 1 show very little variation from the mean compared to classes 2 or 3. A pixel having the same digital count $DC_?$ as used in the minimum distance to the mean example would intuitively be classified not into the closest class (class 1 in this case) but into the class that has the higher *conditional probability*. The normalized histogram is referred to as the conditional probability [$p(DC|i)$] because it describes the probability of a digital count (DC) occurring

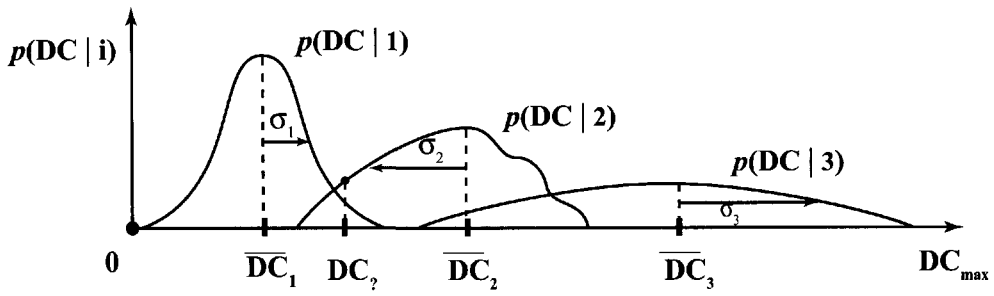


Figure 9.4 Use of simple statistical distances to develop classifiers that are sensitive to variance in the data.

subject to the condition that we are sampling from the *i*th class. Large sample sizes are required to characterize the shape of the normalized histogram. As a result, the shape is often assumed to be approximately Gaussian. Then the standard deviation, together with the mean, define the shape of the normalized histogram. Thus assigning the pixel to the class with the highest conditional probability [$p(DC|i)$] can be approximated by assigning the sampled pixel to the class whose mean is the fewest number of standard deviations away from the gray value of the unknown pixel. This simple statistical distance can be expressed as

$$D_{\sigma i} = \frac{\left[\left(\overline{DC_i} - DC_?\right)^2\right]^{\frac{1}{2}}}{\sigma_i} \tag{9.57}$$

where σ_i is the simple standard deviation of the samples belonging to class *i*. This classifier agrees with our intuitive reaction to data with different variances, stating that we assign the pixel to the class from which it is separated by the fewest number of standard deviations (class 2 in the example in Fig. 9.4). Thus, by accounting for variance within the class, we have introduced a statistical measure of distance that should improve our classification accuracy relative to the simple Euclidean distance metric used in the MDM method.

As shown in Figure 9.4, the use of the conditional probabilities (or an approximation based on the number of standard deviations) to make classification decisions makes intuitive and logical sense when the number of pixels in each class is approximately the same. However, if we anticipate that the number of pixels per class will differ significantly across classes, we might want to adjust our classification algorithm. For example, assume that we know our example scene contains roughly equal numbers of forest and urban pixels but eight times as many water pixels. This foreknowledge of the probability of a randomly sampled pixel being in a given class is called the *a priori probability* [$p(i)$]. In our case, the a priori probabilities are 0.8, 0.1, and 0.1 for the water, forest, and urban classes, respectively. This means that if we want to plot the relative heights of the unnormalized histograms rather than the normalized histograms in Figure 9.4, we would multiply each curve by the a priori probability. In our case, since we are only interested in

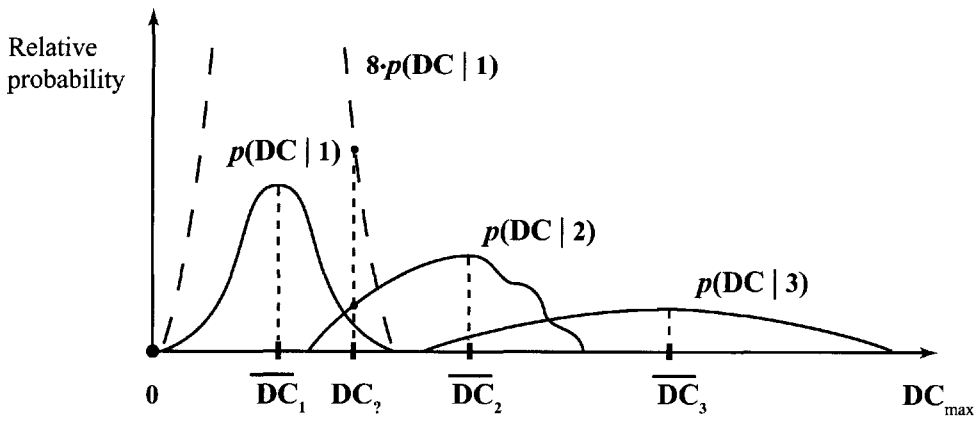


Figure 9.5 Effect of a priori probabilities on class decisions. Use of the a priori probabilities will change our decision about what class pixels with a digital count of $DC_?$ belong to. In this case, many more pixels with digital count $DC_?$ are in class 1 than class 2, so we will be correct more often if we assign pixels with this digital count to class 1.

relative heights of the curves, we could achieve the same end by multiplying the conditional probability for water (class 1) by 8, leaving the others unchanged (cf. Fig. 9.5). This accounts for the fact that there are eight times more water pixels than forest or urban. From Figure 9.5, we see that there will be both class 1 and class 2 pixels with digital count $DC_?$. However, when we factor in the relative number of class 1 pixels, it is clear that a pixel with a digital count value of $DC_?$ is much more likely to be in class 1 than in class 2 or class 3.

9.2.1.3 Univariate Gaussian Maximum Likelihood Classifier

The use of a priori probabilities can be formalized using Bayesian probability theory to define the a posteriori probability that a pixel with digital count DC belongs in class i:

$$p(i|DC) = \frac{p(DC|i)p(i)}{p(DC)} \tag{9.58}$$

where $p(\mathrm{DC})$ is the probability of the digital count occurring anywhere in the image (i.e., it is the normalized histogram of the entire image). The classification process simply involves computing the a posteriori probability of the sampled digital count value $DC_?$ being in each class and assigning pixels with that digital count to the class that yields the maximum a posteriori (MAP) probability (cf. Fig. 9.6). In practice, $p(\mathrm{DC})$ is the same for all classes. As a result, $p(DC)$ scales all the a posteriori probabilities, but it will not change their rank ordering, and so for convenience it is not included in the classifier. The a priori probabilities can be obtained from general knowledge of scene content or from a previous class map of the region [in some cases, a simple MDM classifier is run first to estimate the $p(i)$ values]. If no information is known about a priori probabilities $p(i)$, they can simply be assumed

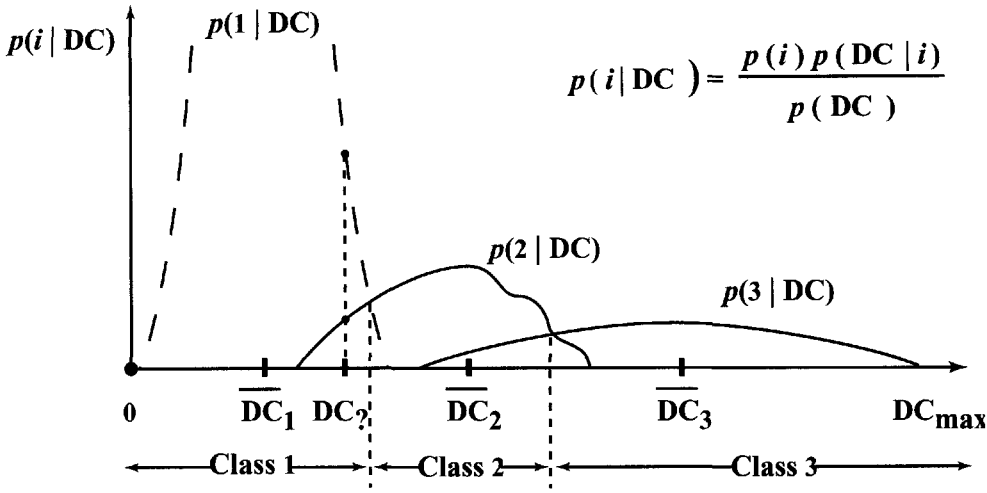


Figure 9.6 Application of probability theory to univariate classification. Using maximum likelihood theory, we would assign a pixel to whatever class has the highest a posteriori probability; $p(i|\mathrm{DC})$. In our case a pixel with a digital count of $\mathrm{DC}_?$ would be assigned to class 1.

to be equal. The conditional probabilities can be estimated from the normalized histograms generated from the training data. This presents a problem in terms of the amount of data to be stored in the algorithm and the amount of training data required to generate a good estimate of the true class histogram. In a univariate case (e.g., what we are considering here), the histogram is not too large (e.g., for an 8-bit image, it would only be 256 values for each class). However, in the multispectral cases we will consider shortly, these histograms are multidimensional and require massive amounts of data to accurately estimate their shape and to store the resulting estimate.

In cases where the shapes of the class histograms can be assumed to be approximately *Gaussian*, this process can be greatly simplified. The shape of the histogram, and therefore the conditional probability, can be estimated from the mean and the standard deviation of the training data according to

$$p(DC|\mathrm{i}) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(DC-\overline{DC_i})^2}{2\sigma_i^2}} \tag{9.59}$$

We can then define a class discrimination metric (D'_i) by combining Eqs. (9.59) and (9.58) and ignoring the $p(\mathrm{DC})$ term. This yields

$$D'_i = p(DC|i)\,p(i) = \frac{p(i)}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(DC-\overline{DC_i})^2}{2\sigma_i^2}} \tag{9.60}$$

For each pixel, the D'_i values for each class are computed, and the pixel is assigned to the class with the highest D'_i value. According to Eq. (9.58) and our analysis, this process will optimize the probability of correctly classifying the pixel

under the assumption of Gaussian distributed data (hence the name for this classifier, *Gaussian maximum likelihood* or *GML*). Note that, for simplicity, we can redefine the discriminant metric by computing the logarithm of D_i:

$$D_i'' = \ln D_i' = \ln[p(i)] - \frac{1}{2}\ln(2\pi) - \ln(\sigma_i) - \frac{(DC - \overline{DC_i})^2}{2\sigma_i^2} \tag{9.61}$$

or

$$D_i = \ln D_i' + \frac{1}{2}\ln 2\pi = \ln[p(i)] - \ln(\sigma_i) - \frac{(DC - \overline{DC_i})^2}{2\sigma_i^2} \tag{9.62}$$

Taking the natural logarithm and adding a constant will change neither the rank ordering of the discriminant metrics nor any resultant decisions about class assignment. Inspection of Eq. (9.62) indicates that only the last term is a function of the digital count and that this term is proportional to the magnitude of the squared value of the number of standard deviations from the pixel gray value to the mean. Thus, for a given class, reducing the number of standard deviations increases the discriminant value in agreement with our intuitively derived simple statistical-distance metric [Eq. (9.57)]. Furthermore, if the a priori probabilities are the same and the standard deviations are the same, Eq. (9.62) reduces to

$$D_i = A_0 - \frac{(DC - \overline{DC_i})^2}{2\sigma^2} \tag{9.63}$$

where σ is the common standard deviation for all classes, $A_0 = \ln(1/k) - \ln\sigma$ and k is the number of classes. In this case, D_i is maximized by minimizing the Euclidean distance (i.e., the problem reduces to an MDM classification).

In summary, if the data can be assumed to be Gaussian, the GML classifier described by Eq. (9.62) will minimize the number of misclassified pixels. In cases where the data are not Gaussian, we would have to use a discriminant function based on maximizing the a posteriori probabilities [cf. Eq. (9.58)] to minimize errors. While this is possible in the univariate case, it becomes operationally difficult to characterize effectively the conditional probabilities as the number of variables (spectral bands) increases. As a result, it is often useful to select classes whose brightness distributions are approximately Gaussian (e.g., a bimodal class could be split into two unimodal approximately Gaussian classes).

9.2.1.4 Risk or Loss Functions Used with Classifiers

To this point, we have assumed that our goal has been to minimize overall classification error. This may not always be the case. For example, consider the error regions indicated in Figure 9.7. In this example, class 1 is a target class that is for some reason more important than the other classes (e.g., a crop of high market value). From the perspective of class 1, the error regions can be separated into *errors of commission* (*false alarms*) and *errors of omission* (failure to identify a true target-false negative). From the perspective of class 2 (i.e., if class 2 is designated as

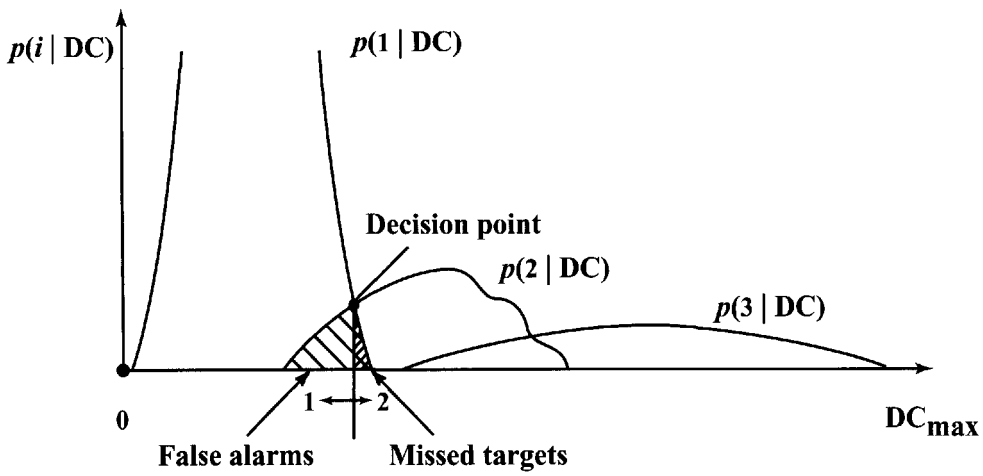


Figure 9.7 Adjustment of decision boundaries to control false alarms or misses.

the priority or target class), these error regions would be reversed. In a case where it is very critical to find all occurrences of a target (e.g., the target is toxic leachate from a waste dump), we might want to shift the decision point in Figure 9.7 to the right. This would increase the number of false alarms but decrease the number of misses. The overall effect would be to increase the number of misclassified pixels. Moving the decision point to the left would have the opposite effect of decreasing false alarms and increasing the number of targets missed. The overall increase in errors might be acceptable if there were a high cost associated with false alarms.

In scenarios such as these, we realize that we want to minimize not the number of misclassified pixels but the cost or loss associated with the overall classification. This can be done by assigning *loss functions* or weights to each type of classification [cf. Duda and Hart (1973)]. For example, let $l(i|j)$ be the loss or cost of assigning a pixel to the ith class when it actually belongs to the jth class. Then the risk or loss associated with assigning a pixel to class i will be related to its digital count by the conditional risk given by

$$R(i|DC) = \sum_{j=1}^{k} l(i|j)\, p(j|DC) \tag{9.64}$$

where k is the number of classes in the image. We minimize risk by assigning the pixel to the class that has the minimum conditional risk. In most cases, the most difficult part of this process is defining the loss matrix. In some cases, this is done based on an economic assessment, and dollar values can be used to define the cost of each type of misclassification. In other cases, costs are estimated in terms of relative loss or "pain" levels associated with different types of misclassification. Figure 9.8 shows a loss matrix for our simple three-class problem where we have defined the water (class 1) to have a high cost associated with false alarms (false positives) in case 1 and a high cost of missing a target (false negatives) in case 2. If we then choose to use one of these loss matrices to minimize our cost using Eq. (9.64), we will make more overall errors in classification but minimize our cost as

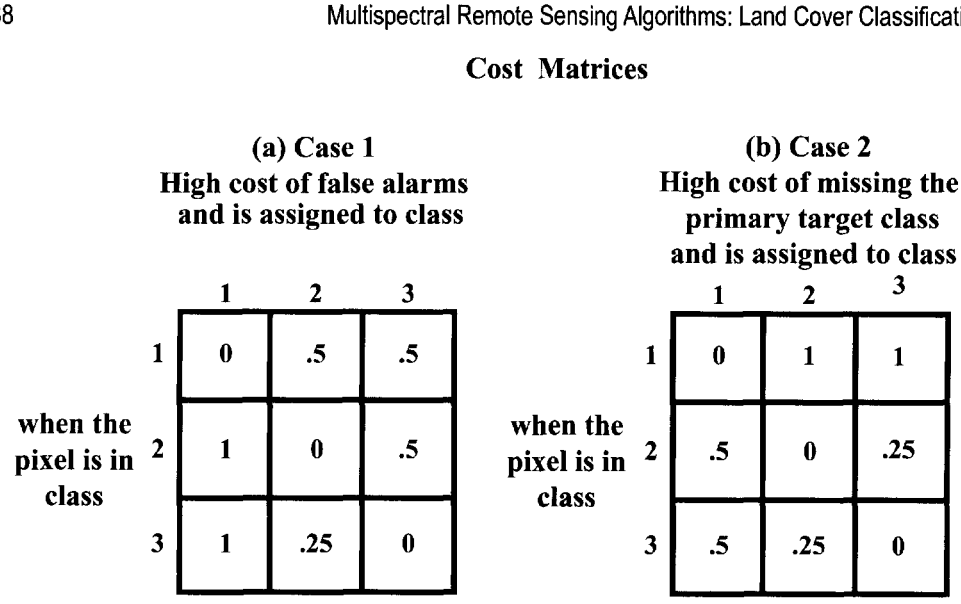


Figure 9.8 Example cost matrices for a three-class scenario: (a) Case 1 is for a high cost of false alarms, with class 1 as the target class. (b) Case 2 is for a high cost of missing the target class.

defined by the matrix. The final classification can be very sensitive to the relative magnitudes of the loss matrix, and the user should carefully evaluate how the loss values are defined. In many cases, there are too few data to develop good loss functions, and the misclassification error is simply minimized by maximizing the a posteriori probability.

9.2.2 Supervised Multispectral Image Classification

The single-band classification case described in the previous section is seldom used in practice because only limited class separability is achieved when only one spectral band is available. However, it is fairly straightforward to extend the principles to many spectral channels. If there are reasonably unique spectral signatures, then the classifiers can provide separation of classes. To keep the illustrations clear, we will generally illustrate only the two-spectral-band case, although the mathematical solutions will be in terms of ℓ spectral bands. Consider a set of training data illustrated in Figure 9.9, where the three classes are plotted with different symbols for two spectral bands. It is clear that if these values were projected onto either axis, discrimination using a single band would result in considerable overlap. Yet, in the two-band case, there appears to be fairly clear separability between the classes. The objective of this section is to define a set of classification algorithms that can take advantage of this apparent separability using multiple spectral bands.

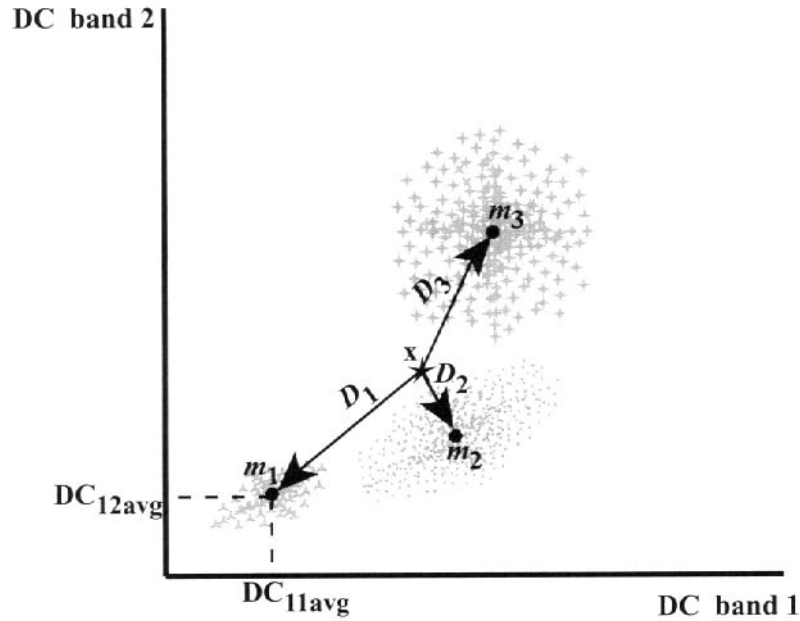


Figure 9.9 Multispectral classifications using a minimum distance to the mean classifier.

9.2.2.1 Multivariate Minimum Distance to the Mean Classifier

To begin, we define for each pixel a column vector (**x**) made up of its digital count values (DC) in each of ℓ spectral bands, i.e.

$$\mathbf{x} = \begin{bmatrix} DC_1 \\ DC_2 \\ \vdots \\ DC_\ell \end{bmatrix} \tag{9.65}$$

Similarly, the center of a cluster of multispectral sample points can be defined by the multivariate mean (**m**) for the class using a column vector composed of the mean digital count values in each band for samples from class (i) according to

$$\mathbf{m}_i = \begin{bmatrix} \overline{DC_1} \\ \overline{DC_2} \\ \vdots \\ \overline{DC_\ell} \end{bmatrix} \tag{9.66}$$

From Figure 9.9, we see that the simple *Euclidean distance* between **x** and $\mathbf{m}_i$ in a two-dimensional space is

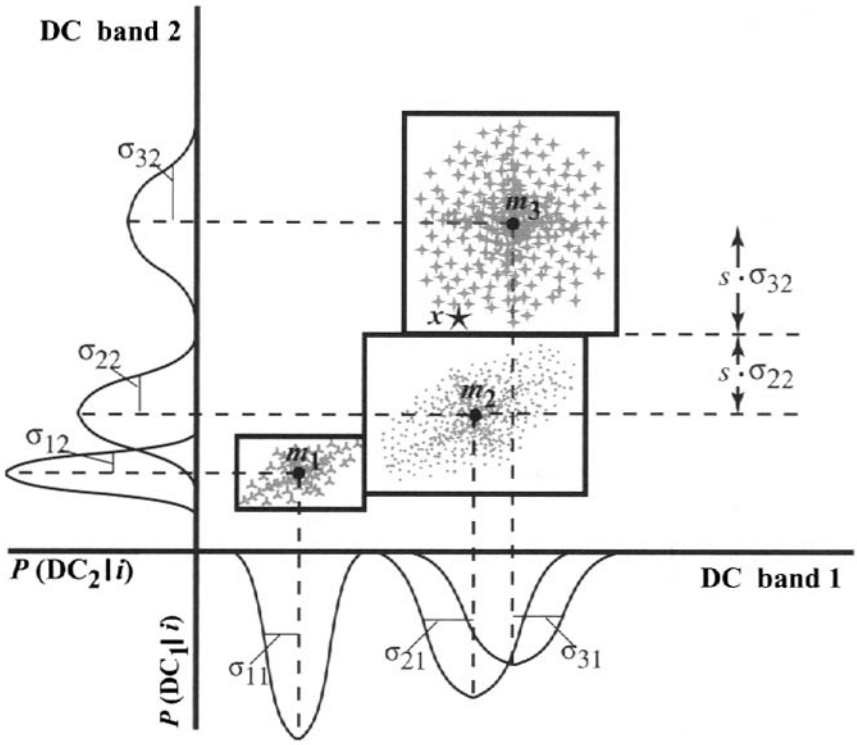


Figure 9.10 Parallelepiped classifier using equal numbers of standard deviations for the boundaries. To locate the boundary in band 2 between class 2 and class 3, we equalize the number of standard deviations ($s \cdot \sigma$) from the mean of each class to the boundary.

$$D_i = \left[(\mathbf{x} - \mathbf{m}_i)^2 \right]^{\frac{1}{2}} \tag{9.67}$$

We could then implement a multivariate minimum-distance-to-the-mean (MDM) classifier by computing D_i or, more practically, D_i^2 values for each pixel relative to the mean vector ($\mathbf{m}_i$) for each class and assign the pixel to that class with the minimum squared distance (D_i^2). This classifier, like its univariate equivalent, is very easy to implement and can be coded to run very rapidly on today's computers. However, it ignores the variability in the data as illustrated in Figure 9.10. The pixel **x** in Figure 9.10 would be classified into class 2 by an MDM classifier, yet our intuition would assign it to class 3 based on the high variability in class 3 and the relatively low variability in band 2 for class 2. Recall that in the univariate case, we used a simple statistical distance employing the number of standard deviations to account for variance.

9.2.2.2 Parallelepiped Classifiers

In the multivariate case, the variance in the data can be included in a classifier by locating boundaries in the classification space with parallelepipeds, as shown in Figure 9.10. Any vector that falls within the multi dimensional parallelepiped as-

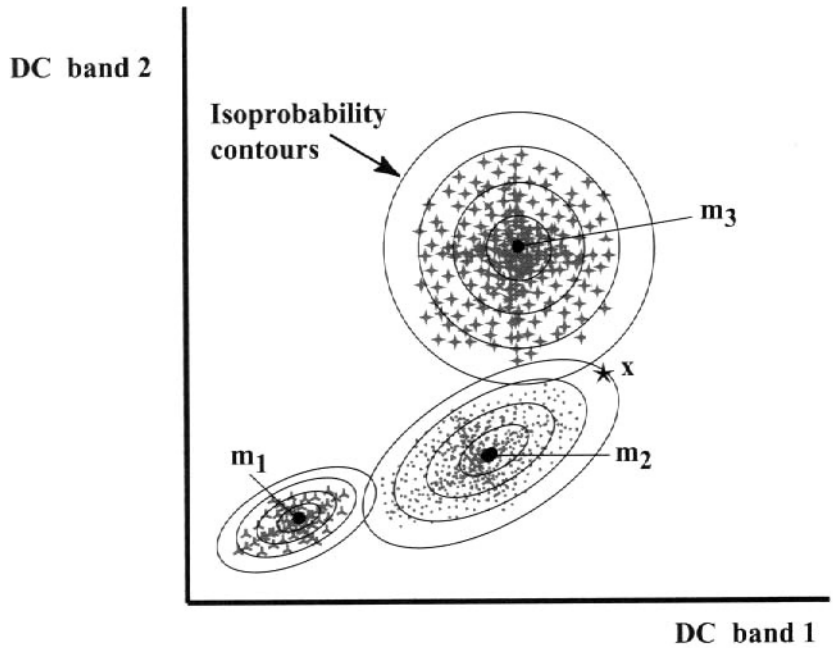


Figure 9.11 Classification using a multispectral Gaussian maximum likelihood classifier.

sociated with class i is assigned to that class. In two dimensions, these parallelepiped boundaries could be defined interactively on a computer screen. However, as the number of dimensions (spectral bands) increases, this approach becomes more cumbersome, as well as yielding unsystematic results. In general, some simple formula is used to define the boundary locations. One method that draws on the simple statistical distance is to place the parallelepiped boundaries at equal numbers of standard deviations from the means, as illustrated in Figure 9.10. The basic parallelepiped method does not account well for the spectral shape of the sample distributions, as illustrated in Figure 9.11. Here it is clear that the pixel x would be classified into class 3 by the parallelepiped method, but our intuition would indicate that the pixel is more likely a member of class 2. Our intuition is responding to the elongated shape of the class 2 data, which is described by the *correlation* between bands 1 and 2 for class 2. We would like to identify a classifier that accounts for the shape of the spectral distribution.

9.2.2.3 Multispectral Gaussian Maximum Likelihood

One way to include the shape of the class sample distribution is to use the probability theory introduced in Section 9.2.1. The a posteriori probability of a pixel with a spectral vector $\mathbf{x}$ belonging to class i is

$$p(i|\mathbf{x}) = \frac{p(i)p(\mathbf{x}|i)}{p(\mathbf{x})} \tag{9.68}$$

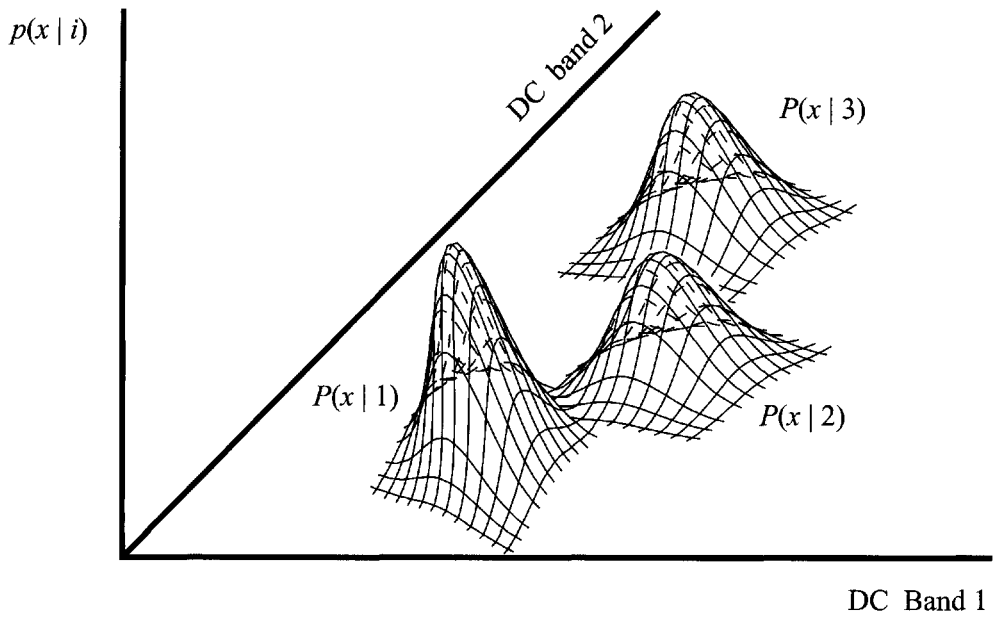


Figure 9.12 Conditional probability surface (multidimensional normalized histogram) for a two-band, three-class case.

where the definitions of the probability terms are the same as introduced previously, only now they are with respect to the spectral vector **x** instead of the scalar digital count value (DC). The conditional probability $p(\mathbf{x}|i)$ can be estimated from the normalized multidimensional histogram for each class as illustrated for a two-band case in Figure 9.12. The simple probability of **x** occurring in the image, $p(\mathbf{x})$ (while not needed in the classifier as we saw earlier), is the value of the normalized multidimensional histogram of the entire image. The pixel is assigned to whatever class has the highest a posteriori probability $p(i|\mathbf{x})$ for the spectral vector **x** associated with the pixel. While very attractive from a theoretical standpoint (i.e., this classifier will minimize classification errors by fully incorporating the spectral shape of the sampled pixels in each class), this approach is cumbersome to implement. It requires a very large set of training pixels to characterize adequately a multidimensional probability surface, and it is difficult to store a description of such a surface unless it is very well behaved.

In cases where the surface is well behaved, it is possible to approximate a solution to Eq. (9.68) by using a multivariate Gaussian maximum likelihood (GML) classifier. This classifier is based on the assumption that the conditional probability $p(\mathbf{x}|i)$ for each class can be approximated by a multivariate normal distribution. This is the multivariate equivalent of the assumption of Gaussian behavior for the conditional probability $[p(\text{DC}|i)]$ that was made in developing the univariate GML classifier. If the conditional probabilities are normal (i.e., if they can be represented by a multivariate Gaussian distribution), then they can be expressed as

$$p(\mathbf{x}|i) = \frac{1}{(2\pi)^{\frac{\ell}{2}} |\mathbf{S}_i|^{\frac{1}{2}}} e^{\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \mathbf{S}_i^{-1}(\mathbf{x}-\mathbf{m}_i)\right]} \tag{9.69}$$

where ℓ is the number of spectral bands, $\mathbf{S}_i$ is the covariance matrix for class i, $|\mathbf{S}_i|$ is the determinant of $\mathbf{S}_i$, $\mathbf{S}_i^{-1}$ is the inverse of $\mathbf{S}_i$, and $(\mathbf{x} - \mathbf{m}_i)^T$ is the transpose of $(\mathbf{x}-\mathbf{m}_i)$. The *covariance matrix* is a matrix of dimension $\ell \times \ell$ comprised of the band-to-band spectral covariance values for the class, i.e.,

$$\mathbf{S}_i = \begin{bmatrix} \sigma_{i11} & \sigma_{i12} & \cdots & \sigma_{i1\ell} \\ \sigma_{i21} & \ddots & & \\ \vdots & & & \\ \sigma_{i\ell 1} & & & \sigma_{i\ell\ell} \end{bmatrix} \tag{9.70}$$

where the *covariance* between spectral bands m and n for the ith class is defined as

$$\sigma_{imn} = \sum_{q=1}^{N} \frac{[DC_{im}(q) - \overline{DC}_{im}][DC_{in}(q) - \overline{DC}_{in}]}{N-1} \tag{9.71}$$

where N is the number of pixels in the sample set for class i and using this notation $\sqrt{\sigma_{mm}}$ is equivalent to σ_m, the standard deviation in band m.

The location of the multivariate normal distribution for a class is fully characterized by the mean vector ($\mathbf{m}_i$), and the shape of the distribution by the covariance matrix ($\mathbf{S}_i$). Duda and Hart (1973) point out that the isoprobability contours are hyperellipsoids centered on $\mathbf{m}_i$ and having a constant Mahalanobis distance from $\mathbf{m}_i$. The square of the *Mahalanobis distance* of $\mathbf{x}$ from $\mathbf{m}_i$ is defined as

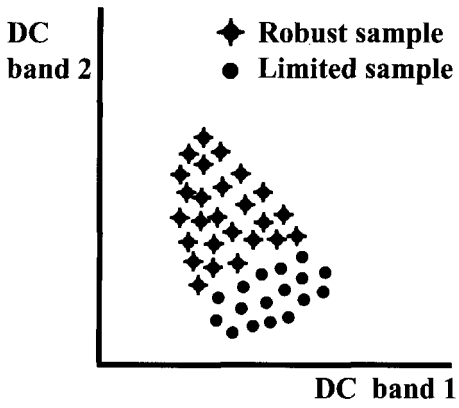
$$d_i = (\mathbf{x}-\mathbf{m}_i)^T \mathbf{S}_i^{-1} (\mathbf{x}-\mathbf{m}_i) \tag{9.72}$$

Combining Eqs. (9.68) and (9.69) yields the a posteriori probability for $\mathbf{x}$ belonging to class i if the data are normally distributed:

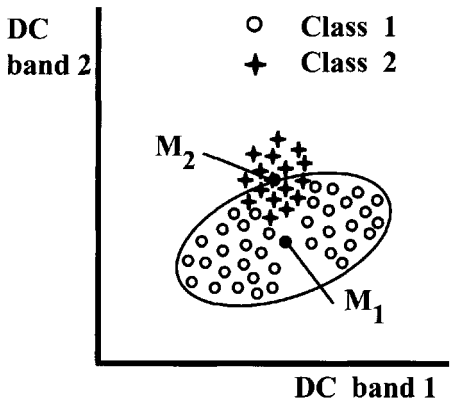
$$p(i \mid \mathbf{x}) = \frac{p(i)}{p(\mathbf{x})(2\pi)^{\frac{\ell}{2}} |\mathbf{S}_i|^{\frac{1}{2}}} e^{\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \mathbf{S}_i^{-1}(\mathbf{x}-\mathbf{m}_i)\right]} \tag{9.73}$$

We can define a discriminant function for the multivariate GML classifier using the same simplifying steps used in the univariate case. The decision regarding which class has the maximum value of $p(i|\mathbf{x})$ is independent of $p(\mathbf{x})$, so this term is dropped from Eq. (9.73). Taking the natural log will further simplify the expression and not change the rank order of the discriminant function, which is then defined as

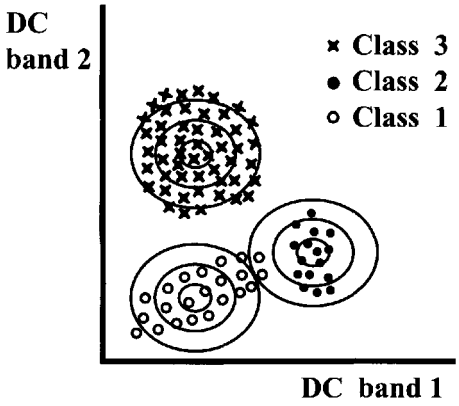
$$D_i = \ln[p(i)] - \frac{\ell}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{S}_i| - \frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \mathbf{S}_i^{-1}(\mathbf{x}-\mathbf{m}_i) \tag{9.74}$$



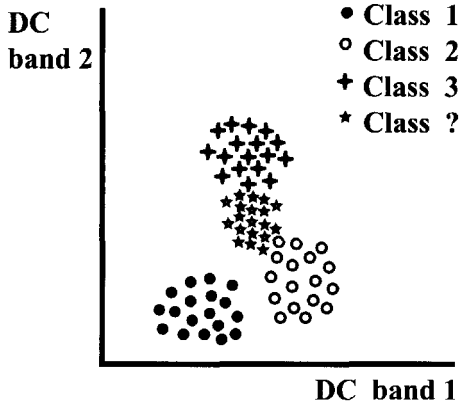
(a) Example of limited sample set in training process; note that both the mean and covariance can be seriously in error.



(b) Example of failing to separate a bimodal distribution in Class 1 into two classes.



(c) Example of using a pooled covariance matrix; note that the isoprobability curves are all the same except for location.



(d) Example of a class not represented in the training process.

Figure 9.13 Considerations in using a GML classifier.

Training data are used to generate estimates of the mean vector ($\mathbf{m}_i$) and covariance matrix ($\mathbf{S}_i$) for each class. The a priori probabilities are estimated from previous data or from a simple preliminary classification. From this information, the discriminant value for any unknown pixel can be computed from its spectral vector ($\mathbf{x}$) using Eq. (9.74). The pixel is then assigned to whichever class yields the maximum D_i value. The process is repeated for each pixel until the entire scene is classified. This will yield a classified scene that has the maximum number of correctly classified pixels based on the sample data available and the assumption of Gaussian distributions for the class populations. When something other than a minimum error solution is required, the risk functions discussed in Section 9.2.1.4 can be applied in exactly the same way as in the univariate case.

Before leaving this subject, we should note some caveats involved in using GML classifiers (many of these caveats apply to any classifier):

1. The user needs to make sure that the training data set is robust enough to characterize fully the class [cf. Fig. 9.13(a)]. For example, a forest class should include samples from different slopes, aspects, stand types, and densities. In addition, the user must ensure that there are sufficient data to estimate adequately the values of $\mathbf{m}$ and $\mathbf{S}$. Swain (1978) points out that $\ell + 1$ samples are a theoretical minimum (just enough to get you into trouble), but that $10\,\ell$ is closer to a practical minimum, and $100\,\ell$ is a desirable objective.
2. The user must ascertain that the data are approximately Gaussian. In particular, any multimodal classes should generally be split into separate classes for classification and merged into a common class afterward [cf. Fig. 9.13(b)]. For example, forest classes on east- and west-facing slopes may be significantly different, and forcing a unimodal Gaussian assumption can cause confusion with another class.
3. Many GML software packages allow the user to assume that the class covariance matrices are similar and to use a pooled covariance matrix for all classes. This assumes that the distributions are all the same shape and simply displaced by the mean vectors [cf. Fig. 9.13(c)]. This can lead to serious misclassification error if the assumption is false, which it generally is.
4. The user must ensure that all classes are included in the training process, or untrained classes will be grossly misclassified [cf. Fig. 9.13(d)]. To some extent this effect can be mitigated by using classification thresholds that will assign a pixel to an unknown class if the discriminant function fails to exceed a user-defined threshold. This is often useful in isolating small numbers of unusual materials and isolating pixels at boundaries. These "mixed pixels" have spectral vectors composed of mixtures of two material classes and can be seriously misclassified by a GML classifier.
5. The quality of the classifier should be evaluated prior to its use by testing on a known subset. If the performance on the subset of data used for evaluation is inadequate, more extensive/careful training should be conducted, alternate classification methods employed, or alternate data sets considered. There are several ways to evaluate the quality of a classifier as discussed in Section 9.2.5.

9.2.3 Unsupervised Multivariate Classifier

In some cases, it is useful to have the computer sort out which pixels have similar characteristics (e.g., spectra) rather than to try to force the pixels into a class based on our culturally driven sense of their similarities. This is done using an unsupervised classifier. Duda and Hart (1973) describe several approaches to unsupervised classification. We will examine only the simplest and most commonly used method here. It is often referred to as the *k*-means or iterative self-organizing data (*isodata*) analysis approach and in its simplest form requires only one piece of user-supplied data. The user must specify the number of classes (k) in the image. The *k*-means algorithm then attempts to locate the mean vector ($\mathbf{m}_i$) for each of the k classes. Normally, the image data are subsampled by selecting every qth pixel in the image to reduce the data volume. Next, k initial estimates of the location of the mean vectors in the ℓ-dimensional spectral space are obtained. These initial estimates of the mean can be made in a variety of ways, including randomly assigning locations in the ℓ-dimensional spectral space, randomly selecting pixels from the scene, or assigning equally spaced locations along the "gray" vector from darkest to brightest scene element. As long as the initial estimates are unique, the method should converge; however, the better the initial estimates, the quicker the convergence. If we designate these initial estimates as ${}^1\mathbf{m}_i$ (cf. Fig. 9.14), then we can tentatively assign each pixel to a class based on how close (minimum distance to the mean) it is to the mean vectors. The mean of all the pixels tentatively assigned to the ith class becomes our new estimate of the class mean (${}^2\mathbf{m}_i$). The sample pixels are tentatively reassigned using the new class means, and the procedure repeats in this fashion until the class means no longer change (i.e., the change is less than some threshold). At this point, the tentative means are assumed to be good estimates of the class mean vectors ($\mathbf{m}_i$). All pixels in the image can then be assigned to a class using the MDM classifier. Alternatively, the pixels assigned in the last iteration to the ith class can be used to generate estimates of the other statistics for the class (e.g., the covariance matrix), and a parallelepiped or GML classifier can be used.

The resultant classes will be indicative of the natural spectral clusters in the data. They may or may not correspond to land cover or material classes as we normally think of them. For example, in high-resolution airphoto data, a shadow class may develop, or a class that is a combination of shadow and water. In some cases, classes are formed with no obvious common characteristic. Because of this limitation, unsupervised classification is often used as a preprocessor for other algorithms. For example, when attempting to locate training regions for supervised classification, it is often useful to know where natural spectral groupings occur. If an unsupervised classifier clearly delineates forests on east- and west-facing slopes as two classes, then it is a good idea to make them two classes in the supervised classifier.

In practice, it is often useful to run an unsupervised *k*-means classifier with the initial estimate of k slightly greater than expected to "see" what types of spectral groupings naturally occur in the image. These "extra" classes often identify spectral features that would confuse a supervised classifier (e.g., clear and turbid

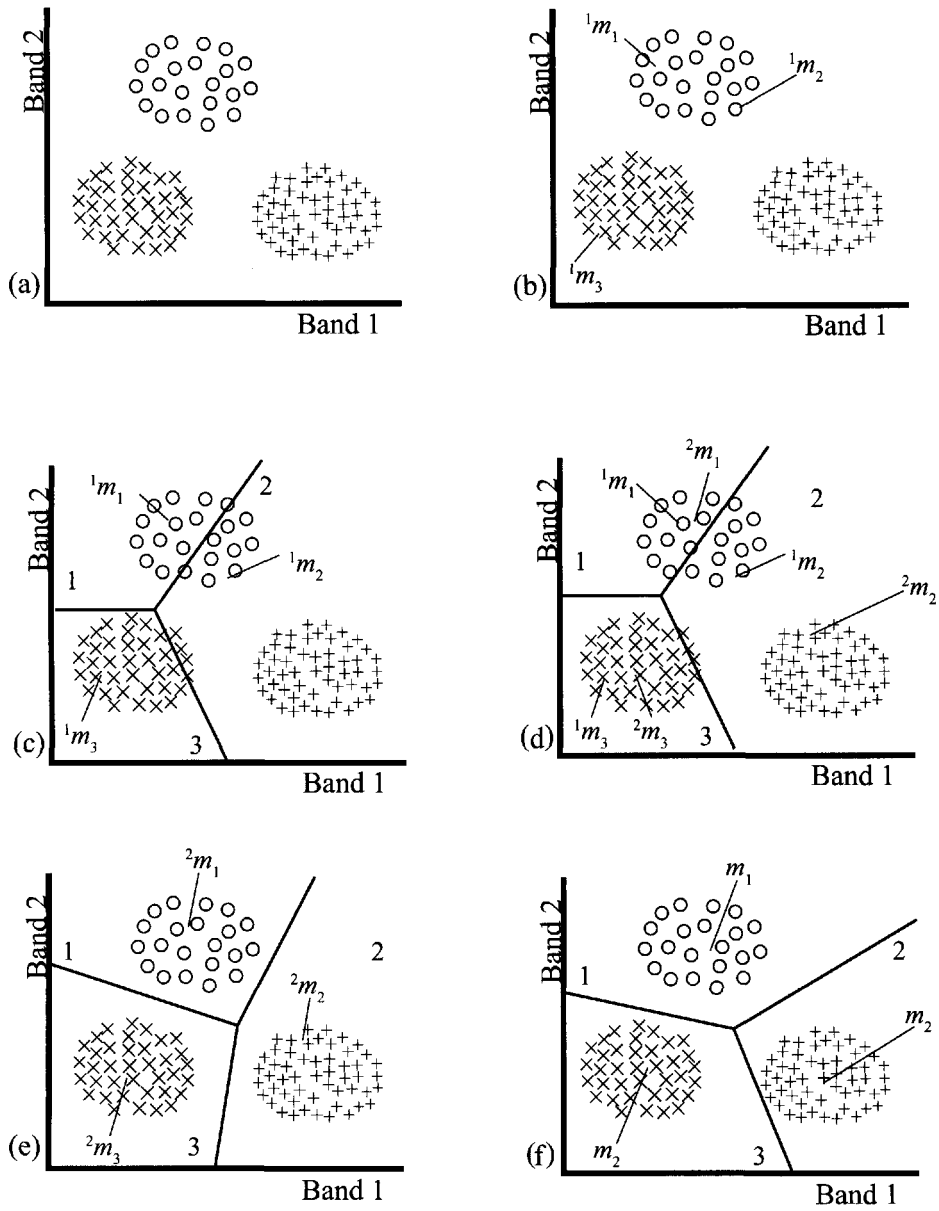


Figure 9.14 Unsupervised *k*-means classifier illustrating a case with two bands and three classes.

water as two separate classes) or point out classes that will be very difficult to separate spectrally.

The basic *k*-means or Isodata approach described here can be extended in a variety of ways to shrink or expand the number of clusters from an initial estimate. For example, once the initial clustering is completed, the number of clusters can be reduced by eliminating clusters with too few pixels (as defined by a user-supplied threshold). The pixels in the small cluster would then be assigned to the closest mean and the means recomputed. To avoid clusters too close together in spectral space, a merging option can be employed that will combine two clusters if their

final means are too close together (again, as defined by a user-supplied threshold). It is also possible to split clusters that have too much variance (as defined by a user-supplied threshold) by defining two means within the cluster and reclustering. While the threshold on all of the parameters in the isodata extensions to the basic k-means approach can be predetermined, they can often be improved by the interaction of a user experienced with the algorithm and the scene content. The final output of the isodata process, no matter how involved, is likely to suffer from the same limitations as the simple k-means approach. This is that the classes represent spectral clusters that may or may not coincide with our general perception of land cover classes. As a result, we may often need to exercise some control over an initial unsupervised classification by merging classes or using the unsupervised results to guide a supervised classification process.

9.2.4 Multivariate Classification Using Texture Metrics

Multivariate classifiers are generally used on multispectral data with from two to tens of spectral bands. The classifiers, however, are "blind" to the source of the data and will work on any data set that looks like spatially registered multiband data. Thus, one can use a monochrome image and several texture metric images (e.g., as described in Sec. 8.3) to form a multiband image that can then be classified (cf. Fig. 9.15). Alternatively, texture metric images can be combined with multispectral data to improve classification accuracy.

One way to do this is to simply compute the texture metrics described in Section 8.3 for one of the bands, a spectral average of the bands or for multiple poorly correlated bands (i.e., where the texture patterns may be different between two bands). An alternative method described by Scanlan et al. (2004) is to use a variation of the gray-level co-occurrence (GLC) matrix introduced in Section 8.3. This spectral co-occurrence matrix (SCM) is produced by comparing the common occurrence of a gray level between two spectral bands for pixels some distance apart in a specified direction (i.e., over two pixels and down two pixels). Thus, computationally it is identical to the GLC except that instead of comparing pixels from the same image, we compare them between spectral bands. The compositing metrics expressed in Eqs. (8.16)–(8.18) can then be used without change. The advantage

Figure 9.15 Example of the image shown in Figure 8.10 classified using texture metric images. See color plate 9.15.

of the SCM-based metrics is that they can simultaneously capture the spatial and spectral character of the local window. An overwhelmingly large number of band combinations, SCM matrices and texture metrics can be generated using this approach. Generally, only very small window sizes are used to reduce blurring of the texture clutter statistics; very local co-occurrence matrices are computed (e.g., one or two pixel shifts), and they are often averaged if orientation is arbitrary and only one or two SCM computed [e.g., one for two spectrally well-correlated bands and one for two spectrally ill-correlated bands (cf. Sec. 13.3.1 for a more general discussion of band selection)]. The resulting metrics [e.g., using Eqs. (8.16–8.18)] are produced on a per pixel basis and appended to the spectral vector for each pixel. The resulting image of augmented vectors can then be analyzed using the supervised and unsupervised tools described in the previous section. Care must be taken when using GML classifiers to ensure that the Gaussian approximations are valid. Furthermore, in some cases, it may be advisable to scale the texture metrics to have variances comparable to the spectral variances to avoid over- or underweighting of the texture metric features in the final classifier.

9.2.5 Evaluation of Class Maps

As indicated in Section 9.2.2, it is highly advisable to evaluate a classifier prior to applying it to a full data set and before accepting the product as final. This allows us to iteratively adjust the classifier if necessary (by retraining, changing control parameters, selecting an alternate image set, augmenting the data set, etc.) and provide an estimate of the expected accuracy of the final classifier to the user. Often this accuracy assessment is as important to the user as the actual values in the class map, and the accuracy can be easily misrepresented. Therefore, the remote sensing analyst must, whenever possible, both carefully assess the accuracy of the classifier and clearly represent to the user how the accuracy was assessed and what the assessment implies for the user. This section will consider several methods to assess classification accuracy.

The classification accuracy is usually assessed using a *contingency table* or more colloquially a *confusion matrix*. This is a $k \times k$ matrix for a k class problem where the rows represent the truth and the columns represent the results of the image-derived classification. Thus, entry (i,j) represents the number (or percentage) of pixels that are actually in class i that were assigned to be in class j. There are a number of ways to generate a confusion matrix, and this leads to one source of variability in how to interpret a statement about accuracy assessment. The first method commonly used in supervised classification is to simply run the training data through the classifier and produce a confusion matrix as shown in Figure 9.16(a). This shows how the training data would be classified by the classifier they trained. It is referred to as a *dependent data set* and can yield inflated performance estimates if the training data are not robust. It is often supplemented by also running the classifier on data from training sites that were not included in the data used to train the classifier. This *independent data* set is a good check of the robustness of the classifier and can also be used with unsupervised classifiers. If both data sets yield comparable results, this is a good estimate of how well the classifier will per-

Actual class	% classified as class 1	2	3	4	5
1	99.3	0.0	0.7	0.0	0.0
2	0.0	86.0	12.8	1.2	0.0
3	0.0	0.2	99.8	0.0	0.0
4	2.5	10.4	7.0	80.0	0.1
5	0.0	0.0	0.0	0.0	100.0

(a) Dependent data set

Actual class	% classified as class 1	2	3	4	5
1	97.3	1.5	1.1	0.0	0.0
2	0.2	74.0	12.2	1.5	12.1
3	0.0	4.1	95.9	0.0	0.0
4	2.4	1.6	6.4	89.6	0.0
5	0.0	0.0	0.0	0.0	100.0

(b) Independent data set

Actual class	% classified as class 1	2	3	4	5
1	93.5	0.0	2.2	0.0	4.3
2	0.0	29.7	65.7	0.8	3.8
3	12.3	0.0	87.4	0.3	0.0
4	0.0	0.0	1.4	98.6	0.0
5	0.0	0.0	0.0	0.0	100.0

(c) Randomly selected data set

Figure 9.16 Typical confusion matrices.

form on the uniform target areas that are typically used for training (cf. Fig. 9.17). In many cases a better estimate of performance is obtained by having the computer randomly select and highlight pixels on an interactive display. The user identifies the class to which the pixel belongs. The pixel is also classified by the classifier and the result entered into a confusion matrix. This process is continued until a good statistical estimate is obtained (typically 50 pixels per class yields a good estimate of performance). In many cases, this test is run after the initial classification so that equal numbers of pixels from each class can be randomly presented to the user. Figure 9.16 contains typical confusion matrices for dependent, independent, and randomly selected data sets. Note that the random classifier generally shows significantly poorer performance because it will include samples from transition regions and mixed pixels.

Thus, when reporting classification accuracy, the analyst needs to carefully specify which type of contingency table was used to generate the assessment. Because contingency tables can become large and cumbersome and because managers and decision makers can often only handle mononumeric solutions, a number of ways to analyze a contingency table have been developed. The simplest composite estimate is to compute the mean of the diagonal values in a contingency table (expressed as the percent classified in each entry) and report it as the accuracy of the classifier. This estimate does not well represent the actual expected accuracy over an image because it treats all classes the same. To overcome this problem, a simple weighted average can be used to estimate the expected classification accuracy. This can be expressed as

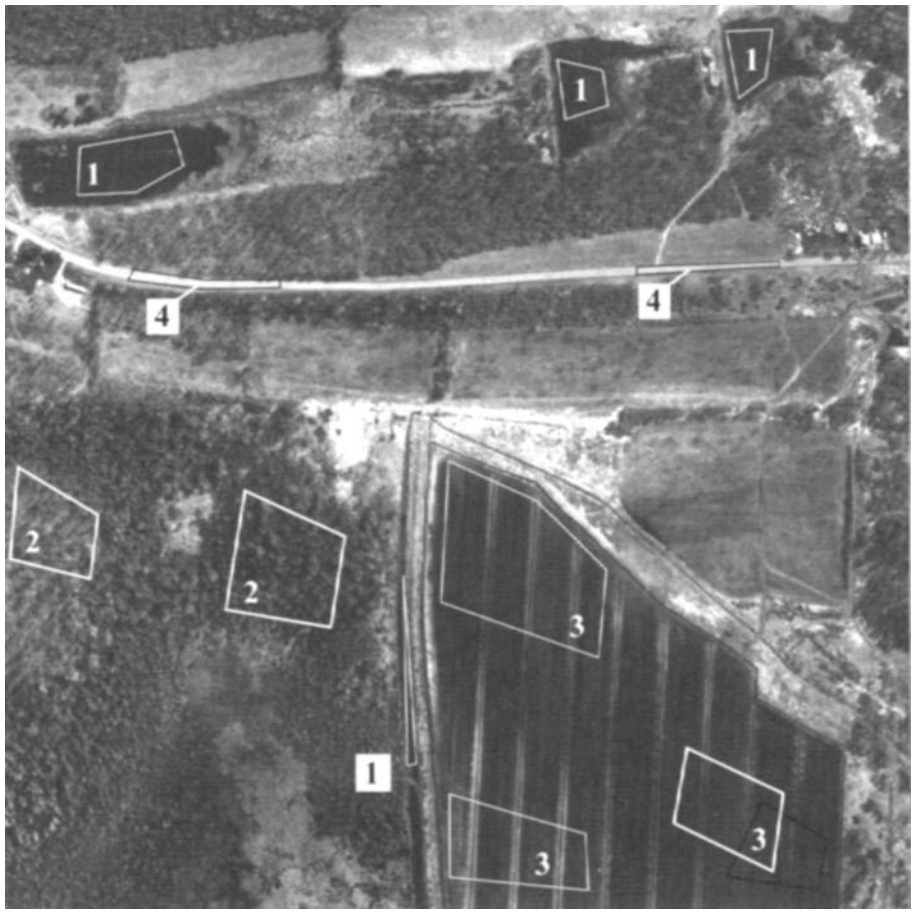
$$\hat{A} = \sum_{i=1}^{k} n_i c_{ii} / \sum_{i=1}^{k} n_i \tag{9.75}$$

where $\hat{A}$ is the expected accuracy, n_i is an estimate of the number of pixels in class i, c_{ii} is the percentage of class i pixels classified into the ith class (i.e., the entry in the percentage contingency table), and the sum is over all k classes. Clearly, one could report accuracy results based on contingency tables built using any of the methods described above. The contingency table based on the random selection approach is likely to generate the most realistic estimates. The accuracy computed using Eq. (9.75) requires an estimate of the number of pixels in each class. The values can come from the a priori estimates or can be computed based on the class maps generated with the classifiers.

Congalton et al. (1983) point out that the simple accuracy estimate based on the diagonal elements ignores the random probability of correct classification and overestimates the accuracy due to the classification process. They suggest using the kappa coefficient to overcome this limitation. The *kappa coefficient* ($\hat{k}$) is based on the simple accuracy and can be expressed as

$$\hat{k} = \frac{p_o - p_c}{1 - p_c} \tag{9.76}$$

where p_o is the simple accuracy expressed as



Class 1 : Open water
Class 2 : Forest
Class 3 : Agricultural muckland
Class 4 : Roads

•
•
•

Figure 9.17 Training data sets for input to a classifier.

$$p_o = \left(\sum_{i=1}^{k} c_{ii} \right) / N \tag{9.77}$$

N is the total number of samples used to construct the contingency table, c_{ii} is the ith diagonal element in a numerical contingency table and p_c is the proportion of the accuracy due to random chance, which can be expressed as

$$p_c = \frac{1}{N^2} \sum_{i=1}^{k} c_{it} c_{ti} \tag{9.78}$$

where

$$c_{it} = \sum_{j=1}^{k} c_{ij} \tag{9.79}$$

and

$$c_{ti} = \sum_{j=1}^{k} c_{ji} \tag{9.80}$$

The kappa coefficient is an estimate of overall accuracy and typically ranges from 0 to 1 just like the simple and weighted accuracy estimates. For large samples, the variance in the kappa coefficient can be expressed as [cf. Congalton and Green (1999)]

$$\sigma_k^2 = \frac{1}{N}\left[\frac{p_o(1-p_o)}{(1-p_c)^2} + \frac{2(1-p_o)(2p_o p_c - a_1)}{(1-p_c)^3} + \frac{(1-p_o)^2(a_2 - 4p_c^2)}{(1-p_c)^4}\right] \tag{9.81}$$

where

$$a_1 = \frac{1}{N^2}\sum_{i=1}^{k} c_{ii}(c_{it} + c_{ti}) \tag{9.82}$$

and

$$a_2 = \frac{1}{N^3}\sum_{i=1}^{k}\sum_{j=1}^{k} c_{ji}(c_{it} + c_{tj})^2 \tag{9.83}$$

Having a measure of the variance in kappa is important as it allows us to compute confidence limits about kappa using standard Gaussian statistics [cf. Congalton et al. (1983)]; e.g., at the 95% confidence interval we expect kappa to be

$$\hat{k} \pm 1.96\sigma_k \tag{9.84}$$

More importantly, if we wish to compare the performance of two classifiers, we can use contingency tables to estimate kappa values for methods A and B and statistically test the hypothesis that they are the same, i.e.,

$$H_0 : \hat{k}_A = \hat{k}_B \text{ or } \hat{k}_A - \hat{k}_B = 0 \tag{9.85}$$

This uses the standard z score given by

$$z = \frac{\left|\hat{k}_A - \hat{k}_B\right|}{\sqrt{\sigma_{kA}^2 + \sigma_{kB}^2}} \tag{9.86}$$

If z exceeds the threshold for the required confidence level, we can reject the null hypothesis and conclude that one of the classifiers is outperforming the other. Figure 9.18 shows an example set of calculations used for comparing two types of classifiers run on a common data set. In this case, we could conclude that classifier B is significantly better than A for this data set at the 95% confidence level.

This section has described a number of metrics for evaluating the performance of a classifier, as well as how the kappa coefficient can be used to compare the relative performance of classifiers. It is important to recall that these analysis

Classifier A

Pixels actually in class

Pixels assigned to class	A	B	C	D	E	F	C_{it}
A	1721			6	15	8	1750
B		448					448
C		2	358			1	361
D				55	2	22	79
E			21		45	1	67
F	133		4	221	240	225	823
c_{ti}	1854	450	383	282	302	257	

Total = N = 3528
Trace = 2852

$$\text{Simple Accuracy} = P_o = \frac{\text{Trace}}{\text{N}} = 0.808$$

$$P_c = 0.31$$

$$\hat{k} = 0.723$$

$$\sigma_k^2 = 7.53 \cdot 10^{-5}$$

Classifier B

Pixels actually in class

Pixels assigned to class	A	B	C	D	E	F	C_{it}
A	1854	103	23	3		3	1986
B		383					383
C			362		5	8	375
D				231	17	2	250
E		1			180	1	182
F				48	105	248	401
c_{ti}	1854	487	385	282	307	262	

Total = N = 3577
Trace = 3258

$$P_o = 0.911$$

$$P_c = 0.33$$

$$\hat{k} = 0.867$$

$$\sigma_k^2 = 4.81 \cdot 10^{-5}$$

95% confidence interval: $\hat{k} \pm 1.96\sigma_k$:

$$H_0 : \hat{k}_A - \hat{k}_B = 0$$

$$z = \frac{|\hat{k}_A - \hat{k}_B|}{\sqrt{\sigma_{k_A}^2 + \sigma_{k_B}^2}} = 12.9$$

If $z \leq 1.96$, accept null hypothesis at 95% confidence level.

In the case we reject the null hypothesis.

Figure 9.18 Sample confusion matrix/contingency table calculations.

tools can be applied to any classification process whose results can be cast in the form of a confusion matrix. Thus, these evaluation methods will be equally applicable to texture based classifiers, multispectral classifiers, and material maps based on imaging spectroscopy as discussed in Chapter 10.

9.2.6 Limitations of Conventional Multispectral Classification

A major limitation of multispectral classifiers, as they are normally used, is that the classifier is scene specific and cannot be used on any other scene. This is because the classifier is based on raw digital count values and has the sensor and atmospheric calibration effects incorporated into the classifier. In some sense, this is a beneficial attribute since we do not need calibrated data to build a classifier, but it limits the applicability of the classifiers. Thus, it is often not cost-effective to build extremely elaborate classifiers only to discard them after a single use. Two alternatives exist to this disposable classifier approach. The first is to calibrate scenes into reflectance units and to build reflectance-based classifiers. These classifiers could be reused and made more and more sophisticated because they are based on stable physical units. The limitation of this approach is that most of the operational atmospheric calibration techniques have residual errors large enough to introduce serious classification errors. The second approach is to use multidate radiometric normalization techniques. In this approach, a standard supervised classifier is constructed using imagery acquired on one day. Images of the same scene on other days are radiometrically transformed using one of the methods discussed in Section 7.5.2 such that the day-2 DC values vary with reflectance in the same way as the day-1 values. The day-1 classifier can then be directly applied to the transformed day-2 image. Using the method of Schott et al. (1988), it was demonstrated that under certain conditions update scenes normalized with the pseudo-invariant feature technique could be classified with a day-1 classifier nearly as well as with a classifier built specifically for the day-2 images. However, extension of classifiers to multiple days using normalization is usually limited to short time spans between acquisitions (e.g., several weeks) or to images within a few weeks of anniversary dates (e.g., one year ± several weeks) to avoid significant seasonal changes in the reflectance spectra.

9.3 IMAGE TRANSFORMS

In most cases, we tend to assume that more spectral bands will yield a better classifier when processing multispectral data. In cases where we have only a few bands, this is often a good assumption. However, more are not always better for two reasons: processing time and interband correlation. In the first case, if we add too many bands, the processing time to run multispectral image analysis algorithms can become staggering. The second reason is that both spatial (texture) and spectral data tend to be highly correlated. Adding another band of data that is highly correlated with previous bands may add little new data and only increase

the noise and the processing time, possibly even reducing the effectiveness of the algorithm.

Image transforms can be used to overcome some of these limitations. These transforms are designed to redefine the multivariate feature vectors into a feature space where the features are more decorrelated. Therefore, fewer features can carry the information needed for processing algorithms such as multispectral classifiers. These transforms are also often used to attempt to reduce some of the variations that result from processing uncalibrated data. In many cases, the transforms tend to normalize out some of the atmospheric and sensor calibration effects.

The most common transform is the *principal component* (PC) transform that is specifically designed to decorrelate the data and maximize the variability in a reduced number of features. The PC transform utilizes the *eigenvectors* and eigenvalues of the image covariance matrix introduced in Section 9.1. Each feature value in the transformed data set is a linear combination of the features in the input data set. If $\mathbf{x}$ is the vector comprised of ℓ digital count values corresponding to the ℓ features (most often spectral bands) in the input image, then an output feature (PC) can be defined as

$$PC_1 = \mathbf{e}_1^T \mathbf{x} = e_{11}DC_1 + e_{12}DC_2 \cdots e_{1\ell}DC_\ell \tag{9.87}$$

where PC_1 is the "brightness" value of the first principal component feature, and e_1 is the first principal component vector (eigenvector) composed of ℓ weights (e_{11}, $e_{12}e_{1\ell}$), i.e.,

$$\mathbf{x} = \begin{bmatrix} DC_1 \\ DC_2 \\ \vdots \\ DC_\ell \end{bmatrix} \text{ and } \mathbf{e}_1 = \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1\ell} \end{bmatrix} \tag{9.88}$$

Recall from vector algebra that Eq. (9.87) can be thought of as projecting the vector $\mathbf{x}$ onto the vector $\mathbf{e}_1$. The vector $\mathbf{e}_1$ can be thought of as a new axis, and PC_1 is the magnitude of the vector $\mathbf{x}$ along the axis defined by the vector $\mathbf{e}_1$ (cf. Fig. 9.19).

There are ℓ principal component features, so we can define a principal component feature vector **PC** for each pixel by transforming the input feature vector through the transform matrix **E** according to

$$\mathbf{PC} = \mathbf{E}^T \mathbf{x} \tag{9.89}$$

where

$$\mathbf{PC} = \begin{bmatrix} PC_1 \\ PC_2 \\ \vdots \\ PC_\ell \end{bmatrix} \tag{9.90}$$

and **E** is composed of ℓ principal component column vectors (**e**), i.e.,

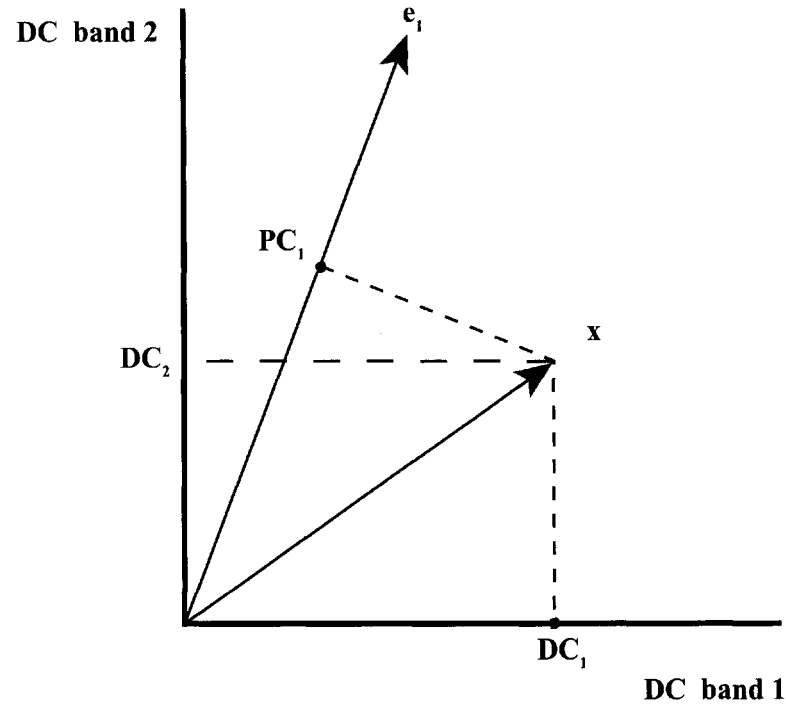


Figure 9.19 Projection of the vector **x** onto the $\mathbf{e}_1$ axis to yield the magnitude of the projected value (PC_1).

$$\mathbf{E} = \underbrace{\begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1\ell} \end{bmatrix}}_{\mathbf{e}_1} \underbrace{\begin{bmatrix} e_{21} \\ e_{22} \\ \vdots \\ e_{2\ell} \end{bmatrix}}_{\mathbf{e}_2} \cdots \underbrace{\begin{bmatrix} e_{\ell 1} \\ e_{\ell 2} \\ \vdots \\ e_{\ell\ell} \end{bmatrix}}_{\mathbf{e}_\ell} \tag{9.91}$$

The discussion to this point is valid for any transformation matrix **P**. What makes the principal component transform unique is that the new axes ($\mathbf{e}_i$) are orthogonal and are selected such that the first principal component axis is the axis that will exhibit the most variance when the input data vectors are projected onto it. The second principal component is the axis orthogonal to $\mathbf{e}_1$ that exhibits the next largest variance, etc. Formally, **E** is defined such that it diagonalizes the covariance matrix, i.e.,

$$\mathbf{E}^T\mathbf{S}\mathbf{E} = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & & \\ 0 & 0 & & & \\ 0 & & & & \lambda_\ell \end{bmatrix} = \mathbf{\Lambda} \tag{9.92}$$

where λ_i is the variance of the ith principal components feature values (PC_i), and the λ values are arranged such that $\lambda_1 > \lambda_2 > \cdots \lambda_{\ell-1} > \lambda_\ell$. The columns of **E** are the ordered eigenvectors ($\mathbf{e}_i$) corresponding to the ordered eigenvalues (λ_i) of the image

covariance matrix. An approach for computing the eigenvectors of a matrix was introduced in Section 9.1. Recall from that discussion that the eigenvectors of a real symmetric matrix will form an orthonormal set. Using the eigenvectors, we can generate a transformed image with ℓ bands where each band is composed of PC_i values. The first band, or feature, in the multiband principal component image will account for most of the variance in the data, with decreasing amounts in the remaining bands. The latter bands in a many-band PC-transformed image tend to contain mostly random noise. Figure 9.20 shows a PC-transformed six-band Landsat TM image. All the data have been contrast stretched to enhance the visual representation of the data. The actual percentage variance in each principal component band is shown in the figure. This is typical of Landsat TM scenes with most of the image data contained in the first three or four principal components.

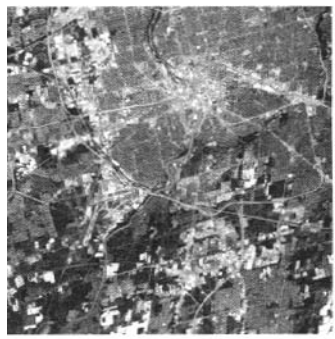
The percentage of the scene variance described by each PC band can be computed from the diagonalized matrix of eigenvalues ($\mathbf{\Lambda}$) according to

$$\text{percent variance in band i} = \frac{\lambda_i}{\text{trace } \mathbf{\Lambda}} \cdot 100 \tag{9.93}$$

Multispectral classification algorithms can be run using only the first few principal component images, greatly reducing run times. In some cases, simpler classifiers can be used because the transform process, by selecting axes with high variance, will tend to increase the separability of classes. This can be seen for a simple case in Figure 9.21. PC-transformed images can also be used for data compression in storing and transmitting multispectral data, since the latter PC bands can often be completely neglected and intermediate bands require fewer bits per pixel to characterize because of their low variance.

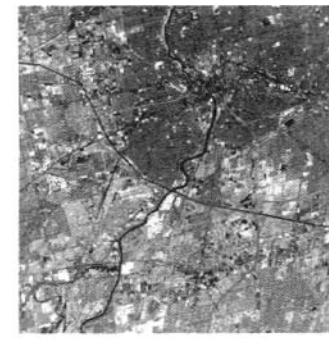
Care must be taken when using the principal component transform. Principal component analysis essentially equates variance with information. Thus, high-variance bands will tend to dominate the first few principal components, with low-variance bands appearing in the later PC bands regardless of information content. For example, in Landsat TM imagery, the thermal band contains unique information not depicted in the other bands. However, because the thermal band usually has low variance, the data from the thermal channel tends to be mixed with noise from the other channels and partially lost in the later principal components. This problem can be avoided by normalizing the data by the variance in each band before performing the principal component analysis (cf. the noise-adjusted PC transform in Chapter 10 for a more rigorous approach to normalization).

A disadvantage of the principal component data is that the **PC** vectors are scene dependent and often have no obvious correlation with a physically interpretable concept. Thus, image interpretation of principal component images can be difficult. Furthermore, a PC_2 image of one scene may represent different information than a PC_2 image from another scene. On the other hand, principal component images of multidate scenes of the same area tend to be correlated with each other and to some extent can be processed in the same way (i.e., some degree of normalization is included in the transform process). As with most multidate approaches, the interpretation of the PC transforms will change if there is a significant change in the information content in the scene (e.g., crop development).



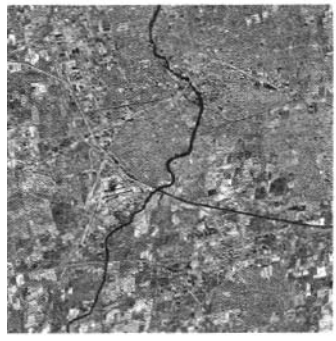
(a)

(58.55)



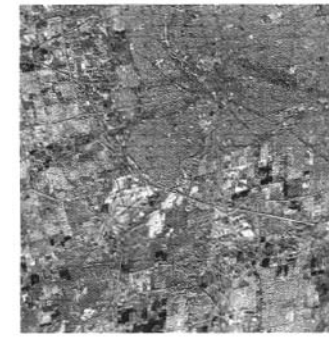
(b)

(32.54)



(c)

(7.28)

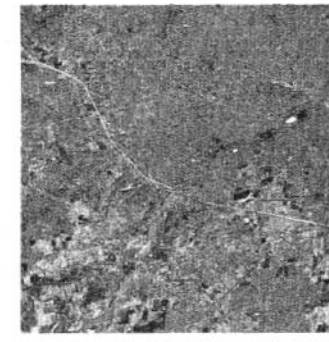


(d)

(0.96)

(e)

(0.59)



(f)

(0.08)

Figure 9.20 Principal component images from a six-band Landsat TM scene of Rochester, NY. Percentage of variance contained in each PC band is shown in parentheses.

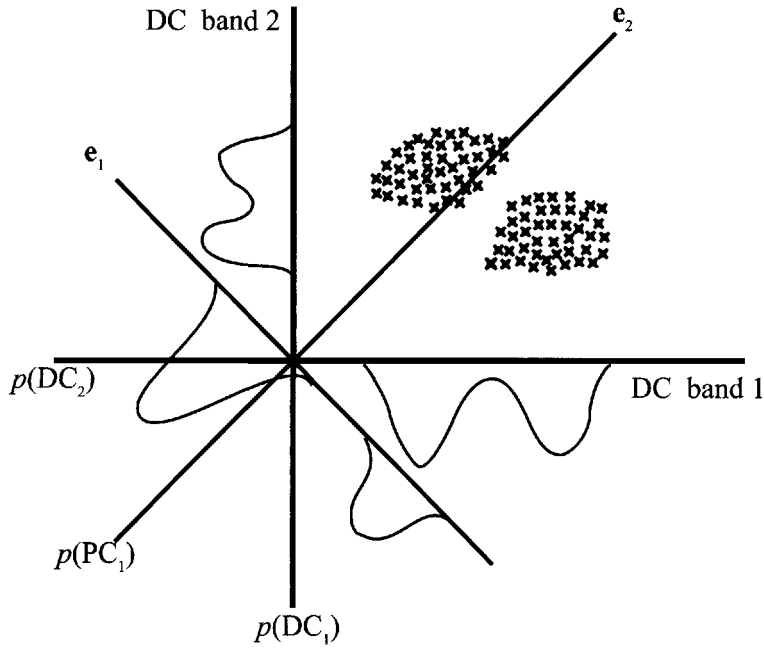


Figure 9.21 Principle component axis showing how increased separability can be realized by maximizing the variance on the projected axis.

A limitation of the principal component transform can arise if there are a small number of pixels associated with a critical class. Since the transform is based on whole-image statistics, such a class may not be well separated in the transform space. This problem can be addressed using the *canonical transform*, which is designed to identify a new coordinate space that maximizes the separation between class means and minimizes the dispersion within the classes [cf. Schowengerdt, (1983)]. Clearly, this transform requires information about the mean and covariance matrix for each class and is therefore only applicable when training data are already available.

An alternative transform approach was developed by Kauth and Thomas (1976). They defined a transform space that was designed to improve the analysis of agricultural scenes using Landsat MSS data. The *tasseled-cap* transform (named after the shape of the population distribution) is designed to project the data along a set of axes where the first three axes correspond roughly with the brightness of soils (brightness axis), the vegetation biomass (greenness axis), and the senescence of vegetation (yellowness axis). The tasseled-cap transform is sensor specific (assuming fixed sensor gain) and is designed to transform the input data to a feature space where the features are more directly correlated with an application parameter and more intercomparable over time. While not an optimized transform for any given image, it has the advantage from the user's standpoint of using a constant precomputed transform matrix. It must, however, be recognized that the transform was designed for particular types of scenes and is sensor specific. Note that the Gram-Schmidt orthogonalization process discussed in Section 9.1 can be used to

help guide the generation of an orthogonal set of transform vectors where the user can explicitly set the first axis by selecting the first Gram-Schmidt vector.

9.4 HIERARCHICAL IMAGE PROCESSING

Many image processing algorithms cannot be run effectively in a single direct application to an image. In most cases, the more sophisticated algorithms are designed to be applicable only to a subset of the image data. For example, the NDVI algorithm is only meaningful in vegetated regions. An algorithm designed to measure the amount of damage to a particular crop based on correlation with field reflectance data will yield false signatures on other crops, as well as on any other targets in the scene. To take advantage of these target-specific algorithms, a hierarchical approach to image processing is often required. For example, in the crop damage case, we might run the following sequence of algorithms. First, the scene would be calibrated to reflectance using ground truth panels. Second, a GML classifier could be trained and run to isolate agricultural fields from other land cover types. Third, a land cover classifier, including texture metrics, could be trained and run on just the agricultural fields to isolate the particular crop of interest. Next, a postclassification algorithm (mode filter) could be used to clean up the classifier. Finally, the crop damage algorithm can be run on the reflectance values for just those pixels that correspond to the crop for which the algorithm was designed. This hierarchical approach allows us to build more and more specialized algorithms based on a priori knowledge about specific pixels, (i.e., only operate on agricultural fields or only a particular crop). In this way the algorithms can become very specialized without becoming extremely complex.

9.5 REFERENCES

Congalton, R.C., & Green, K. (1999). *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. Lewis Publishers, NY.

Congalton, R.S., Oderwald, R.G., & Mead, R.A. (1983). Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. *Photogrammetric Engineering and Remote Sensing*, Vol. 49, No. 12, pp. 1671-1678.

Duda, R.O., & Hart, P.E. (1973). *Pattern Classification and Scene Analysis*. Wiley, NY.

Johnson, R.A., & Wichern, D.W. (1998). *Applied Multivariate Statistics*. 4th ed., Prentice Hall.

Kauth, R.J., & Thomas, G.S. (1976). "The tasseled cap — a graphic description of the spectral-temporal development of agricultural crops as seen by Landsat." Proceedings of the Symposium on Machine Processing of Remotely Sensed Data, Purdue University, W. Lafayette, IN.

Morrison, D.F. (1967). *Multivariate Statistical Methods*. McGraw-Hill, NY.

Press, W., et al. (1992). *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambridge.

Richards, J.A. (1999). *Remote Sensing Digital Image Analysis, An Introduction*. 3rd ed., Springer, NY.

Scanlan, N.W., Schott, J.R., & Brown, S.D. (2004). Performance analysis of improved methodology for incorporation of spatial/spectral variability in synthetic hyperspectral imagery. *Proceedings SPIE*. Vol. 5159, pp. 319-330.

Schott, J.R., Salvaggio, C., & Volchok, W.J. (1988). Radiometric scene normalization using pseudoinvariant features. *Remote Sensing of Environment*, Vol. 26, No. 1, pp. 1-16.

Schowengerdt, R.A. (1983). *Techniques for Image Processing and Classification in Remote Sensing*. Academic Press, Orlando, FL.

Schowengerdt, R.A. (1997). *Remote Sensing Models and Methods for Image Processing*. Academic Press, San Diego, CA.

Swain, P.H. (1978). Fundamentals of pattern recognition in remote sensing. In P.H. Swain & S.M. Davis, eds. *Remote Sensing: The Quantitative Approach*, McGraw-Hill, New York, NY.

Wylie, C.R., & Barret, L.C. (1995). *Applied Enginerring Mathmatics*. 6th ed., McGraw Hill, NY.

CHAPTER 10
SPECTROSCOPIC IMAGE ANALYSIS

In this chapter we will expand the discussion of image processing algorithms to include those algorithms that have been developed or adapted for use with image spectrometer data. We will see that, in some cases, this is a very arbitrary distinction, with the imaging spectroscopy data simply representing more bands to be processed in the same fashion as multispectral data. From this perspective, the multispectral algorithms of Chapter 9 can all be applied to spectroscopic data (albeit with significant run time penalties in many cases). Conversely, several of the algorithms we will introduce (e.g., spectral angle mapper in Sec. 10.3 and RX in Sec. 10.4) are directly applicable to multispectral data. In fact, in the case of RX, the algorithm was first introduced for use with multispectral data. However, it has been much more widely applied to hyperspectral data.

There are also cases where the increased dimensionality of hyperspectral data allows us to solve problems that could not be solved at all with multispectral data or could only be solved in trivial cases. This is because many of the problems we will address have several unknowns. If, for example, we attempt to solve for those unknowns using a system of simple linear equations, then we need as many linearly independent equations as we have unknowns. The number of truly independent equations we can write is a function of the inherent dimensionality of the data. For multispectral data, this is often only two or three bands, meaning we can only solve uniquely for two or three variables. More complex problems are ill-posed with the often unsatisfactory mathematical solution that there are multiple valid answers. The higher dimensionality of hyperspectral data often allows us to solve uniquely for significantly more variables (mathematically we have well-posed or

over determined problems). This allows us to introduce algorithms such as spectral unmixing and orthogonal subspace projection (cf. Sec. 10.3), which for many interesting cases will not yield closed solutions for the classic four- to six- band multispectral data sets.

In addition to the types of algorithms discussed above, which are only different by degree (i.e., numbers of independent variables or bands), we will also examine algorithms that take direct advantage of the spectral structure of the data (cf. Sec. 10.5). These algorithms concern themselves with the spectral location, shape, and depth of one or more absorption features in a manner that could not even be considered with multispectral data.

10.1 PERSPECTIVES ON SPECTRAL DATA

The algorithms that have evolved to analyze spectral image cubes tend to take one of three different, though not necessarily mutually exclusive, perspectives on the nature of the data set. This section introduces these three perspectives in an attempt to provide at least one taxonomy for the multitude of algorithms in the literature. This approach draws on similar efforts by Manolakis [cf. Manolakis et al. (2000) and Keshava et al. (2000)] and Schaum and Stocker [cf. Schaum (2001) and Schaum and Stocker (2002)]. The three perspectives on data representations are the geometric or deterministic approach, the statistical or stochastic approach, and the spectral feature approach.

10.1.1 The Geometric or Deterministic Perspective

The *geometric* or *deterministic approach*, which is also referred to as the *structural model*, is often associated with the concept of end members. An *end member* can be thought of as representing a spectral extremum that is characteristic, for example, of a material type. We can then think of any pixel in the image cube as being made up of an area-weighted mixture of these end members. From this perspective, a pixel brightness vector is thought of as a mixture of end-member brightness vectors. A pixel containing only the spectrum of an end member is thought of as a pure pixel. The entire spectral data set can then be thought of as being made up of a deterministic set of end-member vectors or extrema, with all other vectors being mixtures of these end members. The mixtures, being made up as linear combinations of the extrema, must lie between end members (if a mixture of two) or within the simplex formed by any mixture of three or more. This gives rise to the geometric concept of a spectral data set forming a *convex hull* with the end-member spectra defining the vertices of the hull and mixtures of spectra lying on the surface or internal to the hull (cf. Fig. 10.1). Note that the units in which the data are represented (e.g., reflectance, radiance, or digital count) will change the numerical location of end members but will not change the general concept of the convex hull (i.e., extrema in one representation will still be extreme in another). This geometric perspective on spectral data has given rise to a wide range of image

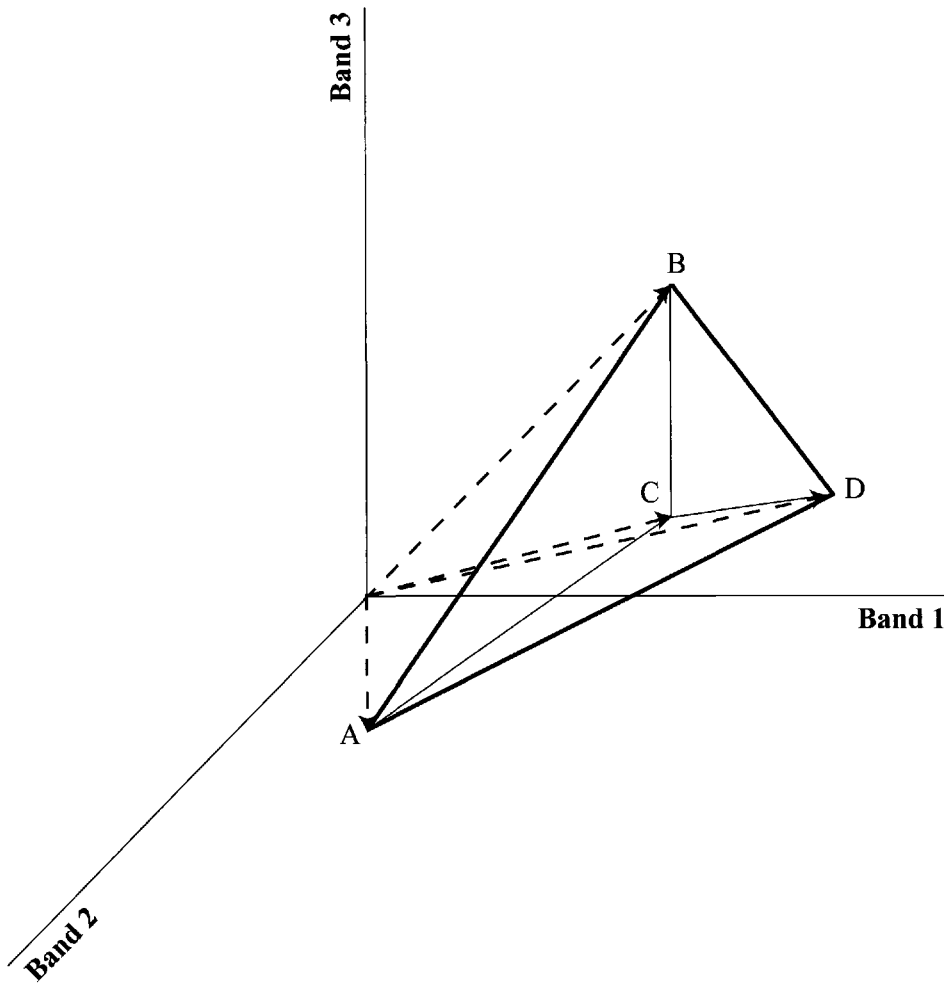


Figure 10.1 Geometric or structural perspective on spectral data. For simplicity, we show a three-dimensional spectral space characterized by four end members (A, B, C, and D) and linear mixtures thereof.

analysis tools, including the spectral angle mapper, spectral mixture models, and orthogonal sub-space projection, which are discussed in Section 10.3.

10.1.2 The Statistical Perspective

The statistical perspective is the logical extension of the *statistical approach* to land cover classification, introduced in Chapter 9, to the spectral analysis problem. From this perspective, the image data can be described collectively as a cluster of vectors in spectral space, which can be characterized by statistical parameters (e.g., the mean vector and the covariance matrix). Refinements to this approach include use of spatially localized mean vectors and covariance matrices to describe a localized background and class- or cluster-specific mean vectors and covariance

matrices (cf. Fig. 10.2). This perspective leads to a number of algorithms based on local or scene-wide probability estimates concerning targets, backgrounds, and/or spectral clusters as discussed in Section 10.4.

10.1.3 Spectral Feature Representation

The *spectral feature* representation draws most directly on the nature of each pixel as a spectrum resulting from spectroscopic dispersion in an electro-optical instrument. As a result, this approach looks at the individual spectrum for specific features. This takes the form of searching for absorption features at a specific wavelength, the shape of an absorption feature, or the presence of combinations of absorption features (cf. Fig. 10.3). The resulting algorithms try to compare the characteristics of a given spectrum to a library representation or to a sample of interest (e.g., target) as discussed in Section 10.5. This approach attempts to look for the electro-optical manifestations of the chemical structure of the materials under study and, as a result, most closely resembles classical spectroscopy.

10.2 ISSUES OF DIMENSIONALITY AND NOISE

Before proceeding to look at the algorithms that result from these various perspectives on spectroscopic imagery, we need to recognize that many of the approaches need to account for two important characteristics of hyperspectral data sets. These are the large dimensionality that can make algorithmic analysis cumbersome and the noise levels that, if significant, can mask features of interest (i.e., cause confusion or reduce performance in an algorithm).

10.2.1 Dimensionality Reduction

Hyperspectral data sets often consist of hundreds to thousands of spectral bands, making them computationally intractable and impossible to fully visualize. Thus, it is often desirable to reduce the dimension of the data to a more manageable number of bands. This section presents some of the most straightforward methods of band reduction. More sophisticated methods incorporating noise effects are discussed in Section 10.2.3.

One of the first steps in reducing the dimensionality of the data is to remove bands that don't carry any information about the scene or target of interest. For example, many sensors acquire data on constant spectral centers right through very strong atmospheric absorption features. It is common practice to drop these bands, containing no signals from the ground, from further analysis. In addition, after atmospheric compensation (if employed), bands heavily influenced by atmospheric absorption may also be removed from further analysis. Finally, for specific targets, bands known to carry little or no information may be removed from subsequent analysis. For example, after atmospheric compensation (cf. Sec. 7.6), the spectral variations of interest for water-quality assessment are typically in the VNIR region of a VNIR-SWIR spectrometer, and the SWIR bands can be removed. In many

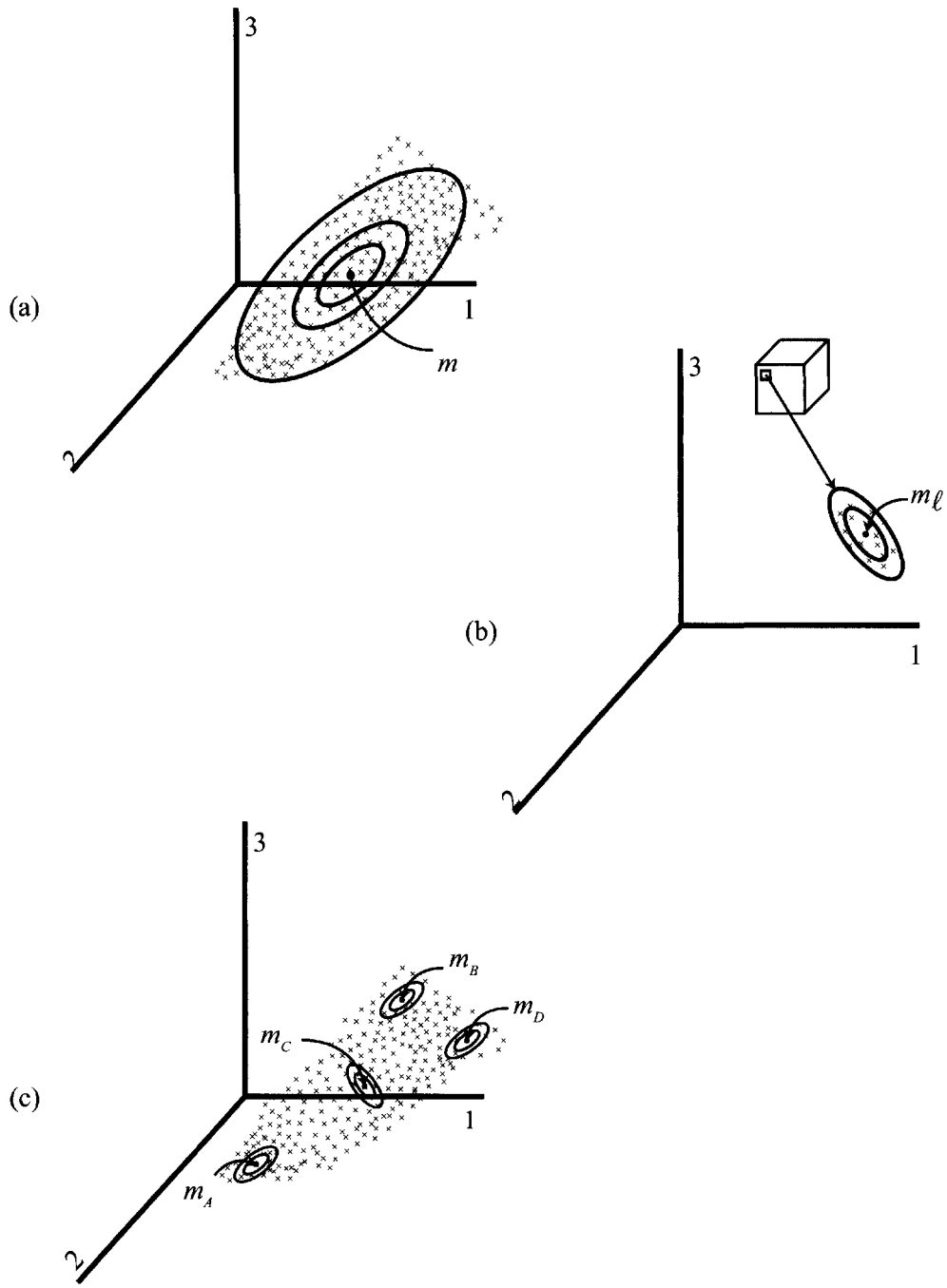


Figure 10.2 Illustration using three bands of a spectral data set to show the stochastic perspective: (a) image wide mean (m) and isoprobability contours based on the covariance matrix and Gaussian assumptions, (b) a spatially local mean and isoprobability surface, and (c) Shows cluster means and isoprobability surfaces.

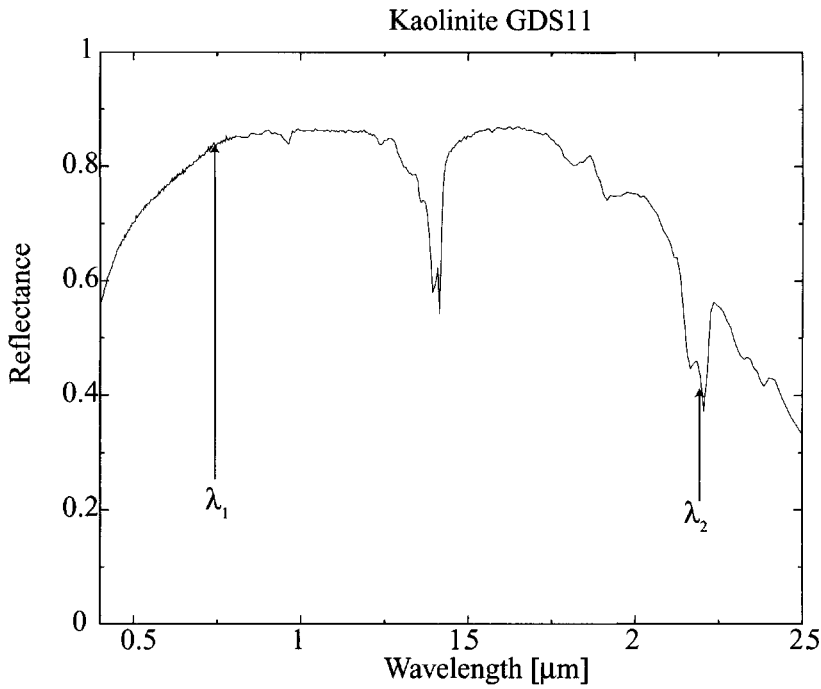


Figure 10.3 Illustration showing the spectral feature perspective. In this case, we might be looking for a spectrum that has no feature at λ_1 and has an absorption feature at λ_2, and the absorption feature at λ_2 must be a doublet. (Spectrum courtesy of U.S. Geological Survey.)

cases, these processes only have a limited impact on reducing the total number of bands still needing to be analyzed. When computationally intensive algorithms are to be employed, additional dimensionality reduction may be necessary.

At this point, one option is the rather primitive step of averaging two or more adjacent bands and forming a smaller number of lower spectral resolution bands. This will reduce the dimensionality and increase the signal to noise in the reduced resolution bands. However, it has the potential downside of spectrally blurring important information that can result in the failure or reduced performance of analytical tools. On the other hand, if the spectral features of interest are sufficiently broad, this approach can be considered.

A more elegant method of band combination uses the *principal component* method to preserve most of the meaningful variability from the original imagery in a reduced set of orthogonal bands (cf. Sec. 9.3). In this process, only the first 10-20 principal component bands are maintained and used for subsequent image analysis. This process is illustrated in Figure 10.4. In the first step, bands in the strong atmospheric absorption windows are removed. In the second step, the remaining bands are transformed into the first ℓ principal component bands, where the value of ℓ is set by the user to preserve, for example, 99.5% of the variability in the data set; i.e., minimize ℓ subject to

$$\sum_{i=1}^{\ell} \lambda_i > 0.995 \text{ trace } (\mathbf{\Lambda}) \tag{10.1}$$

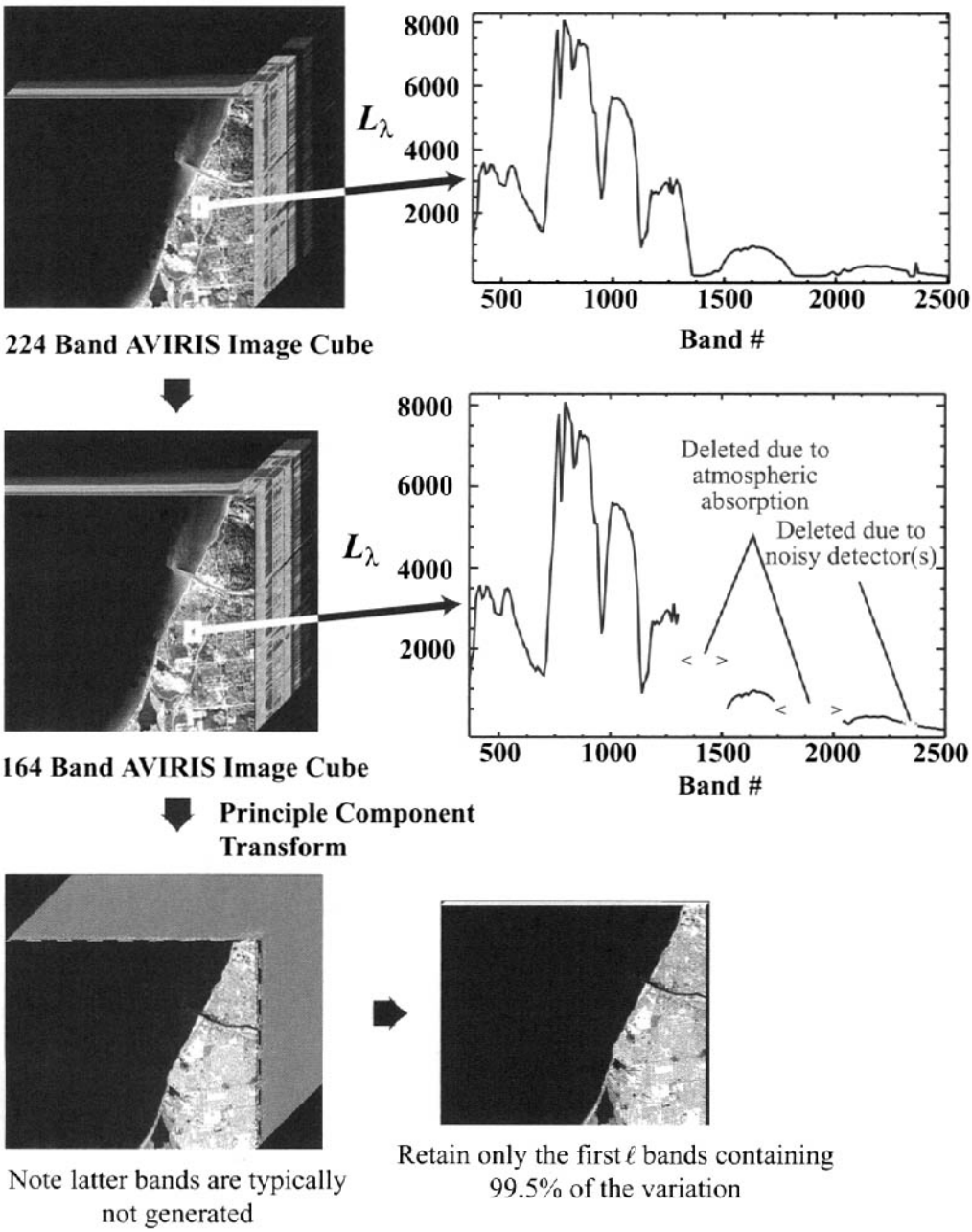


Figure 10.4 Typical steps in simple dimensionality reduction.

where λ_i are the eigenvalues in descending order of the image covariance matrix and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues. The resulting vectors no longer carry intuitive information about spectral content, but they are significantly more manageable computationally. The reader is cautioned that this approach implicitly assumes that information is proportional to variation. This is a reasonable assumption for very low-noise data sets and when the targets of interest are of sufficient number and/or contrast to influence the image covariance [cf. Shen (2002) for an alternative band reduction approach]. The next two subsections address this noise issue in greater detail.

10.2.2 Noise Characterization: Noise versus Clutter

Many panchromatic and multispectral imaging systems have *signal-to-noise* ratios of many hundreds (or even thousands) to one for the average scene element. As a result, the impact of noise on processing algorithms is often negligible. For spectroscopic image data, the push for high spectral resolution, while maintaining some level of spatial resolution, often pushes us to system designs with lower signal to noise. Consequently, we will often find it necessary to first characterize the noise associated with these systems and then attempt to reduce the impact of the noise on subsequent processing stages.

Because the noise associated with an imaging spectrometer can change significantly, depending on operating conditions, it is often desirable to characterize the noise associated with a specific image acquisition. For some systems, a dark-field image is acquired before and after each acquisition. This is equivalent to acquiring an image, or a portion of an image, with the shutter closed. The standard deviation of the resulting image in each band represents a best estimate of the dark noise. In addition, the covariance [cf. Eq.(9.71)] between any two bands of the dark-field image captures the correlated noise between two bands. For visualization purposes, we often show the covariance normalized by the standard deviations in the two bands, which is the correlation, i.e.,

$$\rho_{mn} = \frac{\sum_{q=1}^{N} [DC_m(q) - \overline{DC}_m][DC_n(q) - \overline{DC}_n]}{(N-1)\sigma_{mm}^{1/2}\sigma_{nn}^{1/2}} \tag{10.2}$$

where ρ_{mn} is the correlation between bands m and n, and $DC_m(q)$ is the qth digital counts from a noise sample of N pixels having mean $\overline{DC}_m$ and variance σ_{mm} in the mth band [cf. Eqs.(9.70) and (9.71)]. Figure 10.5 shows several ways to represent the noise from an imaging spectrometer. Note that when the correlations are assembled into a correlation matrix and displayed, we see not only the individual correlations between bands, but patterns associated with correlation levels in the noise within individual spectrometers (recall that AVIRIS has four separate spectrometers and, therefore, four independent detector arrays with separate electronics).

In some cases, a noise image is not available or the noise may be signal dependent such that the dark noise is not characteristic of the operational performance. In these cases, it may be necessary to attempt to derive estimates of the

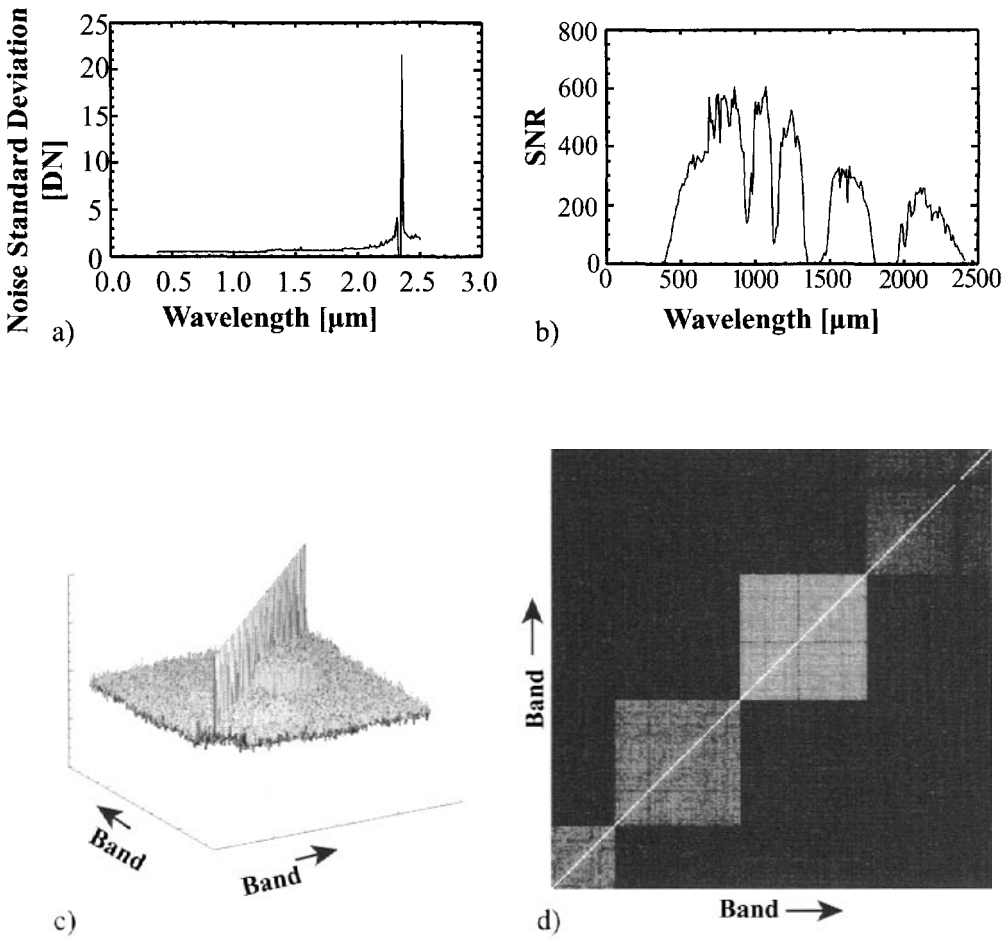


Figure 10.5 Various ways to represent the dark noise from an imaging spectrometer: (a) Plot of noise as a function of wavelength expressed in digital number (DN). (b) Plot of signal to noise as a function of wavelength from the signal associated with an 18% reflector at noon on the summer solstice in Rochester, NY. (c) Visualization of the noise correlation matrix. (d) Absolute value of the noise correlation matrix. All data shown are for an AVIRIS flight over Rochester, NY, 1999.

noise from the image data. If a spatially uniform region exists in the image, we can assume that all the variation in the brightness about the mean is due to instrument noise. In this case, we can simply use the local covariance over the uniform region as an estimate of the noise covariance.

In many scenes, no region with sufficient uniformity exists to permit the use of this approach. In this case, Green et al. (1988) suggest that we may be able to approximate the noise covariance by taking advantage of the local spatial stationary characteristic of most image data (particularly in relatively uniform regions of a

scene). Take the case where the local spatial correlation is quite high, then we can express the difference between adjacent pixels i and j in the mth band as:

$$DC_i(m) - DC_j(m) = \left([S_i(m) + n_i(m)] - [S_j(m) + n_j(m)]\right) \cong n_i(m) - n_j(m) = \Delta n(m) \tag{10.3}$$

where $DC_i(m)$ is the digital count for the ith pixel assumed to be composed of signal $S_i(m)$ plus a signal-independent noise $n_i(m)$. Green et al. (1988) point out that the noise covariance can be approximated from the covariance of the neighboring pixel differences (i.e., a spectral image formed by replacing each image vector with the difference between the pixel vector and the neighboring pixel vector) according to

$$\mathbf{\Sigma}_n \cong \frac{1}{2}\mathbf{\Sigma}_{\Delta n} \tag{10.4}$$

where $\mathbf{\Sigma}_n$ is the desired noise spectral covariance matrix and $\mathbf{\Sigma}_{\Delta n}$ is the matrix formed by computing the spectral covariance of the pixel difference image. In practice, the actual functional form of Eq. (10.4) depends on the spatial correlation of both the noise and the signal. Eq. (10.4) is a good estimate only if the noise is spatially uncorrelated and the signal is spatially highly correlated. Green et al. (1988) include a more complete treatment of the dependency of $\mathbf{\Sigma}_n$ on $\mathbf{\Sigma}_{\Delta n}$.

To this point, we have been characterizing the instrument noise. More specifically, we have suggested that the dark-field spectral noise covariance is a good estimate of the instrument-dependent variability in the image signal. For many of the target detection algorithms discussed in Section 10.4, we are interested in characterizing not just the spectral variability due to the instrument, but also the variability due to naturally occurring fluctuations in the spectral scene content. The imagewide spectral covariance matrix characterizes this variability for the entire scene. However, in many cases we will be interested in the spatially localized spectral variability (i.e., whether our candidate target sufficiently different from the local background to warrant interest). This variability is referred to as *spectral clutter* and can be characterized by the spectral covariance computed over a region of interest (ROI). Usually, these calculations are based on a window region about the target pixel or pixels. Both instrument-induced variability and scene-induced variability can be sources of confusion for analytical algorithms, as we will see in the discussion that follows. Note that spectral clutter, as described above, includes both variability due to instrument noise and scene-generated variability.

10.2.3 Noise-Sensitive Dimensionality Reduction

As discussed in Section 10.2.1, the principal components (PC) transform can be a useful method for dimensionality reduction when the signal-to-noise ratio is quite high (i.e., when essentially all the variability in the data set can be attributed to actual variation in scene brightness). In many cases, instrument noise is sufficiently large that it is of value to adjust the dimensionality reduction approach to account for the noise. Green et al. (1988) suggest using a transform called the maximum

noise fraction to accomplish this. Lee et al. (1990) point out that the maximum noise fraction transform is the equivalent to performing two PC transforms and their *noise-adjusted principal component* (*NAPC*) approach will be used here for clarity and consistency with the previous treatment of principal components. Recall that the PC ftransform uses eigenvectors to transform a spectral image into a new set of orthogonal dimensions that are rank ordered to maximize their ability to optimally account for the variability in the data (cf. Secs. 9.1 and 9.3). Clearly, if noise is a significant source of the variability in one or more of the bands, the weights assigned to that band(s) in the standard PC transform will overemphasize the importance of that band at the expense of others. The result is a set of PC-transformed bands that are sensitive to noise-based variability at a cost of scene-based variability. To overcome this limitation, we seek a preprocessing step that equalizes the amount of noise-related variability across all the bands. Note that ideally we would like to make all the noise-related variability go away, but, since we can't separate it from the scene variability on a per pixel basis, this noise equalization (i.e., *noise whitening*) step is the best alternative.

To understand the *noise equalization* step, we will conceptually process a dark-field image (i.e., an image where all the variation is due to noise). Let $\mathbf{\Sigma}_n$ represent the spectral covariance matrix associated with this noise image. Then, if we compute the eigenvectors of $\mathbf{\Sigma}_n$ (cf. Sec. 9.1) and apply a PC transform to our dark-field image, we obtain a new image where the bands are orthogonal and the variance in each band is equal to the eigenvalues of $\mathbf{\Sigma}_n$. If we divide each transform band by the square root of the eigenvalue ($\sqrt{\lambda_{ni}}$) corresponding to that *i*th band (i.e., the standard deviation of the variability in the band), the resulting bands in the transformed dark-field noise image will all have equal variability. In practice, we don't want to operate on the dark-field image. Instead, we perform the same operation on the actual image, making the assumption that the scene (or signal) covariance $\mathbf{\Sigma}_s$ and the noise covariance are independent, i.e.,

$$\mathbf{\Sigma} = \mathbf{\Sigma}_s + \mathbf{\Sigma}_n \tag{10.5}$$

where $\mathbf{\Sigma}$ is the observed covariance of the image. We then transform the image data through a PC transform based on the eigenvectors of the noise covariance matrix (computed using one of the methods described in Sec. 10.2). The variance in each band of the transformed image due just to noise should be equal to the noise covariance. Thus, if we divide the brightness in each of the transformed bands by $\sqrt{\lambda_{ni}}$, the resulting transformed images should have the same variability due to noise in each band (i.e., we have whitened the noise). We can think of this process as operating on each original image vector $\mathbf{x}$ to generate a corresponding vector $\mathbf{z}$ in the noise-whitened space according to

$$\mathbf{z} = \left(\mathbf{E}_n \mathbf{\Lambda}_n^{-\frac{1}{2}} \right)^{\mathrm{T}} \mathbf{x} \tag{10.6}$$

or for any output band *i* the computation would look like:

$$z_i = \frac{1}{\sqrt{\lambda_{ni}}} \mathbf{e}_{ni}^{\mathrm{T}} \mathbf{x} \tag{10.7}$$

where z_i is the ith element in the noise-whitened PC transform vector, $\mathbf{\Lambda}_n$ is the diagonal matrix of eigenvalues (λ_{ni}) of the noise covariance matrix, and $\mathbf{E}_n$ is the matrix made up of the eigenvectors ($\mathbf{e}_{ni}$) of the noise covariance matrix as columns.

Now that the noise contribution is the same in all bands of z space, we can apply the PC transform to the noise-whitened data set. To do this, we simply compute the covariance ($\mathbf{\Sigma}_z$) of the noise-whitened image (i.e., of the image cube expressed as $\mathbf{z}$ vectors) and use the eigenvectors of $\mathbf{\Sigma}_z$ to transform the $\mathbf{z}$ vectors into the noise-adjusted PC space according to

$$\mathbf{y} = \mathbf{E}_z^T \mathbf{z} \tag{10.8}$$

where $\mathbf{E}_z$ is the matrix made up of the eigenvectors ($\mathbf{e}_z$) of $\mathbf{\Sigma}_z$ as columns. The resulting PC bands will be rank ordered in terms of how much of the variance in signal (i.e., independent of noise) they describe. In fact, Lee et al. (1990) point out that the eigenvalues of $\mathbf{\Sigma}_z$ will equal the signal to noise plus one, and therefore, the corresponding eigenvectors can be readily interpreted in terms of the importance of the resulting NAPC band (i.e., as the eigenvalues approach one, the bands are nearly all noise cf. Fig. 10.6). Thus, the higher order (lower SNR) bands can be removed from subsequent analysis to reduce the dimensionality of the data. Because the NAPC process is reversible (i.e., we can invert NAPC images back to the original spectral space), we can take advantage of the noise segregation process to filter (smooth) the noisier bands in NAPC space and then inverse transform the reduced noise data into the original spectral bands. This might be attractive for use with spectral feature analysis algorithms as discussed in Section 10.5. In many cases, we will find it more convenient to simply truncate the later bands in the NAPC space and work on the lower dimensional data.

Note that the maximum noise fraction algorithm of Green et al. (1988) achieves the same end as NAPC by computing the eigenvectors of $\mathbf{\Sigma}_n \mathbf{\Sigma}^{-1}$ and transforming the data by projection onto the eigenvectors. The resulting bands are ranked by the eigenvectors in terms of lowest SNR to highest. Thus, the band ranking is reversed relative to NAPC. Thus, it has become common convention to use the term *minimum noise fraction* (*MNF*) transform and reverse the order of the PC-transformed bands of $\mathbf{\Sigma}_n \mathbf{\Sigma}^{-1}$ (i.e., the band with the lowest eigenvalue is the first MNF band). This corresponds directly with the NAPC transform described above.

10.2.4 Estimation of the Dimensionality of a Data Set

In many cases, we will find it necessary, or at least convenient, to have an estimate of the inherent *dimensionality* of an image. One simple way to do this is to define the dimension of the data as the number of eigenvector dimensions needed to explain some fraction of the variability in the data (see the discussion of eigenvectors in Chap. 9). When working with data over a period of time, we develop an estimate of the expected level of variability that is likely to be attributable to image informa-

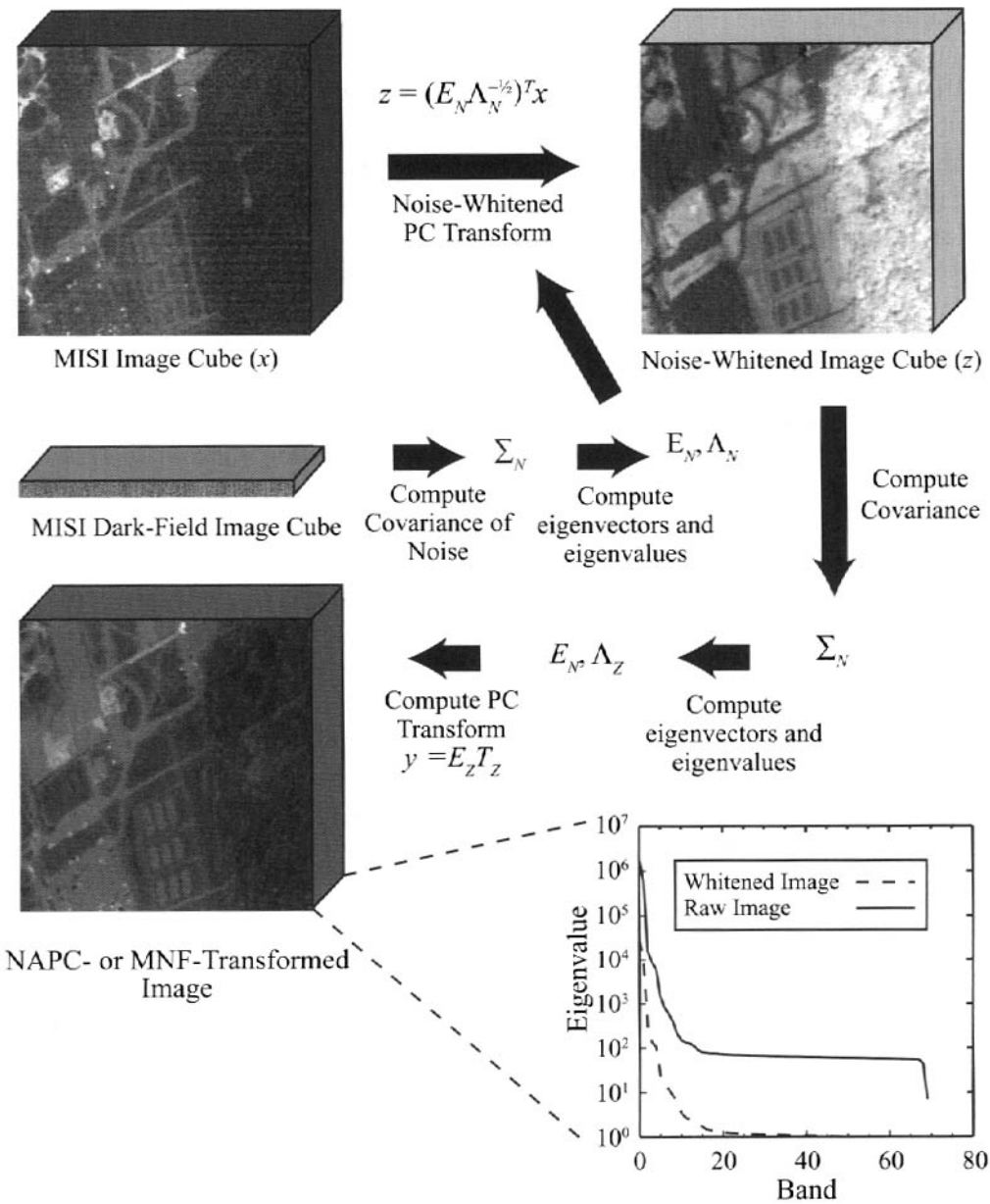


Figure 10.6 Illustration of the NAPC (MNF) process applied to a 72-band image from RIT's Modular Imaging Spectrometer Instrument (MISI) acquired during a noisy flight test. The eigenvalues of the transformed bands are plotted at the bottom of the figure.

tion as opposed to noise for various scene types. For example, given the Landsat TM scene shown in Figure 9.20, we might have set the threshold at 99%. For other systems with very high SNRs or when we are looking for very subtle spectral signatures, this value might be 99.9%. To gain experience with this approach, we often generate PC images and visually assess when we appear to lose significant spatial scene content in the PC transform images. When the NAPC transform described in Section 10.2.3 is used, we can look at the magnitude of the eigenvalues of the noise-whitened covariance matrix. As these values approach one, the SNR of the corresponding bands is approaching zero (these eigenvalues equal the SNR + 1), and we can assume that the bands beyond that point will be dominated by noise. Note that the thresholding here is still somewhat arbitrary since the eigenvalues often slowly approach unity and the user must define a cutoff point.

An alternative method uses the *singular value decomposition* (SVD) to compute the rank of a data set and defines that rank to be the inherent dimensionality [cf. Slater and Healey (1998)]. If we take every pixel in an image and enter it into a matrix as a spectral column vector, then an $m \times k$ dimension image with ℓ spectral bands will form a two-dimensional matrix $\mathbf{M}$ of dimension $[(m \times k) \times \ell]$. If we then apply the SVD process described in Section 9.1.3, we can determine the number of nonzero singular values of the image matrix (i.e., the rank of $\mathbf{M}$) and let the rank define the inherent dimensionality. To make the data processing less cumbersome, the matrix $\mathbf{M}$ can often be made of a characteristic sample of the image vectors (e.g., every nth pixel) with little loss in rigor. This method is attractive in that we do not have to have any previous experience with a data set to compute the inherent dimensionality; however, we do have to concede the processing time.

Before closing this section, it is important to remember that some information can be lost in the dimensionality reduction process. If the objective is to find low probability of occurrence targets with low spectral contrast, we may want to be conservative in any dimensionality reduction (i.e., choose to retain more bands in the reduced dimension space).

10.3 GEOMETRIC OR DETERMINISTIC APPROACHES TO SPECTRAL IMAGE ANALYSIS

As introduced in Section 10.1 and illustrated in Figure 10.1, the geometric perspective on spectral image analysis tends to describe the data space as comprising a convex hull where the extrema are often referred to as end members. In the process of analyzing the data, we can describe at least four types of algorithms that can be used to characterize a scene. The first type of algorithm attempts to classify the scene by assigning every pixel to a class or category. We are familiar with this from the GML classifier introduced in Chapter 9. A second type of algorithm attempts to locate pixels that represent a target class. These *target detection* algorithms can be thought of as classification algorithms where we attempt to classify each pixel into either a target or background class. However, as we will see, most target detection algorithms are aimed at finding scarce or low probability of occurrence targets and tend to take on a different form from classification algorithms where the classes

are assumed to have a relatively high a priori probability. A third type of algorithm attempts to find pixels with unusual or anomalous spectral features. These *anomaly detectors* can be thought of as target detection algorithms where the target is unknown. Thus, we still seek to separate the background into a large class, but we place any pixel whose spectrum is significantly different than the background into the unknown target or anomaly class. The fourth type of algorithm is in some ways the most demanding in that it attempts to describe the fractional abundance of a material type that may be present in each pixel. These *unmixing* algorithms go considerably beyond the classifiers to describe each pixel as being composed of one or more classes and then attempts to describe the fractional makeup of each class in each pixel. While other types of spectral analysis algorithms exists, these four dominate the field, and we will tend to assign most algorithms to one of these categories (cf. Secs. 10.3.2 and 10.3.3).

Before introducing these algorithms, we first need to characterize the image data into a geometric or structured model. Several methods to accomplish this are introduced in the next section.

10.3.1 End Member Selection

The most obvious, and often most difficult, method of *end member* selection is to attempt to interactively select end members from the image by identifying pure examples of each land cover category by inspection of the spatial images. This can be augmented by looking at spectral plots of the data and trying to locate extremes in the spectral plots. The difficulty here quickly becomes apparent when we recognize the enormous number of band combinations that may need to be investigated for either spatial or spectral analysis. As a result, most of the end-member selection methods rely heavily on automated, or at least semiautomated, approaches that are often run on reduced-dimension versions of the spectral data (cf. Sec. 10.2). The three automated methods of end-member selection introduced here can all be run on either full-dimension or reduced-dimension data and in any radiometric space (i.e., digital count, radiance, or reflectance). In addition, since all of these methods are just selecting pixels from the scene to serve as end members, the full spectral detail is available for subsequent processing even when the selection process occurs on the reduced-dimension data.

The first method is referred to as the *pixel purity index* (*PPI*) and uses a two-step process, first to automatically nominate pixels as end members, and second to interactively decide which of the nominated pixels or combinations of pixels to choose [cf. Boardman et al. (1995)]. In the nomination process, a vector is randomly generated in the spectral space to be searched (usually reduced dimension space). All the pixels in the image are then projected onto the random vector, and those with the most extreme values (highest and lowest or most negative) are incremented in an image map of candidate end members. Boardman et al. (1995) suggest that not just the most extreme, but also pixels within the expected noise range of the extremes be included to ensure that the candidate end members are not dominated by noisy pixels. This process is repeated on one random vector after

another until a stopping criterion is met (e.g., some fixed number of random vectors or number of different candidate end members). Boardman et al. (1995) suggest a stopping criteria based on the cumulative number of candidate end members. They suggest that the cumulative number of different candidate pixels should rise relatively rapidly and then approach a constant as all the pixels on the borders of the convex hull have been nominated. By plotting the cumulative number of candidates, a user can interactively stop the process or program a stop when the rate of increase falls below some threshold. At this point, the user may have hundreds of candidate end members. Boardman et al. (1995) suggest the final end members be selected by interactively viewing the candidates in three-band combinations to try to visually isolate clusters that represent vertices of the convex hull. The final end members can be individual pixels or averages of a small cluster that characterize the vertex (cf. Fig. 10.7). The final output from this process is a set of end-member pixels that can be used in either the original or any transformed space.

An alternative approach called *N-FINDR* suggested by Winter (1999) is designed to automatically find the end members in a spectral data set by selecting the set of pixel vectors that form the simplex that occupies the largest volume in the data space. The data are first reduced in dimension using one of the methods discussed in Sections 10.2.1 or 10.2.3. Winter (1999) suggests that the data in this reduced-dimension space will still form a convex hull that can, to first order, be described by a simplex made up of one more vertex than the dimension of the space. The method assumes that the vertices of the simplex can be approximated by end members that occur in the image (i.e., there is a reasonably pure example of the land cover type in the scene). The problem then becomes one of finding the N end members in the $N-1$ dimension space. Winter (1999) points out that the simplex that optimally describes the convex hull using scene-derived end members should enclose the largest volume. This means that the N pixel vectors that form the largest volume should also be the set of end members we seek. The volume enclosed by N vectors $(\boldsymbol{q}_1,\boldsymbol{q}_2 \ldots \boldsymbol{q}_N)$ can be expressed as

$$V = \frac{1}{(N-1)!} abs\left(|\mathbf{Q}|\right) \tag{10.9}$$

where

$$\mathbf{Q} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ \boldsymbol{q}_1 & \boldsymbol{q}_2 & & \boldsymbol{q}_N \end{bmatrix} \tag{10.10}$$

[i.e., $\mathbf{Q}$ is the matrix made up of N column vectors $(\boldsymbol{q}_1,\boldsymbol{q}_2 \ldots \boldsymbol{q}_N)$ augmented by a row of ones] and abs is the absolute value operation. Winter (1999) suggests performing an iterative search to find the set of pixel vectors that maximizes the volume expressed by Eq. (10.9). This is accomplished by randomly selecting N vectors (pixels) to form the $\mathbf{Q}$ matrix [Winter (1999) suggests that, rather than completely random selection, one pixel may be chosen to represent the darkest scene elements commonly referred to as the *shade end member*] and then testing each pixel in each location in the $\mathbf{Q}$ matrix to determine if the volume is increased and leaving it in the location that maximally increases the volume. Because most data sets adhere

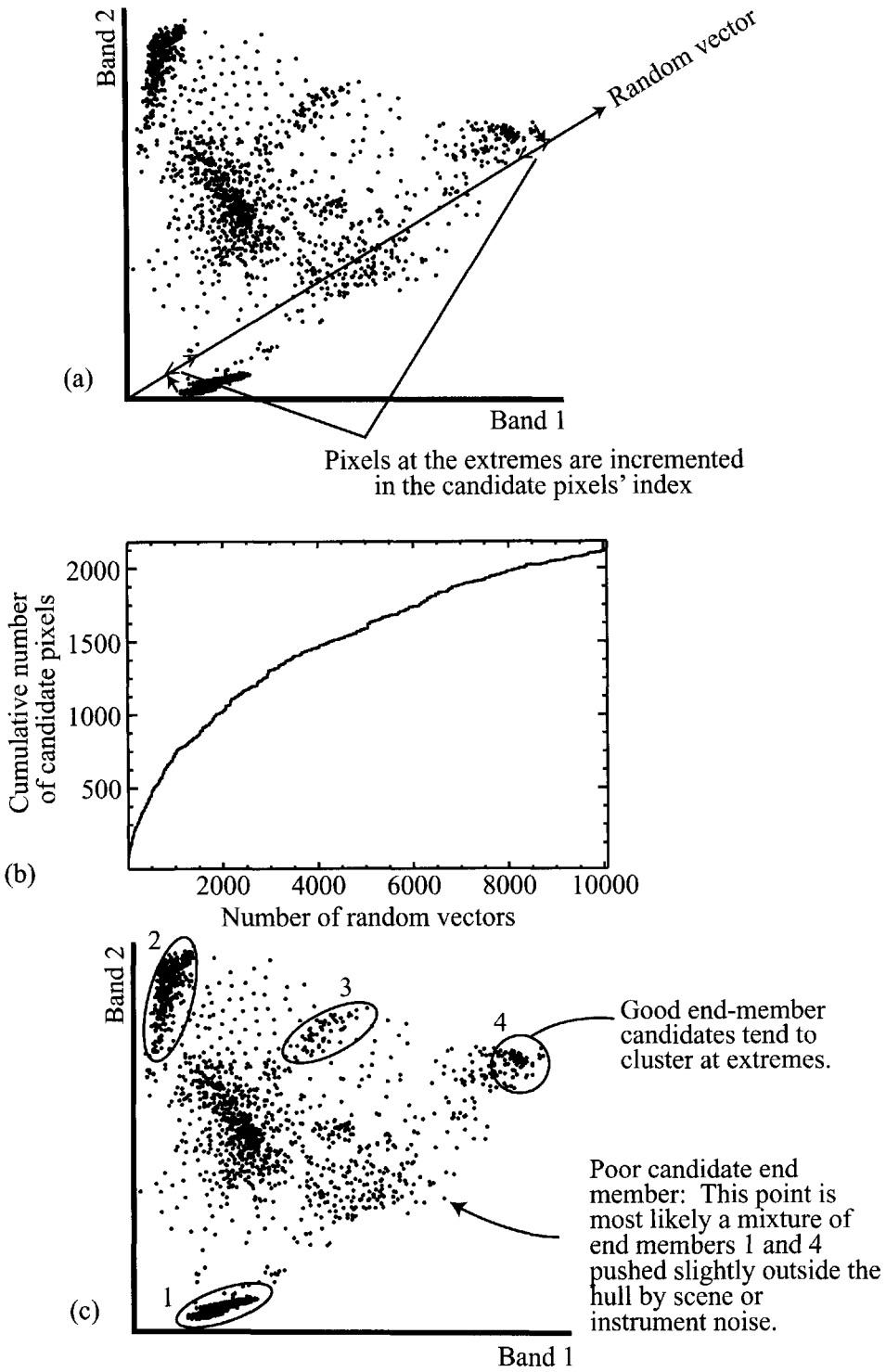


Figure 10.7 Illustration of end-member selection steps using the Pixel Purity Index (PPI) method: (a) projection of pixels onto a random vector and selection of extreme values as candidates (note: only two dimensions shown to simplify the illustration), (b) illustration of cumulative distribution used to determine how many random vectors to use, and (c) 2-D plot of how the candidates and pixel vectors can be visualized in the final end-member selection process.

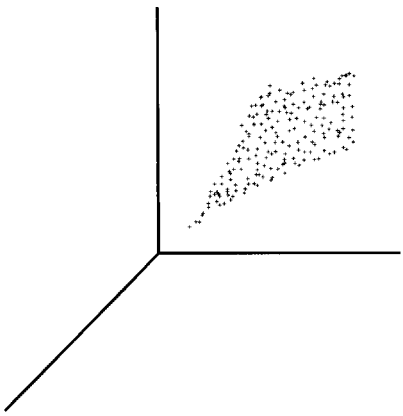
reasonably well to the convex hull model and include good candidate end members in the scene, Winter (1999) suggests that this simple method should be successful without finding local maxima. However, he further suggests that multiple seeds can be employed to initiate the search and the final volumes compared to ensure that a global maximum is achieved.

A third method for end member selection described by Lee (2003) and Schott et al. (2003) attempts to achieve fully automated performance at high speed. Like the PPI and N-FINDR approaches, this *max-D* method uses the convex hull perspective and assumes that representative end members exist within the scene. It uses a dimensionality assessment method such as SVD (cf. Secs. 9.1 and 10.2.4) to define the inherent dimensionality of the data (i.e., $N-1$). The max-D method then assumes that a simplex made of N vertices can be defined to encompass the $N-1$ dimensional data cloud that represents the image data. Note that for computational convenience, the max-D processing is usually performed on the reduced dimension data set. The N vertices that encompass the data space are used as end members in subsequent analysis. The question is how to define the vertices of the simplex. Lee (2003) shows that the pixel vector with the largest magnitude must be a vertex of the simplex for the convex hull model to hold (cf. Fig. 10.8). He postulates that the smallest magnitude vector is also a good candidate vertex, though it can be proven that, for arbitrary data sets, the smallest magnitude vector does not have to be a vertex of the simplex. For real data sets, the spectral vectors tend, due to shadowing effects, to form a cone with an apex near zero radiance [cf. Craig (1994) and Ifarraguerri and Chang (1999)] so that the smallest magnitude vector is generally used as a second vertex. The next step in the max-D process involves projecting the first vertex $\mathbf{v}_1$ onto the second $\mathbf{v}_2$ and projecting all the image data in the same fashion onto the hyperplane perpendicular to the vector $\mathbf{v}_2 - \mathbf{v}_1$. Lee (2003) points out that, after projection, the data will occupy a reduced dimension ($N-2$) and that the vertices of a simplex encompassing the data in the $N-1$ dimension space will also be vertices of a simplex encompassing the data in the $N-2$ dimensional space. Furthermore, the pixel having a maximum distance from any point inside or on the surface of a simplex is a vertex of the simplex. Thus, if in the reduced dimension, we find the pixel most distant from the common vertex $\mathbf{v}_{12}$ (i.e., the projection of 2 onto 1), it must be a third vertex $\mathbf{v}_3$. We then repeat the earlier process of projecting the data onto the hyperplane perpendicular to the $\mathbf{v}_3 - \mathbf{v}_{12}$ vector, which further reduces the dimensionality of the data and projects $\mathbf{v}_3$ onto $\mathbf{v}_{12}$. This process is repeated until all N vertices are located. Note that while the vertices are being located in various projection spaces, they are still pixels in the original data set, and their value in the original space or any of the projected spaces can be used in subsequent analysis.

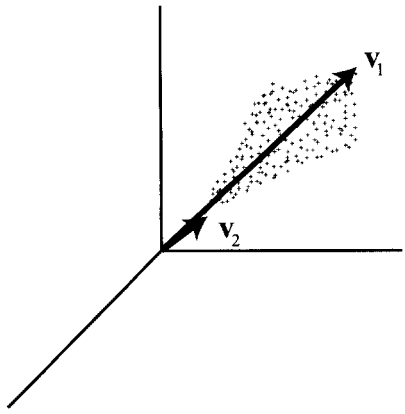
The steps in the max-D process are summarized below with respect to the simplified low, dimensional problem shown in Figure 10.8.

Step 1: Define the inherent dimensionality of the data (e.g., using SVD; cf. Sec. 9.1).

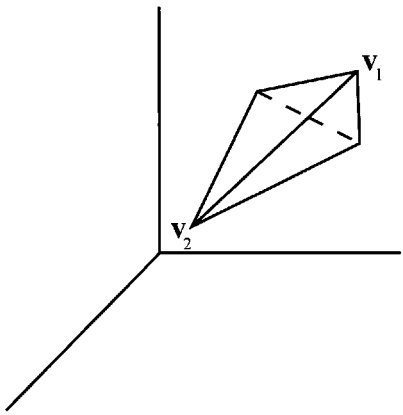
Step 2: Transform the data into a reduced-dimensional space[cf. Fig. 10.8(a)] to simplify further processing (e.g., using an MNF transform; cf. Sec 10.2.3).



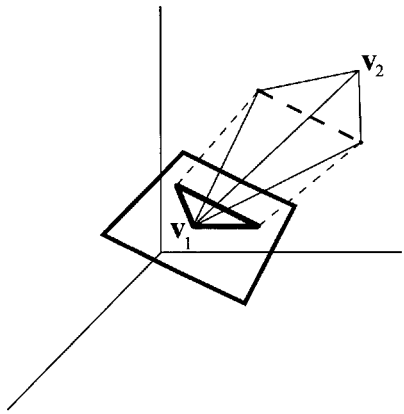
(a) Transform data to a reduced-dimensional space.



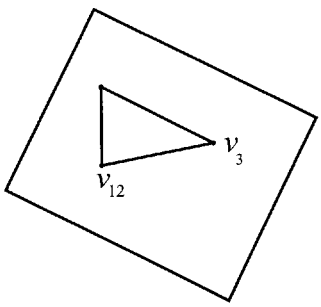
(b) Compute magnitudes and select minimum and maximum as vertices.



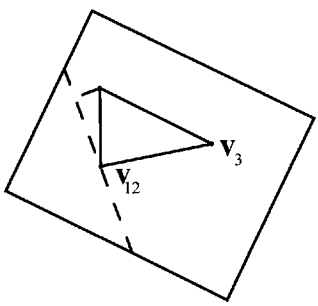
(c)Simplification of the data space for clarity, where we have isolated the convex hull and the vertices we are seeking.



(d) Projection of $\mathbf{v}_2$ onto $\mathbf{v}_1$ and the data onto the hyperplane perpendicular to the vector difference $\mathbf{v}_2$—$\mathbf{v}_1$.



(e) Select the pixel in the projected space more distant from $\mathbf{v}_{12}$ as the next vertex ($\mathbf{v}_3$).



(f) Projection of $\mathbf{v}_3$ onto $\mathbf{v}_{12}$ and the data onto the hyperplane perpendicular to the vector difference $\mathbf{v}_{12}$—$\mathbf{v}_3$.

Figure 10.8 Illustration of the steps in the Max-D process.

Step 3: Compute the magnitude of all vectors in the transform space:

$$\|\mathbf{z}\| = \left(\mathbf{z}^{\mathrm{T}}\mathbf{z}\right)^{1/2} \tag{10.11}$$

and select the vectors with the maximum and minimum magnitude as vertices $\mathbf{v}_1$ and $\mathbf{v}_2$ of the simplex that describes the convex hull [cf. Fig. 10.8(b)].

Step 4: Compute the operator that projects $\mathbf{v}_2$ onto $\mathbf{v}_1$ and apply it to the data set [cf. Fig. 10.8(d)]. The *orthogonal projection operator* will null the difference vector (d) between $\mathbf{v}_1$ and $\mathbf{v}_2$, i.e., for

$$\mathbf{d} = \mathbf{v}_2 - \mathbf{v}_1 \tag{10.12}$$

The *nulling* or orthogonal projection operator ($\mathbf{P}^{\perp}$) is given by

$$\mathbf{P}^{\perp} = \mathbf{I} - \mathbf{d}\mathbf{d}^{\#} = \mathbf{I} - \mathbf{d}(\boldsymbol{d}^{\mathrm{T}}\boldsymbol{d})^{-1}\boldsymbol{d}^{\mathrm{T}} \tag{10.13}$$

where $\mathbf{I}$ is the identity matrix and we have introduced the matrix *pseudo-inverse* $\mathbf{d}^{\#}$ given by

$$\mathbf{d}^{\#}=(\mathbf{d}^{\mathrm{T}}\mathbf{d})^{-1}\mathbf{d}^{\mathrm{T}} \tag{10.14}$$

which has the convenient property that

$$\mathbf{d}^{\#}\mathbf{d} = (\mathbf{d}^{\mathrm{T}}\mathbf{d})^{-1}\mathbf{d}^{\mathrm{T}}\mathbf{d} = \mathbf{I} \tag{10.15}$$

i.e., premultiplication of a matrix (or vector) by its pseudo-inverse yields the identity matrix. From this property, we see that the orthogonal projection or nulling operator $\mathbf{P}^{\perp}$ when applied to the data set will project $\mathbf{v}_2$ onto $\mathbf{v}_1$, nulling the difference according to

$$\mathbf{P}^{\perp}\mathbf{v}_2 - \mathbf{P}^{\perp}\mathbf{v}_1 = \mathbf{P}^{\perp}(\mathbf{v}_2 - \mathbf{v}_1) = \mathbf{P}^{\perp}\mathbf{d} = \left[\mathbf{I} - \mathbf{d}\mathbf{d}^{\#}\right]\mathbf{d} = \mathbf{I}\mathbf{d} - \mathbf{d}\mathbf{I} = \mathbf{0} \tag{10.16}$$

The rest of the data are projected by this operator onto the hyperplane perpendicular to the difference vector.

Step 5: In the projection space, find the pixel vector most distant from $\mathbf{v}_{12}$ and let it be the next vertex $\mathbf{v}_3$:

$$\max\|\mathbf{d}\| = \max\left(\left[\hat{\mathbf{z}} - \mathbf{v}_{12}\right]^{\mathrm{T}}\left[\hat{\mathbf{z}} - \mathbf{v}_{12}\right]\right)^{1/2} \tag{10.17}$$

where $\|\mathbf{d}\|$ is the magnitude of the vector difference, $\hat{\mathbf{z}}$ is a vector in the projected space, and $\mathbf{v}_{12}$ is the projection of $\mathbf{v}_2$ onto $\mathbf{v}_1$ in the composite space (i.e., the common point).

Step 6: Loop on steps 4 and 5 until all N vertices are located [i.e., the next step is to project $\mathbf{v}_3$ onto $\mathbf{v}_{12}$ and the data onto the hyperplane perpendicular to $(\mathbf{v}_3 - \mathbf{v}_{12})$].

In their simplest form, all of the scene-derived end-member selection techniques (i.e., PPI, N-FINDR, and max-D) are sensitive to selection of noisy pixels as end members. Lee (2003) suggests that because max-D runs quite rapidly, the

first set of vertices should be removed from the data set and the process repeated until consistent vertices appear (i.e., the spectral vectors are very close to each other). In practice, Lee (2003) found that for low-noise data from the AVIRIS sensor, the first- and second-run vertices selected by max-D were nearly identical. Even though in this case the initial vertices were valid, the reader should be cautious that with automated approaches, a built-in validity check should be included because a few noisy pixels could severely distort the end-member selection process.

In this section, we have defined three ways to characterize the data space using the convex hull geometric representation of the data introduced in Section 10.1.1. In the next two sections, we will explore algorithms that take advantage of the scene-derived end members to analyze spectroscopic imagery.

10.3.2 Detection and Mapping Algorithms Using the Geometric or Structured Perspective

The *spectral angle mapper* (*SAM*) is one of the most widely used algorithms for making simple comparisons between spectral vectors [cf. Kruse et al. (1993)]. Recall from vector algebra (cf. Sec. 9.1) that the projection of a vector **x** onto a second vector or axis (**a**) is the inner or dot product of **x** with the unit vector in the **a** direction, i.e.,

$$\boldsymbol{x} \bullet \boldsymbol{a}_u = \boldsymbol{x}^T \boldsymbol{a}_u = \boldsymbol{x}^T \left[\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} \right] = \|\boldsymbol{x}\| \cos\theta \tag{10.18}$$

where θ is the angle between the vectors and is referred to as the spectral angle[cf. Fig. 10.9(a)]. From Eq. (10.18), it is clear that the spectral angle can be computed as:

$$\theta = \cos^{-1} \left[\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|} \right]^T \left[\frac{\boldsymbol{a}}{\|\boldsymbol{a}\|} \right] = \cos^{-1} [\boldsymbol{x}_u \bullet \boldsymbol{a}_u] \tag{10.19}$$

and that it is independent of the magnitude of the vectors (i.e., it is only dependent on the spectral direction or "color" and not on the magnitude or brightness). Thus, the spectral angle is commonly used to describe the spectral similarity between vectors. As a result, the spectral angle can be used in a target detection algorithm if the target vector of interest is known. In this case, the angle between every pixel vector and the target vector is computed and pixels with spectral angles below some threshold θ_t are identified as targets [cf. Fig. 10.9(b)]. In a similar fashion, the spectral angle can be used to map pixels to classes if we let a set of vectors define the spectral character of the classes of interest in a scene (e.g., land cover categories). For example, these vectors may be the end-member vectors selected using the approaches described in Section 10.3.1. Then, we can assign each pixel to the class to which it is most similar, as defined by having the smallest spectral angle [cf. Fig. 10.9(c)]. Note that it is also possible to use this approach to find pixels that are simply unlike the background. For example, rather than (or in addition to) mapping each pixel to a class, we could assign a pixel with a spectral angle of more

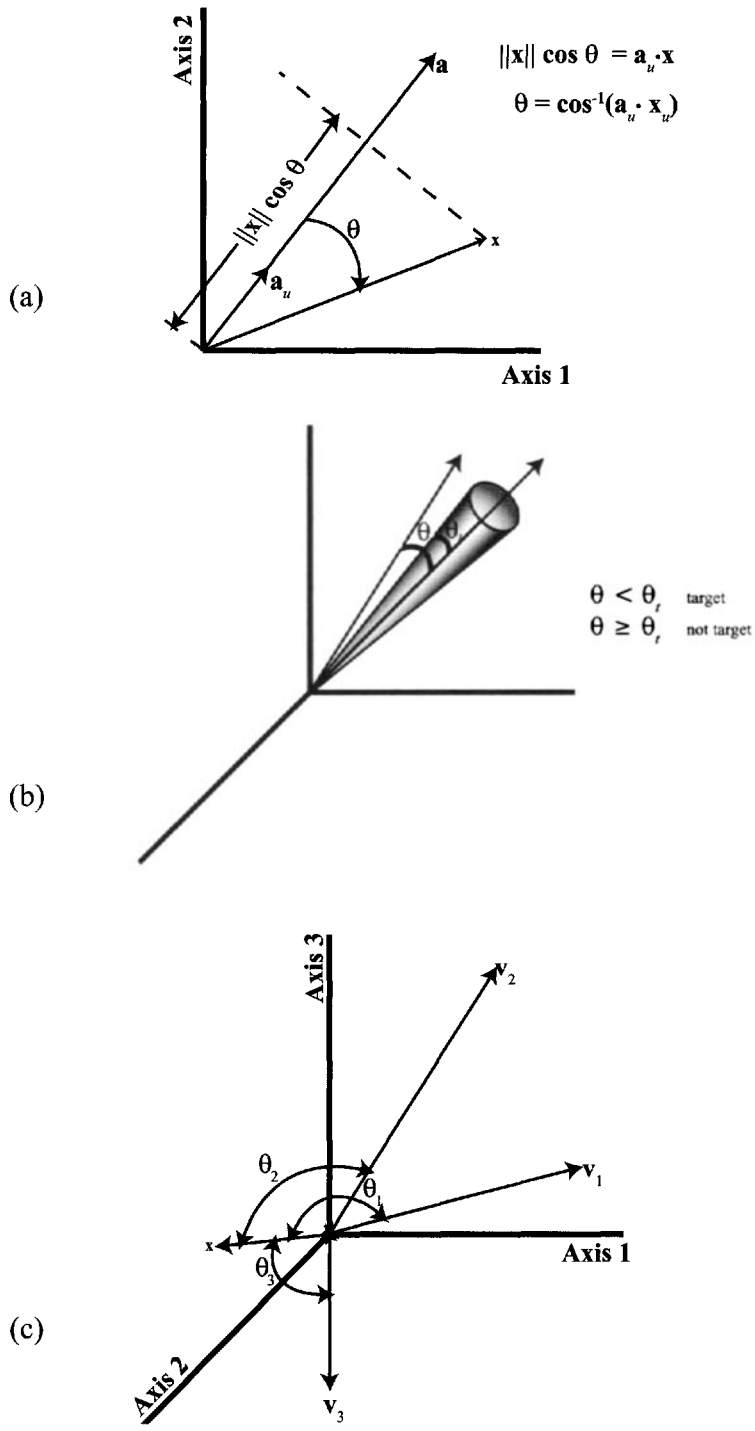


Figure 10.9 Illustration of the spectral angle concept: (a) for the simple two-dimensional case, (b) used in target detection, and (c) used for classification where we assign **x** to the class associated with the end member with the smallest spectral angle.

than some threshold value from any end member to an anomaly category. Thus, we see that the spectral angle is a simple metric that can be used in conjunction with a target vector for target detection or in conjunction with end member vectors for anomaly detection or scene classification. Note that the spectral angle concept can be applied in any of the multidimensional spaces we have discussed (i.e., digital count, radiance, reflectance, PC, MNF, etc.) as long as the target or end-member vectors are known in that same space and the user can define appropriate thresholds if applicable. Also note that, in most cases, these thresholds are set interactively. Because the SAM approach is quite simple to use, it often serves as a starting or comparison point for spectral image analysis. However, the SAM approach uses only a single metric for analysis, and as a result, it tends to ignore brightness variation, interaction (mixing) effects, and within-class variability, which may be important variables impacting certain analysis problems, hence the plethora of other spectral algorithms, some of which will be presented here.

Harsanyi and Chang (1994) suggest a target detection algorithm using a structured background perspective that seeks to simultaneously suppress the effects of the background and match the target spectrum to the observed spectrum. Furthermore, this approach, known as *orthogonal subspace projection* (*OSP*), is designed to be effective against targets at the *subpixel* level (i.e., where the observed signal is a mixture of target and background radiance). The OSP algorithm uses a structured mixing model to describe any given spectral vector (**x**) according to

$$\mathbf{x} = \mathbf{t}a + \mathbf{B}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} = \mathbf{H}\mathbf{f} + \boldsymbol{\varepsilon} \tag{10.20}$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_\ell \end{bmatrix} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_\ell \end{bmatrix} a + \begin{bmatrix} v_{11} & v_{12} & & v_{1k} \\ v_{21} & v_{22} & \cdots & v_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ v_{\ell 1} & v_{\ell 2} & & v_{\ell k} \end{bmatrix} \begin{bmatrix} \mathrm{f}_1 \\ \mathrm{f}_2 \\ \vdots \\ \mathrm{f}_\mathrm{k} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_\ell \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} & & v_{1k} & t_1 \\ v_{21} & v_{22} & \cdots & v_{2k} & t_2 \\ \vdots & & & & \\ v_{\ell 1} & v_{\ell 2} & & v_{\ell k} & t_\ell \end{bmatrix} \begin{bmatrix} \mathrm{f}_1 \\ \mathrm{f}_2 \\ \vdots \\ \mathrm{f}_\mathrm{k} \\ a \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_\ell \end{bmatrix} \tag{10.21}$$

where **x** is an ($\ell \times 1$) spectral vector, **t** is an ($\ell \times 1$) spectral vector characterizing the target, a is the unknown fractional abundance of the target, **B** is an ($\ell \times k$) matrix made up of the k basis vectors ($\mathbf{v}_1$ through $\mathbf{v}_k$) spanning the image space (cf. Sec. 9.1), $\boldsymbol{\alpha}$ is a ($k \times 1$) vector of unknown weights (fractions) applied to each of the k basis vectors in **B**, $\boldsymbol{\varepsilon}$ is the $\ell \times 1$ vector of residual error in the mixture model [assumed Gaussian independent and identically distributed (iid); i.e., the covariance of the $\boldsymbol{\varepsilon}$ vectors is equal to $\sigma^2\mathbf{I}$)], **H** is the [$\ell \times (k + 1)$] matrix formed by concatenating **B** and **t,** and **f** is the [$(k + 1) \times 1$] vector of unknown weights (i.e., it is the k elements making up the vector $\boldsymbol{\alpha}$ augmented by a). The algorithm assumes that **t** is known either by identifying a fully resolved target pixel in the image or from library spectra (e.g., if the image cube is in reflectance units and the target signature exists in an available spectral reflectance library). Furthermore, it assumes that the basis vectors characterizing the background (i.e., spanning the image) can be identified. The end members identified using any of the methods described in Section 10.3.1 form an appropriate set of basis vectors, and in this case, the weights in the

vector $\boldsymbol{\alpha}$ can be thought of as the fractional abundance of each end member when a linear mixing model is applied.

The OSP algorithm does not attempt to solve for the fractional abundance values. Rather, it can be thought of as a two-step process. The first seeks to minimize the effect of the background end-members, and the second applies a matched filter to see if, after background suppression, anything target-like can be found. The background minimization is accomplished by projecting each vector $\mathbf{x}$ onto a sub-space that is orthogonal to the columns of $\mathbf{B}$. This uses the nulling operator from least-squares theory introduced in the max-D algorithm (Sec. 10.3.1) to suppress the background according to

$$\begin{aligned}\mathbf{z} = \mathbf{P}_b^{\perp}\mathbf{x} &= \mathbf{P}_b^{\perp}\mathbf{t}a + \mathbf{P}_b^{\perp}\mathbf{B}\boldsymbol{\alpha} = \mathbf{P}_b^{\perp}\mathbf{t}a + \left[\mathbf{I} - \mathbf{B}\mathbf{B}^{\#}\right]\mathbf{B}\boldsymbol{\alpha} \\ &= \mathbf{P}_b^{\perp}\mathbf{t}a + \mathbf{B}\boldsymbol{\alpha} - \mathbf{B}\left(\mathbf{B}^{\#}\mathbf{B}\right)\boldsymbol{\alpha} = \mathbf{P}_b^{\perp}\mathbf{t}a + \mathbf{B}\boldsymbol{\alpha} - \mathbf{B}\mathbf{I}\boldsymbol{\alpha} = \mathbf{P}_b^{\perp}\mathbf{t}a + \mathbf{P}_b^{\perp}\boldsymbol{\varepsilon}\end{aligned} \tag{10.22}$$

where $\mathbf{P}_b^{\perp} = \mathbf{I} - \mathbf{B}\mathbf{B}^{\#}$ is the projection or nulling operator, we have dropped the error term except for the last expression for clarity of presentation, and $\mathbf{z}$ is the projection of $\mathbf{x}$ onto the subspace orthogonal to the background end-member vectors. Harsanyi and Chang (1994) point out that, to the extent the model is valid, the resulting $\mathbf{z}$ vectors should only have energy (variance) associated with the target signature and the random error term. This is clear from Eq. (10.22), where the $\mathbf{B}$ matrix does not contribute to the new projection space. For the second step in the OSP process, we seek an operator ($\mathbf{w}$) that maximizes the signal-to-noise energy (λ) in the background suppressed scene ($\mathbf{z}$); i.e., we seek an operator $\mathbf{w}$ to maximize

$$\lambda = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{P}_b^{\perp}\mathbf{t}a^2\mathbf{t}^{\mathrm{T}}\mathbf{P}_b^{\perp\mathrm{T}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{P}_b^{\perp}E(\varepsilon\varepsilon^{\mathrm{T}})\mathbf{P}_b^{\perp\mathrm{T}}\mathbf{w}} = \frac{a^2\mathbf{w}^{\mathrm{T}}\mathbf{P}_b^{\perp}\mathbf{t}\mathbf{t}^{\mathrm{T}}\mathbf{P}_b^{\perp\mathrm{T}}\mathbf{w}}{\sigma^2\mathbf{w}^{\mathrm{T}}\mathbf{P}_b^{\perp}\mathbf{P}_b^{\perp\mathrm{T}}\mathbf{w}} \tag{10.23}$$

where E is the expected value and recall that σ^2 is the nominal variance of the iid residual error. Harsanyi and Chang (1994) point out that maximization of Eq. (10.23) is a classic eigenvector problem, which in this case yields the convenient result

$$\mathbf{w}^{\mathrm{T}} = c\mathbf{t}^{\mathrm{T}} \tag{10.24}$$

where c is an arbitrary scalar. The operator $\mathbf{w}^T$ acts like a projection of $\mathbf{z}$ onto the target vector $\mathbf{t}$. Thus, the final overall OSP operator is a $(1 \times \ell)$ vector $\mathbf{t}^{\mathrm{T}}\mathbf{P}_b^{\perp}$ that operates on the original spectral vectors according to

$$T_{OSP}(\mathbf{x}) = \mathbf{t}^{\mathrm{T}}\mathbf{P}_b^{\perp}\mathbf{x} \tag{10.25}$$

where T_{OSP} is the scaler magnitude of the OSP target detection algorithm that can be thresholded to form a conventional binary detection solution (i.e., large values are likely targets). Manolakis et al. (2001) point out that by normalizing the OSP operator by $\mathbf{t}^T\mathbf{P}_b^{\perp}\mathbf{t}$, it yields a product equal to the estimate of the fractional abundance of the target in the pixel according to

$$\bar{T}_{OSP}(\mathbf{x}) = a = \frac{\mathbf{t}^{\mathrm{T}}\mathbf{P}_b^{\perp}\mathbf{x}}{\mathbf{t}^{\mathrm{T}}\mathbf{P}_b^{\perp}\boldsymbol{t}} \tag{10.26}$$

However, they also point out that neither the traditional OSP operator [e.g., Eq.(10.25)] nor the normalized OSP operator [e.g., Eq.(10.26)] has the desirable property of providing a *constant false-alarm rate (CFAR)*. CFAR algorithms are attractive because we can obtain predictable performance by controlling the thresholds. A CFAR detector adapts to the local environment and provides the same expected error levels across varying conditions in the scene. The importance of this can be seen in Figure 10.10, which shows a standard *receiver operator characteristic* (*ROC*) curve. This type of curve can be generated by using a binary threshold on any detection algorithm and plotting the percentage of detections versus the percentage of false alarms as the threshold is varied. Clearly, as the threshold is lowered, the probability of finding targets increases, but so does the probability of false alarms. We often set the threshold in a detection algorithm at some acceptable level for false alarms (e.g., we can afford to investigate one in every 100,000 pixels). However, if our algorithm is not CFAR, it may produce many more false alarms in one region than in another. Thus, we often seek CFAR algorithms where the operator or the threshold value adapts to achieve approximately constant false alarms as the scene characteristics vary. Manolakis et al. (2001) suggest that for the structured or deterministic background case, the *adaptive subspace detector* (*ASD*) (also known as the *F* test) provides such CFAR performance according to:

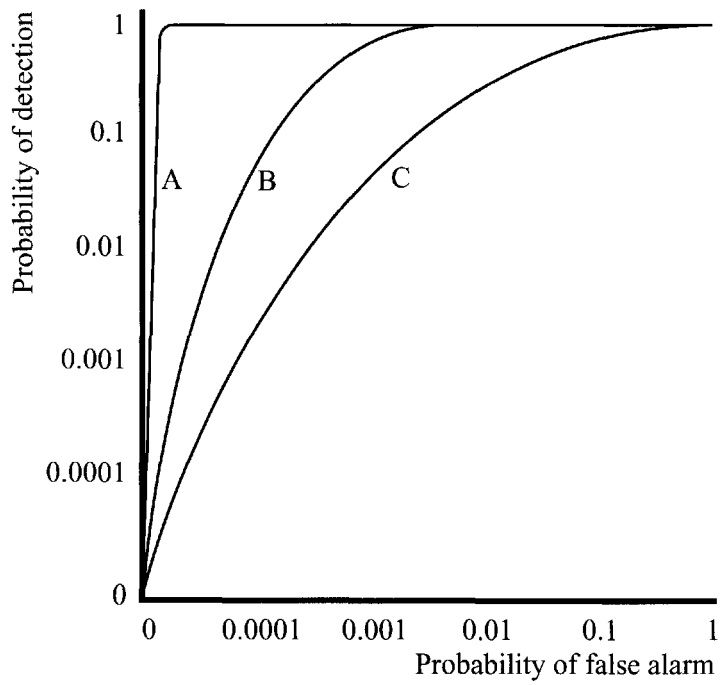


Figure 10.10 Receiver operator characteristic (ROC) curve showing how probability of detection varies with probability of false alarm as a detector threshold is varied. The three curves show the results for three different target classes ranging from high contrast (A) to low contrast (C).

$$T_{ASD}(\mathbf{x}) = \frac{\mathbf{x}^T\left(\mathbf{P}_b^{\perp} - \mathbf{P}_H^{\perp}\right)\mathbf{x}}{\mathbf{x}^T\mathbf{P}_H^{\perp}\mathbf{x}}(\ell - 1 - k) \begin{array}{c} \text{target} \\ \gtrless \eta_o \\ \text{no target} \end{array} \tag{10.27}$$

where $\mathbf{H}$ is the matrix formed by concatenating $\mathbf{B}$ and $\mathbf{t}$, $\mathbf{P}_H^{\perp}=(\mathbf{I}-\mathbf{H}\mathbf{H}^{\#})$, and η_o is the detection threshold that can be determined because $T_{\text{ASD}}(\mathbf{x})$ is F distributed with $(1, \ell - 1-k)$ degrees of freedom in the numerator and denominator, respectively. One can also experimentally determine a threshold by running the algorithm on a large region known to have no targets and adjusting η_o to achieve the desired false-alarm rate.

10.3.3 Linear Mixture Models and Fraction Maps

In the previous section, we emphasized target detection or classification algorithms that employ the geometric or structured perspective. This perspective also gives rise to the concept of spectral unmixing and the generation of *fraction maps* as described by Smith et al. (1990a,b) and Roberts et al. (1993). The linear mixing model assumes that the observed spectral vector is a linear combination of the spectral vectors that would be observed from pure samples of the end members present in the pixel. This results in the same expression used in our discussion of the OSP operator [Eq. (10.20)] for the case where no feature vector is specifically identified as a target, i.e.,

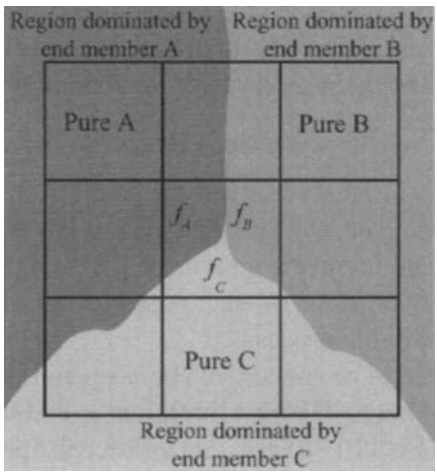
$$\mathbf{x} = \mathbf{B}\boldsymbol{\alpha} + \varepsilon \tag{10.28}$$

where, in the ideal case, the weights in the $\boldsymbol{\alpha}$ vector are the area-weighted subpixel fractions of the pure end members contained in the pixel (cf. Fig. 10.11). We can expand Eq. (10.28) to write ℓ equivalent expressions (i.e., one for each band) of the form

$$\begin{aligned} x_1 &= v_{11}\mathrm{f}_1 + v_{12}\mathrm{f}_2 \cdots + v_{1k}\mathrm{f}_\mathrm{k} + \varepsilon_1 \\ x_2 &= v_{21}\mathrm{f}_1 + v_{22}\mathrm{f}_2 \cdots + v_{2k}\mathrm{f}_\mathrm{k} + \varepsilon_2 \\ &\vdots \\ x_\ell &= v_{\ell 1}\mathrm{f}_1 + v_{\ell 2}\mathrm{f}_2 \cdots + v_{\ell k}\mathrm{f}_\mathrm{k} + \varepsilon_\ell \end{aligned} \tag{10.29}$$

If we know the form of the end members $(\mathbf{v}_1,\mathbf{v}_2\ldots\mathbf{v}_k)$, then these ℓ linear equations can be solved for the k unknown fractions $(\mathrm{f}_1,\mathrm{f}_2\ldots\mathrm{f}_k)$ if k or more of the equations are independent. Since ℓ is typically on the order of tens or hundreds and k is typically much less then 10, this problem is overdetermined and can be solved by conventional least-squares methods (note that for the few-band multispectral case, this problem may not have a unique solution). Thus, we can unmix each spectral vector into the end-member fractions contained in that pixel according to

$$\hat{\boldsymbol{\alpha}} = \boldsymbol{B}^{\#}\boldsymbol{x} = E < \boldsymbol{B}^{\#}\boldsymbol{B}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} > \simeq \boldsymbol{I}\boldsymbol{\alpha} = \boldsymbol{\alpha} \tag{10.30}$$



$L_A = mr_A + L_u = [E_s\pi^{-1}\tau_2 + L_d\tau_2]r_A + L_u$, radiance for end member A

$DC_A = gL_A + b$, digital count for end member A

$L_B = mr_B + L_u$, radiance for end member B

$DC_B = gL_B + b$, digital count for end member B

$L_C = mr_C + L_u$, radiance for end member C

$DC_C = gL_C + b$, digital count for end member C

$r_m = f_A r_A + f_B r_B + f_C r_C$, linear mixing in reflectance

$$L_M = mr_m + L_u = mf_A r_A + mf_B r_B + mf_C r_C + [f_A + f_B + f_C]L_u$$

since $f_A + f_B + f_C = 1$

$$= f_A L_A + f_B L_B + f_C L_C, \text{ linear mixing in radiance}$$

$$DC_m = gL_m + b = gf_A L_A + gf_B L_B + gf_C L_C + [f_A + f_B + f_C]b$$

$$= f_A DC_A + f_B DC_B + f_C DC_C, \text{ linear mixing in digital count}$$

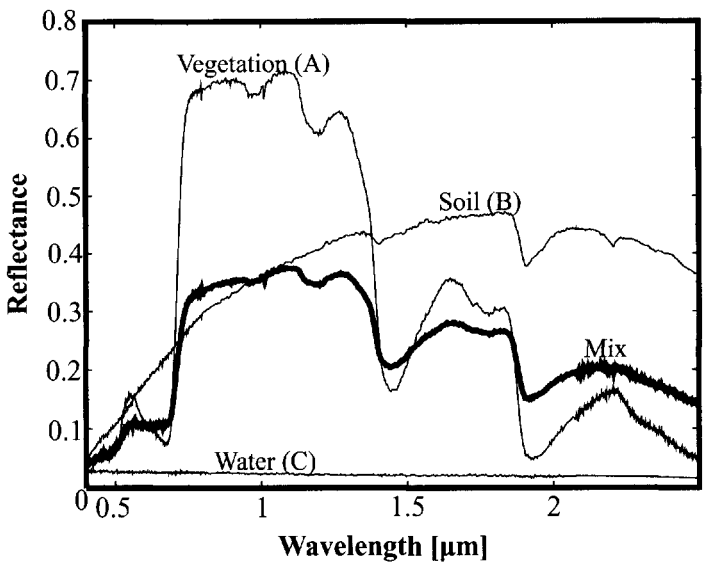


Figure 10.11 Illustration of linear mixing model concept. The mixed center pixel has fractional areas f_A, f_B, and f_C made up of the end members A, B, and C respectively. The derivation demonstrates that linear mixing models can be applied in reflectance, radiance, digital count, or any linear-transformed space.

where the pseudo-inverse operator ($\mathbf{B}^{\#}$) yields a best fit (in a least-squared error sense) solution for the vector of end-member fractions ($\hat{\alpha}$), E represents the expected value operator, and the residual error vector can be computed as

$$\boldsymbol{\varepsilon} = \mathbf{x} - \mathbf{B}\hat{\boldsymbol{\alpha}} \tag{10.31}$$

Operationally, Eq. (10.30) is solved for each pixel, and k fraction planes or fraction images are produced whose brightness is proportional to the fractional value of each of the k end members computed for each pixel. Thus, the output is the subpixel fractional abundance of each possible end member, assuming all end members were included in the process.

In practice, we need to be concerned about several potential limitations of the simple model presented here. The first limitation is that the unconstrained unmixing process captured in Eq. (10.30) does not enforce the physical constraint that the fractional contributions of the end members sum to one, i.e.,

$$\sum_{i=1}^{k} f_i = 1 \tag{10.32}$$

Furthermore, neither Eq. (10.30) nor Eq. (10.32) will constrain the fractional abundances to be fractions, i.e.,

$$0 \le f_i \le 1 \tag{10.33}$$

Draper and Smith (1981) discuss how both the equality and inequality constraints can be included in the least-squares analysis to yield a partially constrained solution for the fractional abundance (i.e., incorporating the equality constraint) or a fully constrained solution (i.e., incorporating both equality and inequality constraints). However, as shown in Figure 10.12, fractions slightly outside the 0 to 1 range may be reasonable, and rather than force alternative solutions, a partially constrained solution is often used. The partially constrained results can be filtered to show values significantly outside the expected range as poor fits to the model. These high error pixels are often pixels that contain materials not in the end-member model. Note that for these pixels, the fully constrained case would generate a mathematically valid, but unreasonable solution.

Roberts et al. (1993) point out the value of using one end member to account for brightness variation within a scene. They suggest using a *shade* end member composed of the darkest pixels within a scene (e.g., a zero reflector in reflectance space or the upwelled radiance vector in a radiance image). Varying the fractional amount of the shade end member will tend to account for brightness variations within a scene due to both shadowing and illumination [e.g., ($\cos\sigma$)] effects. If a partially or fully constrained case is used, the shade fractions can be reapportioned to the other fractions after unmixing, if desired. In the example in Figure 10.12, the A end member would be shade and point F would be unmixed as (1/3, 1/3, 1/3) and then reapportioned between B and C as (1/2, 1/2) for mapping purposes (one can imagine more exotic cases where we might choose to map all or most of the shadow to a single class [e.g., soil or grass)].

One of the most difficult problems with conventional unmixing is the classic least-square regression problem of overfitting. For any unmixing model, we can

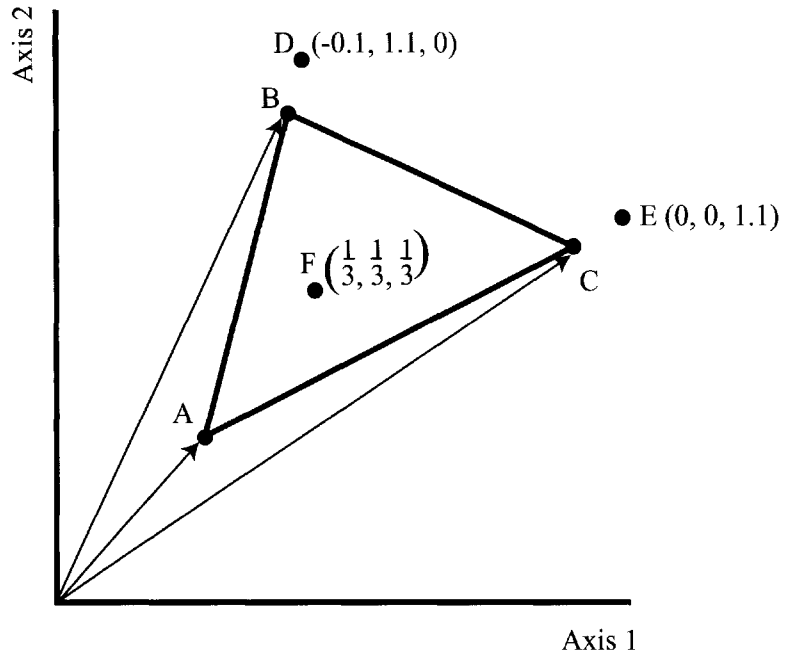


Figure 10.12 Illustration of cases where fractions slightly outside the zero to one range might be tolerable. Point D illustrates a partially constrained case, and E an unconstrained case. Note: fractions are designated as (f_A, f_B, f_C).

reduce the residual error by adding an additional end member. However, at some point we start to fit to the noise, not the substance. Thus, Roberts et al. (1993) suggest using the smallest number of end members in the model that drives the overall residual error below some tolerable threshold (e.g., one might attempt to achieve some small multiple of the instrument noise). The overall error can be expressed as

$$\bar{\varepsilon} = \left[\sum_{p}^{P} \left(\sum_{m}^{\ell} \left| \sum_{n}^{k} v_{mn} f_{pn} - x_{pm} \right| \right) / \ell \right] / P \tag{10.34}$$

where p, m, and n denote pixel, band, and end member, respectively, and P, ℓ, and k are the total number of pixels, bands, and end members, respectively. Once a satisfactory model is achieved as denoted by a small number of end members, low residual errors, and spatially plausible fraction maps, an error image and individual error vectors [cf. Eq. (10.31)] can be analyzed for anomalies. This anomaly analysis can be particularly productive if we are interested in low probability of occurrence materials (i.e., unsuitable as end members in our model) that have distinct spectral features that should appear as residual spectral structure.

Clearly this approach of keeping the number of end members small can make detailed mapping difficult. This can be dealt with by using either a spatial or spectral hierarchy, or both. With a spatial hierarchy, we might presegment the image and design a different mixing model for different spatial regions. With the spectral

hierarchy one could unmix with a simple model, for example, a class that included exposed soils. Then for those pixels with any significant fraction of exposed soil, a more soil-specific model could be employed.

To overcome the limitations of the fixed model approach to unmixing, enhancements have been suggested that involve development of pixel specific models [cf. Gross and Schott (1996) and Roberts et al. (1998)]. We will consider the per pixel unmixing approach of Gross and Schott (1998) since it forms the basis for some subpixel unmixing discussed in Chapter 12. This approach assumes the linear mixing model can change on a per pixel basis, in terms of both how many end members and which end members. Thus, it involves a comparison between alternative models and a search to find the best model that is based on *stepwise regression* theory [cf. Draper and Smith (1981)]. The stepwise approach is an iterative method that first solves for the most likely single end member. It then checks for all two-end member combinations (containing the first) to determine if a "better" model exists. If so, it checks to see if a three-end-member model containing the first two is "better." If a third member is added, it checks to see if the model would be "better" with only two of the current three. The process continues until no "better" model can be found by adding or removing end members. At this point, the end members included in the pixel are assumed known, and any of the conventional linear unmixing algorithms (unconstrained, partially constrained, or fully constrained) can be employed to unmix the pixel. This approach is considerably more compute intensive than a fixed unmixing model; however, it can yield considerably improved results (cf. Fig. 10.13).

To use the stepwise method, we need a library of candidate end members in the same units (e.g., reflectance, radiance, or raw counts) as the image to be analyzed. The end members may be from a reflectance library or scene-derived end members (cf. Sec. 10.3.1). In general, this should be a relatively small set of independent end members to optimize run time and performance. The set should be large enough to include all relevant end members, yet small enough to allow reasonable run times.

The metric used to control the stepwise process uses the errors associated with the regression which can be expressed as

$$\sigma_{tot} = \mathbf{x}^{\mathrm{T}}\mathbf{x} = SSR + SSE = \hat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{x} + [\mathbf{x}^{\mathrm{T}}\mathbf{x} - \hat{\boldsymbol{\alpha}}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{x}] \qquad (10.35)$$

where σ_{tot} is the total variation in the pixel (about the origin), SSR is the sum squared variation explained by the regression, and the difference is the sum squared error SSE. We are using the same parameters to describe the mixing model as introduced in Eq. (10.28), where $\mathbf{B}$ is the $\ell \times k$ matrix of end members and $\hat{\alpha}$ is the column vector made up of the k mixing fractions. We seek a model that explains as much of the variation as possible (maximize SSR) without overfitting to the measurement noise. The solution is based on an analysis of variance, which assumes that if the model is good, the errors will be Gaussian with zero mean, the SSR and SSE are chi squared distributed, and their mean square ratio is $F_{k,\ell-k}$ distributed (i.e., k degrees of freedom in numerator, $\ell - k$ degrees of freedom in the denominator) according to:

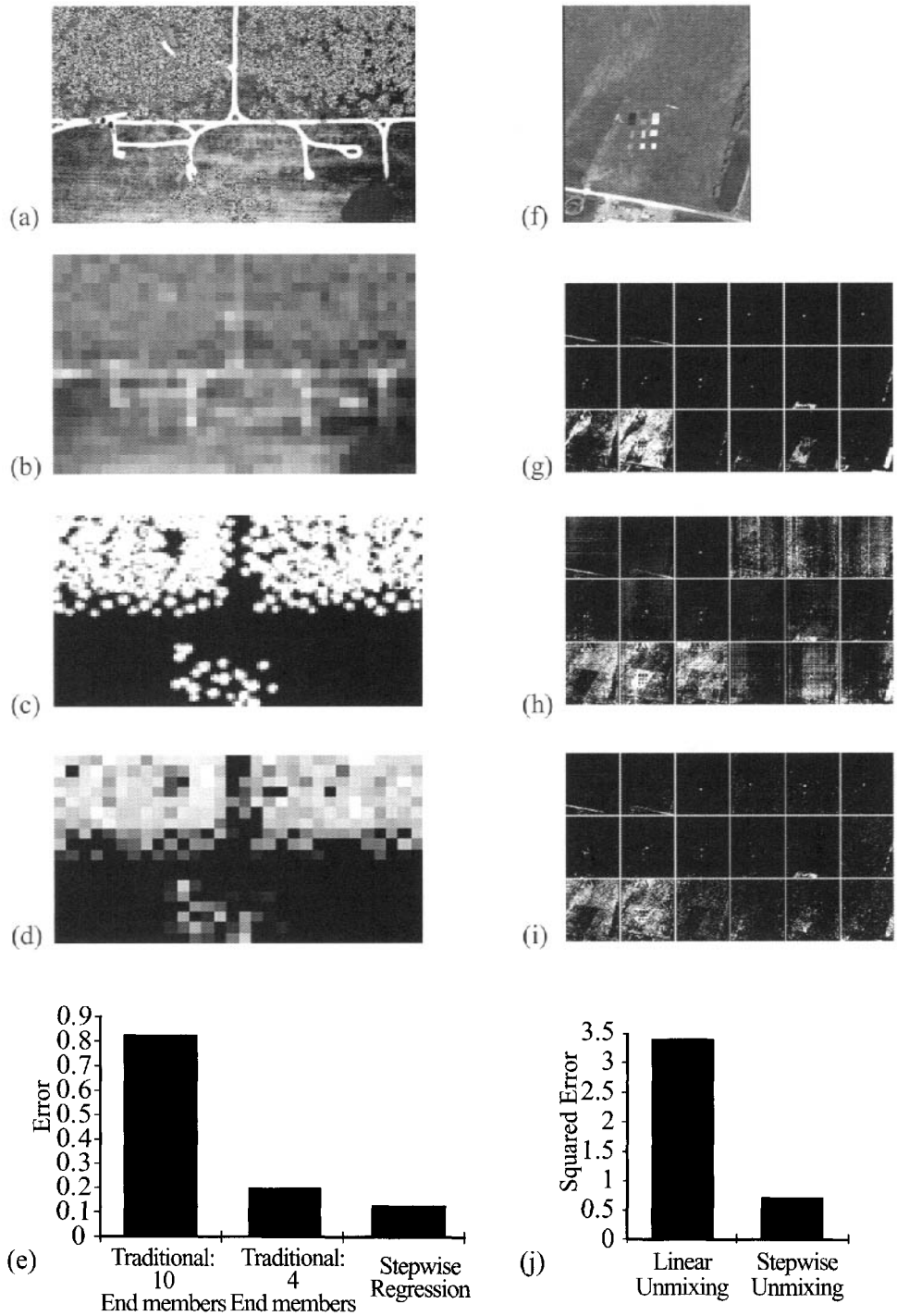


Figure 10.13 Results of linear unmixing. A 15-band synthetic image (a) was spatially degraded (b) and unmixed (d) for comparison against truth fractions (c). Sample results are shown in (e) for three cases: the first is a fixed model with 10 end members, the second is a fixed model with the four most populous end members, and the third is a stepwise regression with all the end members in the available library. A 210-band Hydice image is shown in (f) with truth maps for 18 classes in (g) and mixing maps for conventional (h) and stepwise (i). Final performance is shown in (j).

$$F_{k,\ell-k} = \frac{MSR}{MSE} = \left(\frac{SSR}{k}\right) / \left(\frac{SSE}{\ell - k}\right) \tag{10.36}$$

where MSR is the mean square variation explained by the regression and MSE is the mean square error (residual error).

In general, we want to compare the difference between two models, one of which has k end members and the other has $k - 1$ end members. The models can be expressed as

$$\mathbf{x} = \mathbf{B}_{k-1}\boldsymbol{\alpha}_{k-1} + \boldsymbol{\varepsilon}_{k-1} \tag{10.37}$$

$$\mathbf{x} = \mathbf{B}_{k}\boldsymbol{\alpha}_{k} + \boldsymbol{\varepsilon}_{k} \tag{10.38}$$

where the subscript indicates the number of end members. If we are looking to add a new end member (forward stepwise regression), k represents the new model and $k - 1$ the previous one. If we are looking to remove an end member (backward stepwise regression), k represents the previous solution and $k - 1$ a potential new one. The decision rule used to compare the augmented (k) model to the $k - 1$ model uses the mean square value for the extra term expressed as

$$MS_{extra} = \frac{\hat{\boldsymbol{\alpha}}_k^{\mathrm{T}}\mathbf{B}_k^{\mathrm{T}}\mathbf{x} - \hat{\boldsymbol{\alpha}}_{k-1}^{\mathrm{T}}\mathbf{B}_{k-1}^{\mathrm{T}}\mathbf{x}}{k-(k-1)} \tag{10.39}$$

This represents the difference in the amount of variation explained by the model with k terms and $k - 1$ terms. The ratio of MS_{extra} to MSE is F-distributed with one $[k - (k - 1) = 1]$ degrees of freedom in the numerator and $\ell - k$ in the denominator. We can test to see if the model improvement is significant by checking if

$$F_{1,\ell-k} = \frac{MS_{extra}}{MSE} \tag{10.40}$$

exceeds the tabulated value for the corresponding F-test. If the improvement is significant, the new end member should be included. If, during the backward stepwise test, the smallest F value is below the threshold, then that end member is removed from the model since its inclusion no longer significantly increases the amount of variation explained by the model. Note that by using a constant F value, cycling of end members in and out is precluded. The process continues until no end members are added or removed. The final fractions are then computed using conventional unmixing (i.e., partial or full constraints may be applied). In practice, the F table is not used. Instead, a fixed F value is selected regardless of degrees of freedom. At each step in the analysis, the F statistic is calculated using each end member in the library to augment the model. If the largest F value exceeds the threshold, the model is augmented. Then each end member is sequentially removed and the appropriate F statistic computed. Note that Gross and Schott (1996) suggest that a check can be employed before a model is accepted. If fractions much greater than one or large negative values occur, a model may be rejected. The end member with the largest fraction may then be temporarily removed from the model and the process restarted for that pixel to force a dramatically different solution. An alter-

native per pixel unmixing approach described by Roberts et al. (1998) uses the root mean square residual error and residual vectors to ensure a robust solution.

Gross and Schott (1996) used a squared error metric averaged by pixel to report the performance of various unmixing methods. This can be expressed as

$$E = \sum_{p=1}^{P}\left[\sum_{n=1}^{k}\left(\mathrm{f}_{tpn} - \mathrm{f}_{pn}\right)^2\right] / P \tag{10.41}$$

where f_{tpn} is the true fractional amount of the nth end member in the pth pixel and f_{pn} is the corresponding value estimated by unmixing. Note that this error metric penalizes both errors of omission and of commission equally for all possible end members. Their results on synthetic imagery and subsequent studies on real imagery [cf. Konno (1999) and Newland (1999)] showed that the stepwise method consistently outperformed fixed unmixing models against this metric. These studies also pointed out that partially constrained unmixing generally out-performed unconstrained unmixing, but that fully constrained unmixing, while often performing slightly better against the squared error metric, was probably not worth the increased computational cost. Figure 10.13 shows examples of unmixing results in the form of fraction planes and the squared error metric. Note that fractional error assessment is very difficult to perform because of the difficulty in defining truth, which is why many of the results in the literature use synthetic data or report only anecdotal findings.

10.4 STATISTICAL APPROACHES TO SPECTRAL IMAGE ANALYSIS

In this section we explore a number of spectral image analysis algorithms that are based on a statistical representation of the image data (cf. Sec 10.4.2). In anticipation of this, we must first define some methods for gathering the necessary statistical descriptions of images, subimages, and classes (cf. Sec. 10.4.1).

10.4.1 Estimation of Relevant Statistical Parameters

Many of the algorithms of interest will require us to estimate the mean and spectral covariance of a data set. This can be accomplished using the formalism introduced in Section 9.2.2.3 for Gaussian maximum likelihood classification. However, it is also often convenient to express the covariance matrix as the average of the outer products of the demeaned data set. Computationally, we first compute the mean spectral vector ($\mathbf{m}$) for a data set and subtract it from each spectral vector ($\mathbf{x}$) in the data set, yielding

$$\mathbf{z}_j = \mathbf{x}_j - \mathbf{m} \tag{10.42}$$

The average of the outer products of the demeaned data ($\mathbf{z}$) values then yields the covariance according to

$$\mathbf{S} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{z}_j\mathbf{z}_j^{\mathrm{T}} = \frac{1}{N}\sum_{j=1}^{N}\left(\mathbf{x}_j - \mathbf{m}\right)\left(\mathbf{x}_j - \mathbf{m}\right)^{\mathrm{T}} \tag{10.43}$$

where N is the number of pixels in the data set and $\mathbf{S}$ is computationally the same covariance as introduced in Eq. (9.70) except that in that case the data set consisted of the ith class.

Clearly, we can compute sample means and covariances for samples ranging from whole images or window regions around a pixel of interest to spectral classes within an image or region of interest. In the case where class statistics are needed, we must first somehow segment the scene into classes. This could be accomplished using the supervised or unsupervised multispectral methods discussed in Section 9.2 or some of the structural methods discussed in Section 10.3.2. However, an interesting alternative approach described by Masson and Pieczynski (1993) uses the scenewide statistics to drive the automated computation of class statistics under the assumption of Gaussian behavior for the spectral classes. The *stochastic expectation maximization* (*SEM*) algorithm attempts to describe the overall image spectral probability distribution function as a sum of multivariate Gaussian distribution for some number of classes according to

$$p(\mathbf{x}) = \sum_{i=1}^{k} p(i)p(\mathbf{x}|i) = \sum_{i=1}^{k} \frac{N_i}{N} \frac{HIS_i(\mathbf{x})}{N_i} = \frac{HIS(\mathbf{x})}{N} \tag{10.44}$$

where $p(\mathbf{x})$ is the probability of the spectral vector $\mathbf{x}$ occurring in the data set that is equal to the value of the normalized histogram [HIS($\mathbf{x}$)/N], k is the number of classes in the image, each of which is designated by the variable i, $p(i)$ is the a priori probability of the ith class occurring, which can be computed from N_i/N if the number of pixels in the ith class (N_i) is known, $p(x|i)$ is the conditional probability of $\mathbf{x}$ occurring given that we are sampling from the ith class [which is equal to the normalized class histogram, HIS$_i$($\mathbf{x}$)/N_i], and N is the total number of pixels in the data set. The conditional probabilities are treated as multivariate Gaussian distributions that can be expressed as [cf. Eq. (9.69)]:

$$p(\mathbf{x}|i) = (2\pi)^{-\frac{\ell}{2}} |\mathbf{S}_i|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \mathbf{S}_i^{-1}(\mathbf{x}-\mathbf{m}_i)} \tag{10.45}$$

There is a persistent question regarding how well class statistics match a Gaussian model. However, Jimenez and Landgrebe (1998) point out in a discussion of the structure of hyperspectral data sets that if hyperspectral data are linearly projected into lower dimensional spaces, the resulting data can be characterized by normal distributions or sums of normal distributions. A simple one-dimensional projection of the SEM concept is shown in Figure 10.14. With the explicit multivariate Gaussian assumption, Eq. (10.44) becomes

$$p(\mathbf{x}) = \sum_{i=1}^{k} p(i)(2\pi)^{-\frac{\ell}{2}} |\mathbf{S}_i|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T \mathbf{S}_i^{-1}(\mathbf{x}-\mathbf{m}_i)} \tag{10.46}$$

The SEM algorithm attempts to solve for the number of classes or clusters (k), the a priori probabilities of each class, and the mean and spectral covariance of each class. This is done using an iterative segmentation process on a reduced dimension data set (cf. Sec. 10.2) to expedite processing. The only user inputs are the initial estimate of the number of classes (note that this should be an upper limit

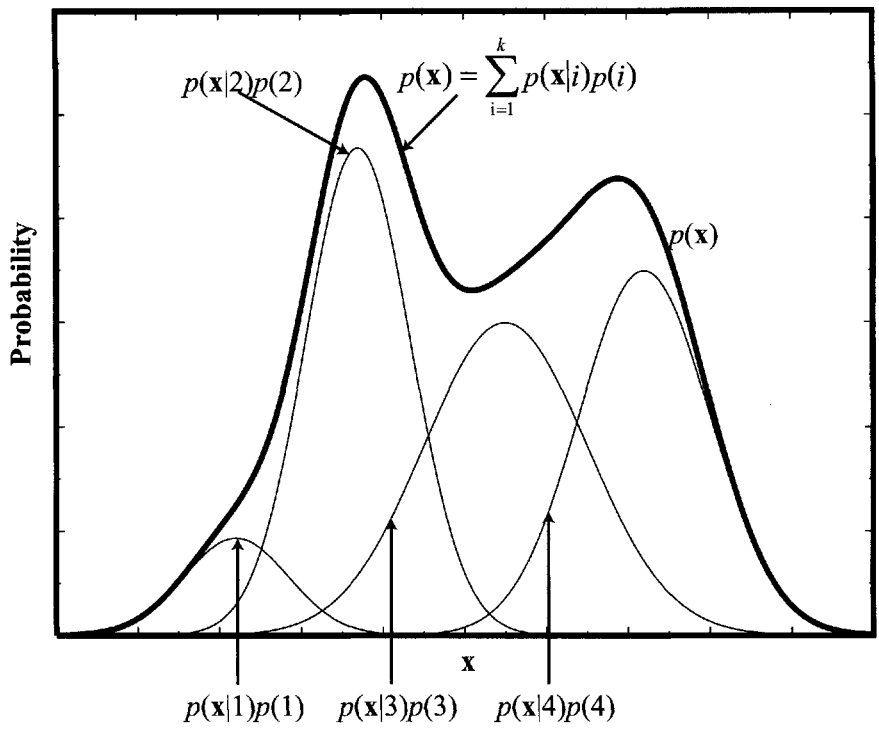


Figure 10.14 One-dimensional projection to illustrate how Gaussian distributions sum to produce overall scene probability distribution function.

since the algorithm can only reduce it) and a threshold (δ) on the a priori probability a class must have to remain relevant [i.e., if $p(i)$ falls below this value, reduce the number of classes by one]. Note that because the SEM algorithm is driven by scene wide statistics, it is designed to identify clutter or background classes with significant abundance and will not be effective in characterizing classes with a low probability of occurrence. The algorithm proceeds as follows:

Initialization: $p(i)=1/k$ and randomly assign each pixel to a class.

Compute class statistics: Based on class assignments from the previous step, compute the prior probabilities $[p^s(i)]$, the class means m_i^s, and the class covariance $\mathbf{S}_i^s$ where the superscript s indicates the iteration level.

Remove small classes: If $p(i) < \delta$, reduce k by 1 and reinitialize. Note that since we start with equal a priori probabilities and drift away from this initial assumption, the algorithm ensures that classes will not be removed too early in the process.

Assign pixels to classes: Each pixel is assigned to the class with the highest a posteriori probability based on the most recently computed class statistics according to

$$p^{s+1}(i|\mathbf{x}) = \frac{p^s(i)p^s(\mathbf{x}|i)}{p(\mathbf{x})} = \frac{\frac{N_i^s}{N}(2\pi)^{-\frac{\ell}{2}}\left|\mathbf{S}_i^s\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-m_i^s)^T\mathbf{S}_i^{s-1}(\mathbf{x}-\mathbf{m}_i^s)}}{\frac{HIS(\mathbf{x})}{N}} \tag{10.47}$$

where we note that $p(\mathbf{x})$ is a constant for any pixel and can be removed from the computation as it will not influence class assignment.

Return to compute class statistics step or stop: There are a number of ways to decide on convergence of this algorithm. One is to stop when the number of pixels changing classes is below some threshold; another is when all the class means change by less than some amount, or one could use the much more computationally intensive approach of computing the error in how well the modeled probability matches the observed histogram as expressed in Eq. (10.44).

The output from the SEM process includes the number of spectral clusters, as well as the mean and covariance of each spectral cluster. In addition, we can obviously use the SEM final a posteriori classification to generate an unsupervised classification of the image by spectral cluster (i.e., this is an unsupervised GML classifier). It is important to note, however, that this process, like the k means process introduced in Chapter 9, does not necessarily group the data into intuitively meaningful clusters. Also, class names, if relevant (possible), must be assigned either by manual interpretation of the class map or by some form of spectral matching of library material signatures against the cluster means.

Beaven et al. (1999) point out that there are substantial similarities between scene segmentation based on applying GML classifiers to SEM data and geometric classifiers (e.g., mixing models) to the same data sets, despite the fundamental differences in the perspective from which the algorithms are derived. For our purposes, we will focus on the use of the cluster statistics in some of the target detection algorithms in the next section.

10.4.2 Target Detection Using Statistical Characterization of the Image

In this section we will look at *target detection* algorithms that are based on the statistical perspective (cf. Sec. 10.1.2). In many cases, these algorithms are based on concepts similar to those introduced in Section 9.2 in the treatment of Gaussian maximum likelihood classifiers and draw on the mathematical tools introduced in Section 9.1.

10.4.2.1 Unknown Target/Anomaly Detection

We will begin with the special case where the target's spectral character is unknown. We assume that the target has an additive spectral feature that distinguishes it from the mean of the background. Our algorithm must then find pixels that

are anomalous from the background, and we must typically use some additional information [e.g., context or shape (cf. Sec. 10.4.2.3)] to decide if the anomalies are targets or not. Anomaly detectors are not limited to the stochastic perspective. For example, if we were using any of the geometric approaches presented in Section 10.3, we could use a threshold to call any pixel that was very distinct from our chosen features an anomaly. Using spectral angle, for instance, a pixel more than some threshold angle from any of our known end members could be flagged as an anomaly. The problem with many ad hoc anomaly detectors is that the number of false alarms is very unpredictable. Reed and Yu (1990), taking the stochastic approach, derived a CFAR detector (RX) for anomaly detection. Based on the assumption that anomalous targets can be treated as having an additive spectral signature relative to a varying background mean with normally distributed background spectral clutter, they derived what has become known as the RX algorithm. In its simplest form, it can be expressed as

$$R(\mathbf{x}) = (\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}-\mathbf{m}) \begin{matrix} \geq \eta \Rightarrow \text{target} \\ < \eta \Rightarrow \text{no target} \end{matrix} \tag{10.48}$$

where $\mathbf{m}$ is the local mean, $\mathbf{S}$ is the local spectral covariance matrix, and η is the threshold set by the user to control the false alarm rate. While the RX algorithm can be applied using whole image statistics, it is designed to be used in an adaptive fashion where the mean and covariance matrix are computed on a local window about the pixel under test. Refinements include excluding a region about the pixel being tested from the calculation of the background statistics to avoid potential contamination of the background statistics with the target, and incorporation of the target size/shape (cf. Sec. 10.4.2.3). Considering the RX algorithm in the context of the GML discussions in Section 9.2, we recognize it as the squared *Mahalanobis* or statistical distance of the pixel from the local mean. In the context of the target detection theory introduced in the next section, we can also think of the RX algorithm as a matched filter of the demeaned pixel with itself normalized by the local statistical variability. Figure 10.15 illustrates the RX concept and shows the results applied to a spectral image. Note that, in practice, the algorithm is often applied to the data after dimensionality reduction (e.g., an MNF transform) to reduce noise and processing time and to increase the validity of Gaussian estimates of the clutter statistics. In some cases, the processing speed can be further increased with little loss in performance if we can assume that while the local mean may vary, the covariance is slowly varying and can be treated as approximately constant over a region. In this case, we can demean locally (e.g., on a window region about the pixel of interest) but use a common covariance computed over a much larger window or even an entire image if the scene content is relatively constant.

One common use of anomaly detection algorithms is to find candidate targets that can be analyzed by other tools (including possibly photo interpretation) such as target detection algorithms where the target spectral signature must be known.

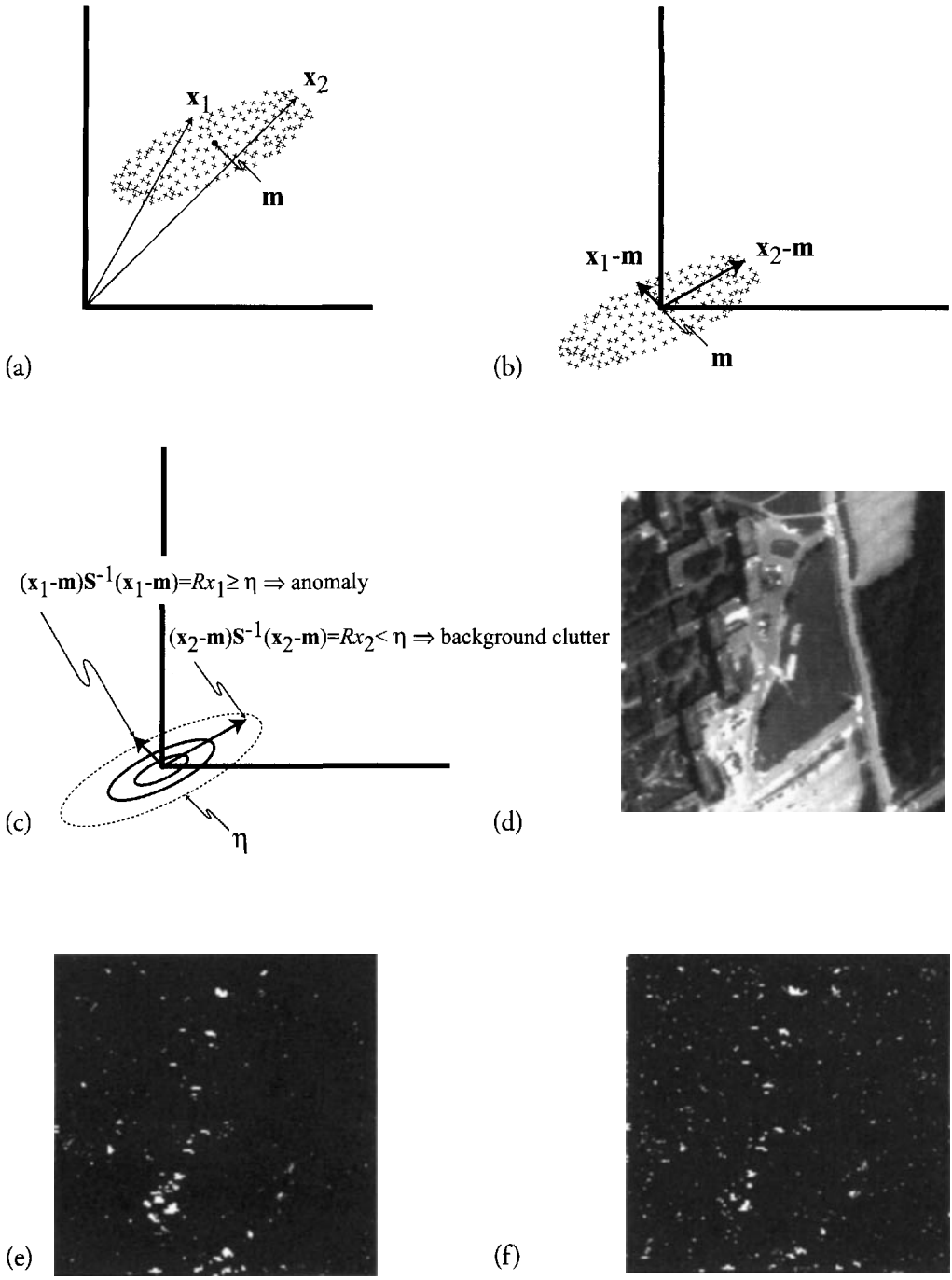


Figure 10.15 Illustration of the RX anomaly detection algorithm. (a) Simplified 2-D plot of the data space highlighting two pixels. (b) The data from (a) after mean subtraction. (c) The isoprobability contours associated with the normalization by the covariance showing that in statistical distance x_1 is outside the anomaly threshold and x_2 is inside. (d) A single band from a 72-band MISI image that was processed using an image-wide RX algorithm and (e) a spatially adaptive RX algorithm (f). The results show that the spatially adaptive algorithm is more sensitive to local anomalies.

10.4.2.2 Target Detection When the Target Is Known

In many cases, we may know what a target looks like in a spectral space (e.g., reflectance, radiance or digital counts). If we can then transform the image and the target spectrum into a common space, a variety of stochastic target detection algorithms can be applied. The radiometric transformations to common spaces have been extensively covered in earlier sections, and will not be repeated. However, recall that residual artifacts of any transformation process can appear to the algorithms as spectral signatures and these artifacts must be avoided or suppressed (cf. Sec. 7.6.2). The target signature may be scene derived [e.g., using photo interpretation, end member selection (cf. Sec. 10.3.1 and 10.4.1), or anomaly detection (cf. Sec. 10.4.2.1)], selected from a spectral library of reflectance spectra, or produced by a phenomenological model (cf. Chaps. 11 and 14). In the discussions that follow, we will assume that an example target spectrum is known and that the image and the target spectrum have been transformed into a common space.

The most common target detection algorithm based on a stochastic description of the data is the *spectral matched filter* in the presence of noise (clutter), which is expressed as

$$SMF(\mathbf{x}) = (\mathbf{t}-\mathbf{m})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}-\mathbf{m}) \quad \begin{matrix} \geq \eta \Rightarrow \text{target} \\ < \eta \Rightarrow \text{no target} \end{matrix} \tag{10.49}$$

where $\mathbf{t}$ is the target vector, $\mathbf{x}$ is the sample vector, and $\mathbf{m}$ and $\mathbf{S}$ are the background mean and covariance matrix. We threshold SMF($\mathbf{x}$) at η to control the false alarm rate. Schaum (2001) points out that when first introduced into the radar/signal detection community, the matched filter was designed to optimize the signal-to-noise ratio (SNR) when noise was the dominant source of variability in the signal (i.e., $\mathbf{S}$ would be the noise covariance). In the spectral remote sensing case, the variability is usually dominated by scene variation, not noise, and so we are more appropriately seeking to maximize the signal-to-clutter ratio (SCR). Thus, $\mathbf{m}$ and $\mathbf{S}$ are the mean of the background (clutter) and the covariance matrix of the background, which includes variation due to scene variability and noise.

To intuitively understand the spectral matched filter (SMF), it is often easier to think of Eq. (10.49) as a two-step process. The first step transforms the demeaned target vector and image data into a space that is normalized or whitened by the square root of the background covariance matrix. This is followed by a projection of the whitened image vector onto the whitened target vector. This can be expressed as

$$SMF(\mathbf{x}) = (\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{t}-\mathbf{m}) = \left[\mathbf{S}^{-\frac{1}{2}}(\mathbf{x}-\mathbf{m})\right]^{\mathrm{T}}\mathbf{S}^{-\frac{1}{2}}(\mathbf{t}-\mathbf{m}) \quad \begin{matrix} \geq \eta \Rightarrow \text{target} \\ < \eta \Rightarrow \text{no target} \end{matrix} \tag{10.50}$$

The SMF assesses how target-like a pixel is in a transform space where we have normalized for the expected variability in the background (cf. Fig. 10.16).

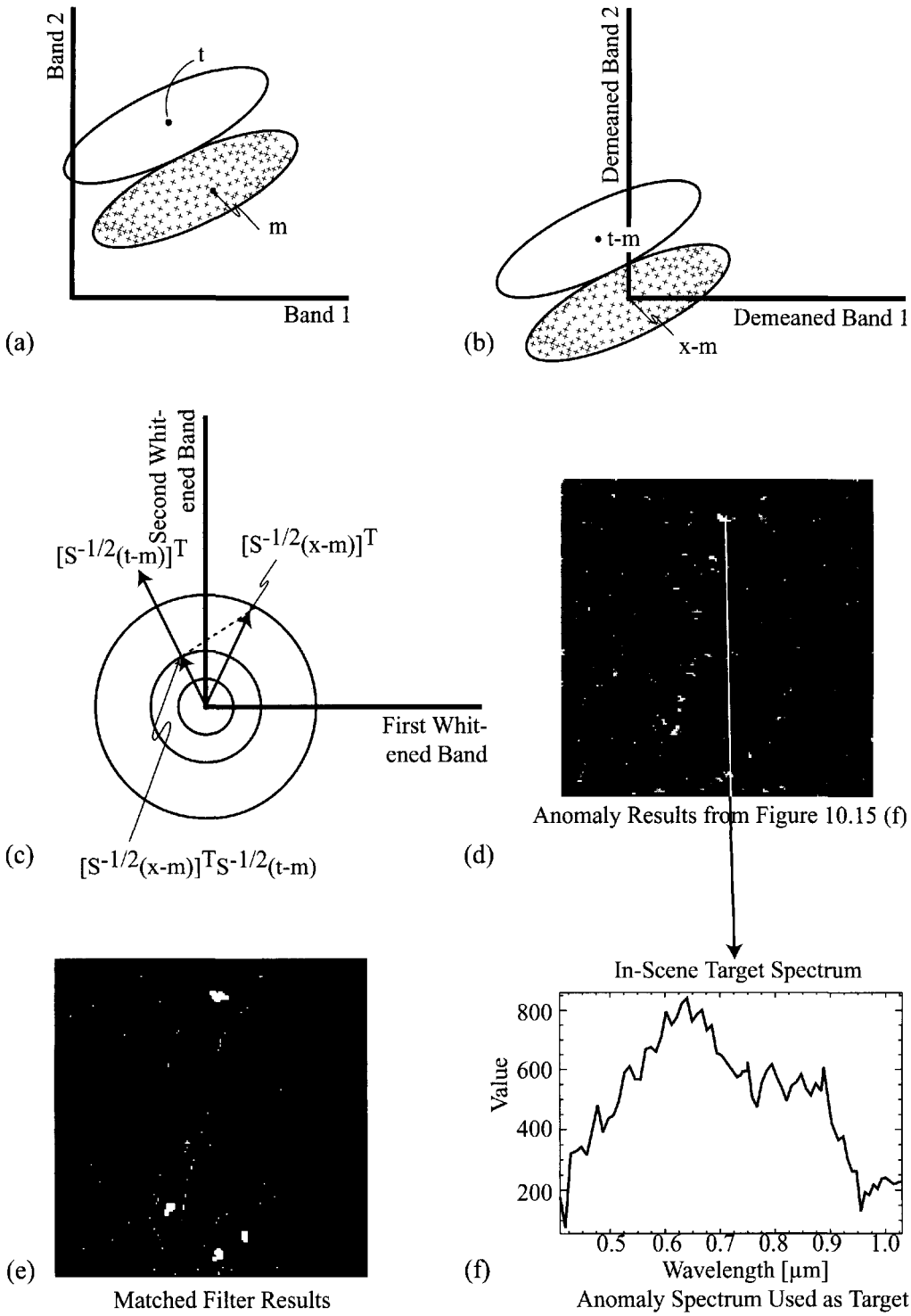


Figure 10.16 The spectral matched filter for a simplified two-dimensional problem is shown conceptually as a two-step process in (a), (b), and (c) (see text). One of the anomalous pixels (f) found using the RX algorithm on the image in Figure 10.15(d) is used to yield the results shown in (e). Note that the target detection process suppresses many non-target-like features that are found by the more general anomaly detector.

We will see this *whitening* concept often in spectral signal processing, so it is worth looking briefly at a simplified two-band case to illustrate the normalization. In the two-band case, we can express the covariance matrix is

$$\mathbf{S} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \tag{10.51}$$

and its inverse is

$$\mathbf{S}^{-1} = \frac{1}{|\mathbf{S}|}\begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{bmatrix} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{bmatrix} \tag{10.52}$$

where σ_{11}, σ_{22} are the variances of bands 1 and 2, respectively, and $\sigma_{12} = \sigma_{21}$ is the covariance of band 1 with band 2. For the simplified case where the bands are uncorrelated (i.e., $\sigma_{12} = \sigma_{21} = 0$), this yields

$$\mathbf{S}^{-\frac{1}{2}} = \frac{1}{\sigma_1\sigma_2}\begin{bmatrix} \sigma_2 & 0 \\ 0 & \sigma_1 \end{bmatrix} \tag{10.53}$$

where σ_1 and σ_2 are the square roots of the variance (i.e., the standard deviations of the variability in the bands). The demeaned whitened vectors ($\mathbf{x}_w$) thus become

$$\mathbf{x}_w = \left[\mathbf{S}^{-\frac{1}{2}}(\mathbf{x}-\mathbf{m})\right] = \frac{1}{\sigma_1\sigma_2}\begin{bmatrix} \sigma_2 & 0 \\ 0 & \sigma_1 \end{bmatrix}\begin{bmatrix} x_1 - m_1 \\ x_2 - m_2 \end{bmatrix} = \begin{bmatrix} \dfrac{x_1 - m_1}{\sigma_1} \\ \dfrac{x_2 - m_2}{\sigma_2} \end{bmatrix} \tag{10.54}$$

where, for our uncorrelated case, we see that the whitening process scales each band by the standard deviation in that band. For correlated bands, the math is messier, but conceptually we are still scaling each sample by its expected variability to yield a whitened result. For this simple case, the final result is just the projection operation expressed as

$$SMF(\mathbf{x}) = \mathbf{x}_w{}^{\mathrm{T}}\mathbf{t}_w = \begin{bmatrix} \dfrac{x_1 - m_1}{\sigma_1} \\ \dfrac{x_2 - m_2}{\sigma_2} \end{bmatrix}^{\mathrm{T}}\begin{bmatrix} \dfrac{t_1 - m_1}{\sigma_1} \\ \dfrac{t_2 - m_2}{\sigma_2} \end{bmatrix} \quad \begin{matrix} \geq \eta \Rightarrow \text{target} \\ < \eta \Rightarrow \text{no target} \end{matrix} \tag{10.55}$$

Schaum (2001) shows that the squared target or *signal-to-clutter ratio* (*SCR*) for a particular target can be expressed as

$$SCR^2(\mathbf{t}) = (\mathbf{t}-\mathbf{m})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{t}-\mathbf{m}) \tag{10.56}$$

and can be used as an estimate of the detectability of target ($\mathbf{t}$). He also points out that this term can be overly optimistic since it assumes that the targets, as they actually occur in the scene, match our estimates of the target spectrum ($\mathbf{t}$) as used in the SCR calculation, and this is seldom the case. Nevertheless, the SCR can be

very useful in assessing the relative detectability of various targets against various backgrounds.

It is important to recognize that the spectral matched filter result [i.e., Eq. (10.49)] is the solution to an optimization problem [cf. Manolakis et al. (2000)] where the target and background are characterized in the following rather restricted fashion. The demeaned background probability distribution function (PDF) is assumed to be multivariate Gaussian with zero mean and covariance matrix **S** expressed as

$$\mathbf{x}-\mathbf{m} \sim N(0,\mathbf{S}) \tag{10.57}$$

where $\sim N(\mathbf{0},\mathbf{S})$ indicates that we are sampling from a multivariate Gaussian distribution with zero mean and covariance **S**. The demeaned target is assumed to be an additive bias to this same clutter yielding a PDF expressed as

$$\mathbf{t}-\mathbf{m} \sim N(\mathbf{t}-\mathbf{m},\mathbf{S}) \tag{10.58}$$

This means that the spectral covariance of the target is assumed to be the same as the background covariance as illustrated in Figure 10.16(a).

Schaum (2001) points out that, even though our data seldom adhere closely to these assumptions (i.e., multivariate backgrounds are seldom multivariate normal unless they include only one spectral class and the target covariance is poorly described by the background covariance), the spectral matched filter may perform well since it is also effective and even optimal for several other (though usually more restrictive) cases. Schaum (2001), for example, points out that the spectral matched filter is optimal for the case where the background is Gaussian and the target is fixed in direction but varies in magnitude to one side of the mean. In this case, the background is still described by Eq. (10.57) and the target is a variable **T** described by a fixed direction $\mathbf{t}_u$ and a variable scaler magnitude t according to

$$\mathbf{T} = (\mathbf{t}-\mathbf{m})_u t = \mathbf{t}_u t \text{ where } 0< t \tag{10.59}$$

This suggests that the spectral matched filter may not be sensitive to the details of the form of the target PDF for at least some range of problems, which explains its widespread use.

In many cases, it is convenient to scale the matched filter to have a value of one in the target direction. This can be achieved by normalizing the filter by its value when operating on the demeaned target according to

$$CEM(\mathbf{x}) = \frac{(\mathbf{t}-\mathbf{m})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}-\mathbf{m})}{(\mathbf{t}-\mathbf{m})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{t}-\mathbf{m})} \tag{10.60}$$

where the algorithm is often referred to in the literature as *constrained energy minimization* (*CEM*) and the output [CEM(**x**)] yields an abundance that is scaled to unity when **x** = **t** [see Farrand and Harsanyi (1997)]. Recognize that for any fixed mean and covariance, the denominator in Eq. (10.60) is a constant and will only scale the matched filter results. It is also important to recognize that the matched filter can be operated in an adaptive fashion just like the RX algorithm described in the previous section. Figure 10.16 illustrates the spectral matched filter concept

and shows the results of running the normalized matched filter [Eq. (10.60)] on the image from Figure 10.15 when the target is chosen to be one of the anomalous pixels found with the RX algorithm.

Manolakis et al. (2001) describe a number of stochastic target detection algorithms related to the SMF. The first is attributed to Kelly (1986) as the solution to *Neyman-Pearson decision/detection theory* for maximizing the probability of detection of a target with a fixed probability of false alarm. The solution uses a *likelihood ratio* (*LR*) *test* expressed as

$$LR(\mathbf{x}) = \frac{PDF_T(\mathbf{x})}{PDF_B(\mathbf{x})} \quad \begin{matrix} \geq \eta \Rightarrow H_1 \text{ i.e. target} \\ < \eta \Rightarrow \mathrm{H}_0 \text{ i.e. background} \end{matrix} \tag{10.61}$$

where $\mathrm{PDF}_T(\mathbf{x})$ and $\mathrm{PDF}_B(\mathbf{x})$ are the target and background probability functions, respectively, and the problem is couched in terms of a binary solution to the hypothesis test where the null hypothesis (H_0) is that the target is absent and the alternate (H_1) is that the target is present. Because we seldom, if ever, know the target probability distribution and often cannot practically characterize the background PDF due to the high dimensional data space, we must, in general, work with estimates of the PDFs. This leads to the *generalized likelihood ratio* (*GLR*), expressed as

$$GLR(\mathbf{x}) = \frac{P_T(\mathbf{x})}{P_B(\mathbf{x})} \tag{10.62}$$

where $P_T(\mathbf{x})$ and $P_B(\mathbf{x})$ are the maximum likelihood estimates of the target and background PDFs. Under the Gaussian assumptions expressed in Eqs. (10.57) and (10.58), the GLR can be solved in the form of a detection metric expressed [after Kelly (1986) and Manolakis et al.(2001)] as

$$GLRT(\mathbf{x}) = \frac{\left[(\mathbf{t}-\mathbf{m})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}-\mathbf{m})\right]^2}{\left[(\mathbf{t}-\mathbf{m})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{t}-\mathbf{m})\right]\left[1+\frac{1}{N}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}-\mathbf{m})\right]} \quad \begin{matrix} \geq \eta \Rightarrow H_1 \\ < \eta \Rightarrow H_0 \end{matrix} \tag{10.63}$$

where N is the number of pixels in the estimate of the covariance ($\mathbf{S}$) and η is a generic threshold variable [i.e., not numerically equal to the value of η used in Eq. (10.61)]. Variations of the GLR*T* are the *adaptive matched filter* (*AMF*) attributed by Manolakis et al. (2001) to Robey et al. (1992) and the *adaptive coherence estimator* (*ACE*) attributed to Conte et al. (1995) and Scharf and McWhorter (1996). They can be expressed as

$$AMF(\mathbf{x}) = \frac{\left[(\mathbf{t}-\mathbf{m})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}-\mathbf{m})\right]^2}{(\mathbf{t}-\mathbf{m})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{t}-\mathbf{m})} \quad \begin{matrix} \geq \eta \Rightarrow H_1 \\ < \eta \Rightarrow H_0 \end{matrix} \tag{10.64}$$

and

$$ACE(\mathbf{x}) \frac{\left[(\mathbf{t}-\mathbf{m})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}-\mathbf{m})\right]^2}{\left[(\mathbf{t}-\mathbf{m})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{t}-\mathbf{m})\right]\left[\frac{1}{N}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}-\mathbf{m})\right]} \begin{matrix} \geq \eta \Rightarrow H_1 \\ < \eta \Rightarrow H_0 \end{matrix} \quad (10.65)$$

where it is important to recognize that the value of the threshold (η) is algorithm specific. Note that all of the detectors (GLR*T*, *AMF*, and *ACE*) have the magnitude of the squared matched filter in the numerator and are normalized by terms proportional to the magnitude of the target in the matched filter space $[(\mathbf{t}\text{-}\mathbf{m})^T \mathbf{S}^{-1} (\mathbf{t} - \mathbf{m})]$ and in some cases the magnitude of the sample in the matched filter space $[(\mathbf{x} - \mathbf{m})^T \mathbf{S}^{-1} (\mathbf{x} - \mathbf{m})]$.

Because real data do not, in general, match the assumptions from which these algorithms are derived [i.e., Eqs. (10.57) and (10.58)], we generally cannot expect one of them to be optimal or even for one to consistently outperform another (which is why we introduce so many detectors). Nevertheless, Manolakis et al. (2001) show, that for a limited data set all these algorithms show some success at target detection, with the ACE algorithm showing superior performance for the limited data set tested.

None of the algorithms we have explored thus far have included an estimate of the distribution of the target signature. This is not because we believe the target signature will have a very low variability, but rather because we seldom have sufficient information to make a good estimate of the target variability. In some cases, there may be locations where there are enough target pixels to characterize the target's spectral covariance, or we may be able to model the range of ways the target may appear to an extent where we can estimate the covariance using physics-based models (cf. Chap. 11). On these rare occasions, if we assume the target distribution is approximately Gaussian, we can use the optimum Nieman-Pierson detector known as the quadratic detector, which takes the form

$$Q(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{m})^{\mathrm{T}}\boldsymbol{S}_b^{-1}(\boldsymbol{x}-\boldsymbol{m}) - (\boldsymbol{x}-\boldsymbol{t})^{\mathrm{T}}\boldsymbol{S}_t^{-1}(\boldsymbol{x}-\boldsymbol{t}) \begin{matrix} \geq \eta \Rightarrow H_1 \\ < \eta \Rightarrow H_0 \end{matrix} \quad (10.66)$$

where $\mathbf{t}$ is the mean target vector and $\mathbf{S}_t$ is the target spectral covariance matrix. Note that the quadratic detector can be described as the difference between the squared Mahalanobis distance to the target and the squared Mahalanobis distance to the background.

Before leaving the topic of stochastic detectors, it is important to recognize that these operators can, in theory, be applied in any space where the target and background can be characterized (i.e., counts, radiance, or reflectance). However, for practical reasons, it is often desirable to perform the analysis after transforming the image and the target vector into a reduced dimension. This can speed processing, reduce noise, better condition the data (i.e., the covariance matrices are more likely to invert in the transform space), and push the data to better match the assumptions of the operators (e.g., Gaussian PDFs). Furthermore, all these operators can be applied at varying levels of background adaptation by computing the background statistics anywhere between globally and very locally in a spatial sense

or by using cluster statistics for the appropriate background class for a scene that has been presegmented [cf. West et al. (2005)]. Finally, it should be recognized that while we have tended to treat these detection algorithms as though they are designed to find only a single target class, they can, in fact, be used repeatedly to map various features. For example, the normalized SMF or CEM algorithm can be used to generate abundance maps of many target classes with results that are conceptually, and often quantitatively, comparable to mixture fraction maps generated by the structured geometric models discussed in Section 10.3.2.

10.4.2.3 Target Detection When the Target Shape Is Known

Up to this point in this chapter, we have focused on analytical approaches that rely solely on the spectral character of the individual pixel to characterize the target. For very large spatial features or features of the order of a pixel or less, this may be the most reasonable approach. However, if we are seeking targets whose spatial extent is predictable and finite (e.g., 5-12 contiguous clustered pixels) or perhaps even of a particular spatial pattern (e.g., approximately rectangular covering 5 × 9 pixels), it may pay to include spatial processing with the spectral algorithms.

Reed and Yu (1990), in developing the unknown target RX anomaly detection operator, first describe an approach to target detection when the spatial and spectral character of the target is known. The spectral part of the algorithm is consistent with the assumptions introduced in Section 10.4 for the stochastic algorithms and expressed in Eqs. (10.57) and (10.58). For convenience, we can write an expression for the N pixels in an operating window of interest as

$$\mathbf{X} = \begin{bmatrix} x_{11} - m_1 & x_{21} - m_1 & & x_{N1} - m_1 \\ x_{12} - m_2 & x_{22} - m_2 & \cdots & x_{N2} - m_2 \\ \vdots & \vdots & & \vdots \\ x_{1j} - m_j & x_{2j} - m_j & & x_{Nj} - m_j \end{bmatrix} = [\mathbf{y}_1, \mathbf{y}_2 \cdots \mathbf{y}_N] \tag{10.67}$$

where each column is a demeaned spectral pixel vector ($\mathbf{y}$) and sequential rows in the image window are concatenated to form the matrix $\mathbf{X}$. Thus, a ℓ row by m column window region with j spectral bands will form a j by N matrix, where $N = m \times \ell$. The target has an additive spectral character described as an additive spectral bias ($\mathbf{d}$) to the demeaned data, i.e.,

$$\mathbf{d} = \mathbf{t} - \mathbf{m} = \begin{bmatrix} t_1 - m_1 \\ t_2 - m_2 \\ \vdots \\ t_j - m_j \end{bmatrix} \tag{10.68}$$

and a spatial shape characterized by a vector made up of the target intensities (abundances) in each pixel according to

$$\boldsymbol{a}^{\mathrm{T}} = [a_1, a_2 \cdots a_{\mathrm{N}}] \tag{10.69}$$

where the target intensities (abundances) are normalized such that

$$\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}=1 \tag{10.70}$$

Thus, the additive signal in the nth demeaned pixel would be $a_n\mathbf{d}$, a background (clutter) pixel would be $\mathbf{y}_n$, and a pixel with some target present would be $\mathbf{y}_n + a_n\mathbf{d}$. Reed and Yu (1990) derive an expression for target detection for the case where $\mathbf{d}$ and $\mathbf{a}$ are assumed known using the following logic.

We once again seek to decide between the target absent hypothesis (H_0) and the target present hypothesis (H_1). Based on the work of Margalit et al. (1985) and Hunt and Cannon (1976), Reed and Yu (1990) suggest that we approximate the spectral data set as a nonstationary Gaussian random process with a rapidly varying mean and a slowly varying spectral covariance. Thus, if we demean the data in a small window, we should be able to assume that the local clutter has stationary Gaussian statistics. Thus, the demeaned data in a local window with spectral covariance matrix $\mathbf{S}$ can be approximated to have a Gaussian joint probability distribution:

$$p_o(\mathbf{X})=p_o[\mathbf{y}_1,\mathbf{y}_2\cdots\mathbf{y}_N]=(2\pi)^{-N\frac{j}{2}}|\mathbf{S}|^{-\frac{N}{2}}e^{-\frac{1}{2}\sum_{n=1}^{N}\mathbf{y}_n^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y}_n} \tag{10.71}$$

When a target is present, the mean will be shifted by the additive signal and the joint distribution would be expressed as

$$p_1(\mathbf{X})=p_1[\mathbf{y}_1,\mathbf{y}_2\cdots\mathbf{y}_N]=(2\pi)^{-N\frac{j}{2}}|\mathbf{S}|^{-\frac{N}{2}}e^{-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{y}_n-a_n\mathbf{d})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{y}_n-a_n\mathbf{d})} \tag{10.72}$$

Note that the joint probability is the product of the individual pixel probabilities resulting in the summation in the exponent. Stocker et al. (1990) point out that the log likelihood ratio of the probability density functions will yield an optimal detector (given that our assumptions are valid) according to

$$\ell(\mathbf{X})=\ell n\left[\frac{p_1(\mathbf{X})}{p_0(\mathbf{X})}\right]\begin{matrix}\geq\eta\Rightarrow H_1\\ <\eta\Rightarrow H_0\end{matrix} \tag{10.73}$$

Substituting Eqs. (10.71) and (10.72) into (10.73) and simplifying yields

$$\ell(\mathbf{X})=\ell n\left[e^{-\frac{1}{2}\sum_{n=1}^{N}(y_n-a_n\mathbf{d})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{y}_n-a_n\mathbf{d})+\frac{1}{2}\sum_{n=1}^{N}\mathbf{y}_n^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y}_n}\right]\begin{matrix}\geq\eta\Rightarrow H_1\\ <\eta\Rightarrow H_0\end{matrix} \tag{10.74}$$

$$=-\frac{1}{2}\sum_{\mathrm{n=1}}^{\mathrm{N}}(\mathbf{y}_n-a_n\mathbf{d})^{\mathrm{T}}\mathbf{S}^{-\frac{1}{2}}\mathbf{S}^{-\frac{1}{2}}(\mathbf{y}_n-a_n\mathbf{d})+\frac{1}{2}\sum_{n=1}^{N}\mathbf{y}_n^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y}_{\mathrm{n}}\begin{matrix}\geq\eta\Rightarrow H_1\\ <\eta\Rightarrow H_0\end{matrix} \tag{10.75}$$

$$=-\frac{1}{2}\sum_{n=1}^{N}\left(\mathbf{S}^{-\frac{1}{2}}\mathbf{y}_n-\mathbf{S}^{-\frac{1}{2}}a_n\mathbf{d}\right)^{\mathrm{T}}\left(\mathbf{S}^{-\frac{1}{2}}\mathbf{y}_n-\mathbf{S}^{-\frac{1}{2}}a_n\mathbf{d}\right)+\frac{1}{2}\sum_{n=1}^{N}\mathbf{y}_n^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y}_{\mathrm{n}}\begin{matrix}\geq\eta\Rightarrow H_1\\ <\eta\Rightarrow H_0\end{matrix} \tag{10.76}$$

$$= -\frac{1}{2}\sum_{n=1}^{N}\mathbf{y}_n^{\mathrm{T}}S^{-1}\mathbf{y}_n + \sum_{n=1}^{N}a_n\mathbf{d}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y}_n - \frac{1}{2}\sum_{n=1}^{N}a_n\mathbf{d}^{\mathrm{T}}\mathbf{S}^{-1}a_n\mathbf{d} + \frac{1}{2}\sum_{n=1}^{N}\mathbf{y}_n^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y}_n \begin{matrix} \geq \eta \Rightarrow H_1 \\ < \eta \Rightarrow H_0 \end{matrix} \tag{10.77}$$

$$= \sum_{n=1}^{N}a_n\mathbf{d}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y}_n - \frac{1}{2}\sum a_n\mathbf{d}^{\mathrm{T}}\mathbf{S}^{-1}a_n\mathbf{d} \begin{matrix} \geq \eta \Rightarrow H_1 \\ < \eta \Rightarrow H_0 \end{matrix} \tag{10.78}$$

Finally, by moving all terms not dependent on the sample to the right-hand side, they produce the following *spatial spectral matched filter* (*SSMF*) solution:

$$SSMF(\mathbf{X}) = \sum_{n=1}^{N}a_n\mathbf{d}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y}_n \begin{matrix} \geq \overline{\eta} \Rightarrow H_1 \\ < \overline{\eta} \Rightarrow H_0 \end{matrix} \tag{10.79}$$

where the threshold $\overline{\eta}$ is modified to account for non-sample-specific operations employed in the simplification process. This SSMF expression can be further simplified to yield

$$SSMF(\mathbf{X}) = \mathbf{d}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{X}\boldsymbol{a} \tag{10.80}$$

Inspection of Eq. (10.80) indicates that we can break this process down into an operation by the spectral matched filter on the demeaned data to yield a gray-level image followed by a spatial matched filter that can be thought of as a kernel operation with weights a_n at each location (cf. Fig. 10.17) to yield the final scaler value for the window to be compared to the detection threshold.

In this section, we have rigorously derived an expression for the SSMF. However, the general approach of first applying a spectral operator followed by a spatial kernel operation can be adopted to essentially all of the spectral operations presented thus far, including both the structured operators of Section 10.3.2 and the stochastic operators of Section 10.4.2.1 and 10.4.2.2.

Before leaving this section, it is important to note several important implementation variations with the stochastic operators. In many cases, the algorithms are run using imagewide means and covariances. However, most spectral data have rapidly varying means and slowly varying covariances. Furthermore, the imagewide data set is generally not well represented by a simple multivariate Gaussian model. As a result, it may, in some cases, be worth the incremental complexity to compute the means much more locally (i.e., on a window about the pixel of interest) and to also use a window region (though possibly not as local) to compute the covariance matrix estimate. Note that many of the algorithms presented above assume that the covariance is only indicative of the variation about the background mean. Therefore, we often seek to remove any target-like features from the estimate. This can be attempted by not including (masking) the pixel under test (or a region around the pixel under test) in the covariance matrix estimate.

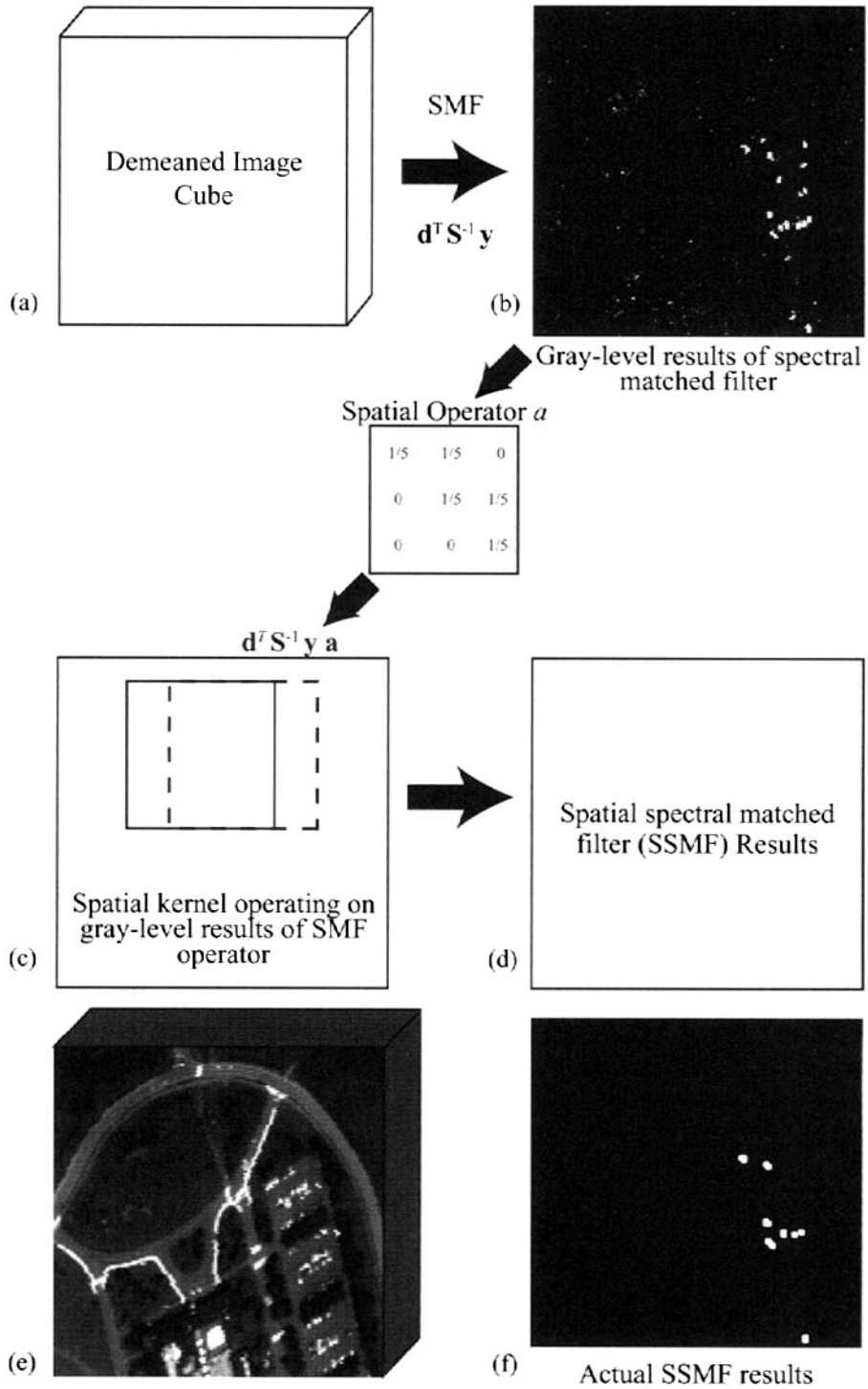


Figure 10.17 Illustration of the spatial spectral matched filter concept. A demeaned image cube (a) is operated on with a spectral matched filter (often employing a slowly varying or fixed covariance and a locally varying mean) to yield a gray-level image product (b), which is operated on with a spatial kernel, (c) and thresholded to yield the final SSMF product (d). An actual image cube (e) and final product (f) are shown for a MISI 64-band image.

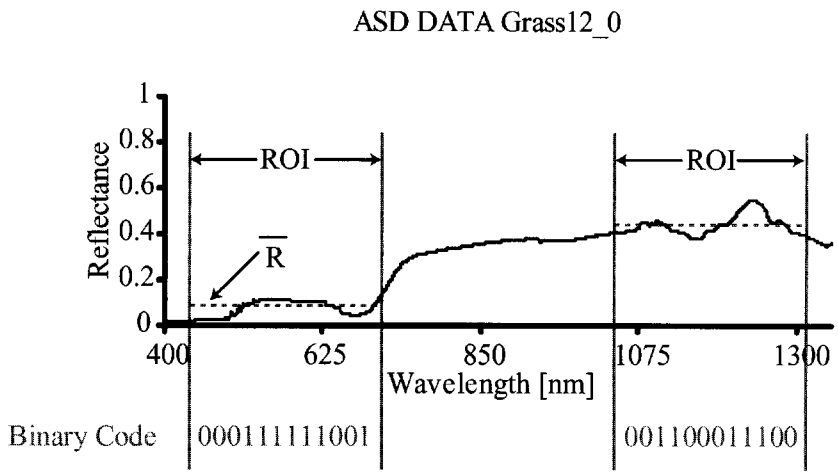
10.5 SPECTRAL FEATURE APPROACHES TO SPECTRAL IMAGE ANALYSIS

As introduced in Section 10.1.3, the *spectral feature approach* uses localized spectral features as indicators of the type or condition of a material. An early example of this approach that was first applied to multispectral data and has been extended with greater spectral detail to imaging spectroscopy is the *normalized difference vegetation index* (*NDVI*) (cf. Sec. 7.5.1). Recall that the NDVI presumes that we are looking at vegetation, often even at a specific species, and uses the strength of the red chlorophyll absorption feature relative to the infrared continuum as a measure of vegetation health/abundance. With access to finer spectral detail, other absorption-feature-based indices have been developed. For example, Gao (1996) describes the *normalized difference water index* (*NDWI*) that is designed to characterize the relative amount of liquid water in vegetation. In this case, we again assume the pixel contains vegetation and use the index to characterize the strength of the liquid water absorption feature expressed as

$$NDWI = \frac{R_{860} - R_{1240}}{R_{860} + R_{1240}} \tag{10.81}$$

where R_{860} and R_{1240} are the estimated surface reflectances at 860 nm and 1240 nm ,respectively. Like the NDVI, the NDWI uses differences and ratios to reduce illumination, calibration, and atmospheric correction effects. Because most indices of this type are very application specific, they are not emphasized here but the reader is encouraged to consider them for specific applications where an absorption feature in a preidentified class may be indicative of class condition.

Of more global interest is the use of spectral features as a means of mapping material classes based on libraries of spectral features. We first introduce a relatively crude approach to this as a means of illustrating the concept. The *binary encoded matched filter* (*BEMF*) is a simple means of analyzing spectral features. For convenience, we will assume the image has been converted to reflectance spectra and that any library reference spectra have been spectrally convolved and resampled to match the imaging sensor's spectral response characteristics. The library spectra can then be processed to characterize one or more spectral features of interest. This is accomplished in the following fashion (cf. Fig. 10.18). Over each spectral window of interest (ROI), the local mean is computed and the brightness for each band in the window is compared to the mean. Values greater than or equal to the mean are assigned a value of one, and those less than the mean are assigned a value of zero. If multiple windows are used, the binary results are concatenated to form a single binary vector. The same process is then conducted on the identical bands of each image pixel to be analyzed. The library- and image-derived binary vectors are then compared to determine what percentage of the binary values in the vectors are the same. The results can then be gray-scale encoded or thresholded. The user must expertly select the spectral regions of interest for each material to be analyzed. The choice requires consideration, not only of which spectral features are indicative of the target, but also of what other materials might have similar features and therefore cause false alarms. This approach is attractive because different



Binary Encoding and Matching

Reference:	000111111001001100011100
Image:	001110110011101100010000
Match:	110110110101011111110011
Match %:	17/24 = 71%

Figure 10.18 Illustration of the binary encoded matched filter (BEMF) concept: (a) binary encoding process and (b) matching process.

spectral regions can be chosen for each material and multiple detections per pixel are possible (i.e., mixed pixel solutions are possible). A downside to this approach is that the percent match often does not give a good indicator of the expected abundance within a pixel and may not give a strong indicator of the quality of the fit relative to more subtle features masked by the binarization process.

A much more sophisticated algorithm described by Clark et al. (2003) forms the basis for the *Tetracorder* processing software developed by the U.S. Geologic Service (USGS) for spectroscopic image analysis. The Tetracorder approach maintains all the advantages of the BEMF but overcomes many of the limitations of the simpler approach. We will once again assume the image and spectral library data have been converted into a common reflectance space. The Tetracorder approach requires an expert to define a small number of absorption features of interest for any material, including the spectral regions that define the wings or continuum outside the feature(s) and the band center(s) (i.e., deepest point in the feature). From these data, the band depth (D) can be computed for both the library spectrum and the spectrum under test (cf. Fig. 10.19) according to

$$D = 1 - R_B / R_C \tag{10.82}$$

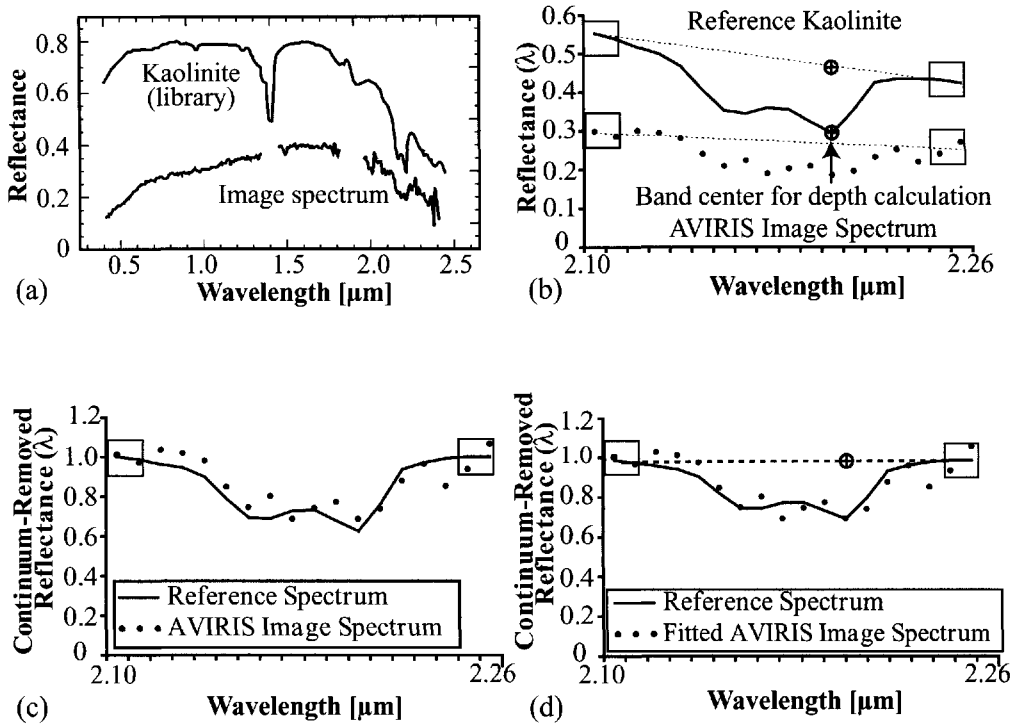


Figure 10.19 Illustration of the Tetracorder approach to spectral featureanalysis: (a) image and library reflectance spectra; (b) detail of a spectralabsorption feature showing expert defined wing regions (boxes), the band center, continuum interpolated band center, and the continuum line; (c) continuum removed library spectrum [$R_L(\lambda)$] in solid line and continuum removed image spectrum $R_I(\lambda)$ (dots); and (d) the results of fitting the continuum removed library spectrum to the continuum-removed image spectrum.

where R_B is the reflectance at the absorption band center (strongest absorption feature) and R_C is the continuum interpolated reflectance at the same wavelength. The continuum is linearly interpolated between the mean values of the wings as defined by the expert to obtain R_C. The continuum-interpolated values for each band location are also used to generate a continuum-removed spectrum for the library and image spectra according to

$$\hat{R}_L(\lambda) = R_L(\lambda) / R_{LC}(\lambda) \tag{10.83}$$

$$\hat{R}_I(\lambda) = R_I(\lambda) / R_{IC}(\lambda) \tag{10.84}$$

where $\hat{R}_L(\lambda)$ and$\hat{R}_I(\lambda)$ are the continuum-removed spectral values for the library (L) and image (I) spectra, $R_L(\lambda)$ and$R_I(\lambda)$ are the spectral library and image reflectance values, and $R_{LC}(\lambda)$ and$R_{IC}(\lambda)$ are the library and image continuum interpolated spectral values. The continuum-removed spectra are linearly regressed against each other to remove simple linear contrast differences and estimate a goodness of fit from the R^2 value of the regression. The regression equation takes the form

$$\hat{R}_I(\lambda) = a_0 + a_1 \hat{R}_L(\lambda) \tag{10.85}$$

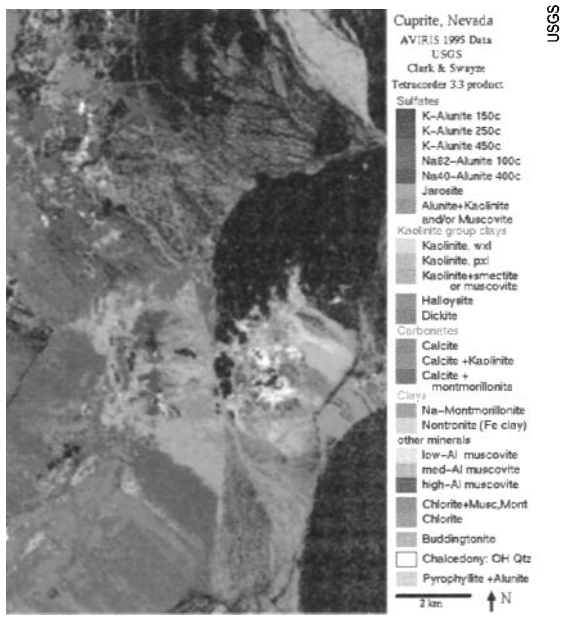


Figure 10.20 Illustration of material map produced by Tetracorder where each pixel is assigned to material with the highest R^2D value. See color plate 10.20.

where a_0 and a_1 are regression coefficients and the R^2 value of the regression is related to how much of the variation in $\hat{R}_f(\lambda)$ is explained by the model of the absorption feature in the library. Thus, for each spectral feature we now have two metrics, the band depth, which we can associate with an abundance, and the goodness of fit or R^2 value, which we can associate with a confidence. Clark et al. (2003) suggest that the product of these two terms (R^2D) may be a good parameter to use to visualize/map the material of interest. This map can be thresholded to generate target detection maps if desired. When more than one feature is used for a material, the Tetracorder approach uses an area-weighted average based on the area under the continuum-removed library spectrum for each feature. Each metric D, R^2 and R^2D is computed as the area-weighted average of each feature included in the analysis. The resulting combined metrics can be used to estimate abundance and confidence and to visualize/map each material considered. The Tetracorder analytical approach includes not only the steps described here, but also the expert values (e.g., band centers, and wings) for a large number of materials (particularly minerals). Finally, some additional rules are employed for certain materials to ensure consistency with mean brightness levels, continuum slope, presence or absence of other absorption features, etc., or to return a null solution if the spectral match is too low. Figure 10.20 shows the result of running Tetracorder in a mineral mapping mode on an AVIRIS image of Cuprite, Nevada. In this mode, each material of interest is mapped to compute abundance/likelihood, and then multiple maps can be combined to generate color-coded mineral maps.

In closing this section, it is important to recognize that the spectral feature approach requires a user who can identify not only the spectral features in the target

of interest but also the features in the potential false alarms so that an algorithm can be tuned to differentiate between them. Particularly in the area of mineral mapping, the Tetracorder software available from the USGS includes much of this expert knowledge, relieving the user of defining the critical spectral parameters.

10.6 HYBRID APPROACHES TO SPECTRAL IMAGE ANALYSIS

In the discussions above, we have tended to treat the various perspectives on spectral data independently and have presented algorithms largely based on a single view at a time. In this section, we want to make it clear that these perspectives are not mutually exclusive. In fact, at times we can use one perspective to help in visualizing another or combine multiple perspectives to try to develop advanced algorithms.

We begin by viewing the spectral matched filter from the perspective of geometric or vector transformations as a way to help to visualize the stochastic algorithm [cf. Stocker et al. (1990)]. We begin by reexpressing the spectral matched filter using the decomposition of the covariance matrix according to

$$\mathbf{S} = \mathbf{E}\Lambda\mathbf{E}^{\mathbf{T}} \tag{10.86}$$

which leads to

$$\begin{aligned} SMF(\mathbf{x}) &= (\mathbf{t}-\mathbf{m})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}-\mathbf{m}) = (\mathbf{t}-\mathbf{m})^{\mathrm{T}}(\mathbf{E}\Lambda\mathbf{E}^{\mathrm{T}})^{-1}(\mathbf{x}-\mathbf{m}) \\ &= (\mathbf{t}-\mathbf{m})^{\mathrm{T}}(\mathbf{E}\Lambda^{-1}\mathbf{E}^{\mathrm{T}})(\mathbf{x}-\mathbf{m}) \\ &= (\mathbf{t}-\mathbf{m})^{\mathrm{T}}(\mathbf{E}\Lambda^{-\frac{1}{2}}\Lambda^{-\frac{1}{2}}\mathbf{E}^{\mathrm{T}})(\mathbf{x}-\mathbf{m}) \\ &= \left[(\mathbf{t}-\mathbf{m})^{\mathrm{T}}\mathbf{E}\Lambda^{-\frac{1}{2}}\right]\left[\Lambda^{-\frac{1}{2}}\mathbf{E}^{\mathrm{T}}(\mathbf{x}-\mathbf{m})\right] \end{aligned} \tag{10.87}$$

where we first take advantage of the real-symmetric nature of the covariance matrix to reexpress it using an eigenvector decomposition with the matrix $\mathbf{E}$ composed of eigenvectors as columns and Λ the diagonal matrix of eigenvalues. Next we recall that the inverse of $(\mathbf{E}\Lambda\mathbf{E}^{T})$ is $(\mathbf{E}\Lambda^{-1}\mathbf{E}^{T})$ and that $(\Lambda^{-1}) = (\Lambda^{-½}\Lambda^{-½})$, where $\Lambda^{-½}$ is a diagonal matrix with one over the square root of the eigenvalue (i.e., $1/\sqrt{\lambda}$) as each entry along the diagonal. Thus, we can view the SMF as a two-step process. In the first step, each demeaned vector $(\mathbf{x} - \mathbf{m})$ is projected onto the eigenvectors of the covariance matrix, and each transformed band in this new rotated space is normalized by the standard deviation (i.e., the expected variability) in that band i.e., $(1/\sqrt{\lambda})$. This is a clutter, whitening process where the resulting transformed data should have the same variance in each band. The second step in the algorithm involves projecting each transformed vector onto a similarly transformed target vector to see how "target-like" it is in the transformed space (cf. Fig. 10.21). While this stochastic perspective does not treat the clutter as structured, it gives us a way to view the stochastic data from a vector projection or vector space perspective that

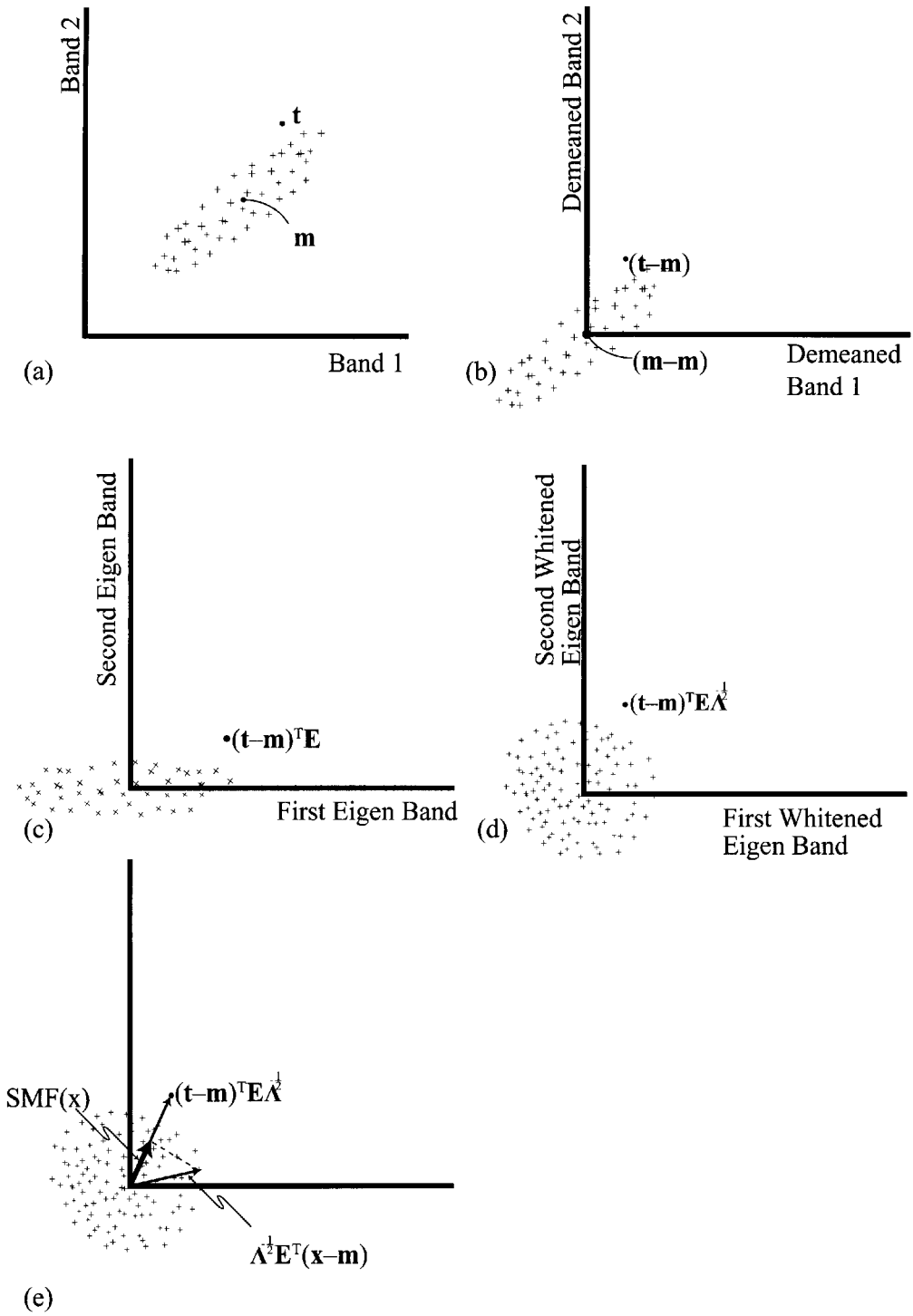


Figure 10.21 Vector geometry interpretation of the spectral matched filter: (a) original data, (b) demeaned data, (c) Eigen projection space (rotation), (d) Eigen projection and whitened space, and (e) the matched filter in the whitened space is simply the projection onto the transformed target vector.

is similar to the geometric perspective and that may provide improved insights for some readers.

A more explicit combination of the structured and stochastic approaches is suggested by Funk et al. (2001). They point out that the clutter will seldom have the Gaussian properties implicitly assumed in using the spectral matched filter (SMF). However, if the clutter can be segmented into a single spectral class, the clutter may be much better represented by a Gaussian model, and the target may be more readily separable using the SMF or its variants. Thus, they suggest beginning from a structured perspective and assigning each pixel to a class using a spectral clustering algorithm. The authors use a *k*-means approach (cf. Sec. 9.2.3); however, alternate spectral clustering approaches (e.g., SEM presented in Sec. 10.4.1) might be considered. Based on this clustering, the mean ($\mathbf{m}_i$) and covariance ($\mathbf{S}_i$) of each clutter class can be computed. At this point, we switch to a stochastic target detection approach and seek to detect a target in the presence of Gaussian background clutter. This points us to the spectral matched filter. However, rather than use a global or spatially local covariance, we use the covariance of the spectral cluster (i) that the pixel is assigned to based on the clustering algorithm. The matched filter for this case is expressed as

$$SMF(x) = (\mathbf{t} - \mathbf{m}_i)^{\mathrm{T}} \mathbf{S}_i^{-1} (\mathbf{x} - \mathbf{m}_i) \tag{10.88}$$

As illustrated in Figure 10.22, this reduces the problem to looking for the target in spectrally local clutter that can be more compactly compressed and whitened by the matched filter operator increasing the likelihood of detection. Funk et al. (2001) report significant increases in the signal-to-clutter ratio (factors of 2 to 3) when the cluster-specific statistics are used in place of the global statistics for the synthetic LWIR spectral data cubes they tested.

It is also important to recognize that this hybrid approach can be applied using any of the stochastic anomaly or target detection algorithms introduced in Section 10.4 by substituting cluster statistics for the global or spatial window statistics normally used. Furthermore, the hybrid approach can also be applied to reduced dimension data to reduce noise and processing time and to better condition the data.

Stocker and Schaum (1997) present another interesting hybrid approach using *stochastic mixing models* (*SMM*) that enables us to generate statistical descriptions of linear mixtures of structured backgrounds. This approach suggests that the background clutter (after appropriate dimensionality reduction) is composed of structured clusters distributed about what the authors call hard end members. Fractional mixtures of hard end members can form a dispersed continuum of spectral vectors between the hard end members. They suggest describing this spectral space using Gaussian distribution to describe the clusters about the hard end members based on cluster-specific statistics (i.e., $\mathbf{m}_i$ and $\mathbf{S}_i$) and also using discrete Gaussian distributions located at fractional steps between the hard means to describe the mixed pixels. To illustrate this approach, consider the two-class problem shown in Figure 10.23. If we let $\mathbf{m}_1$ and $\mathbf{m}_2$ to be the mean vectors of the two classes and $\mathbf{S}_1$ and $\mathbf{S}_2$ the covariance matrices; then we can approximate the cloud of data between using Gaussians based on mixing fractions f_i for each cluster. In general, the vector

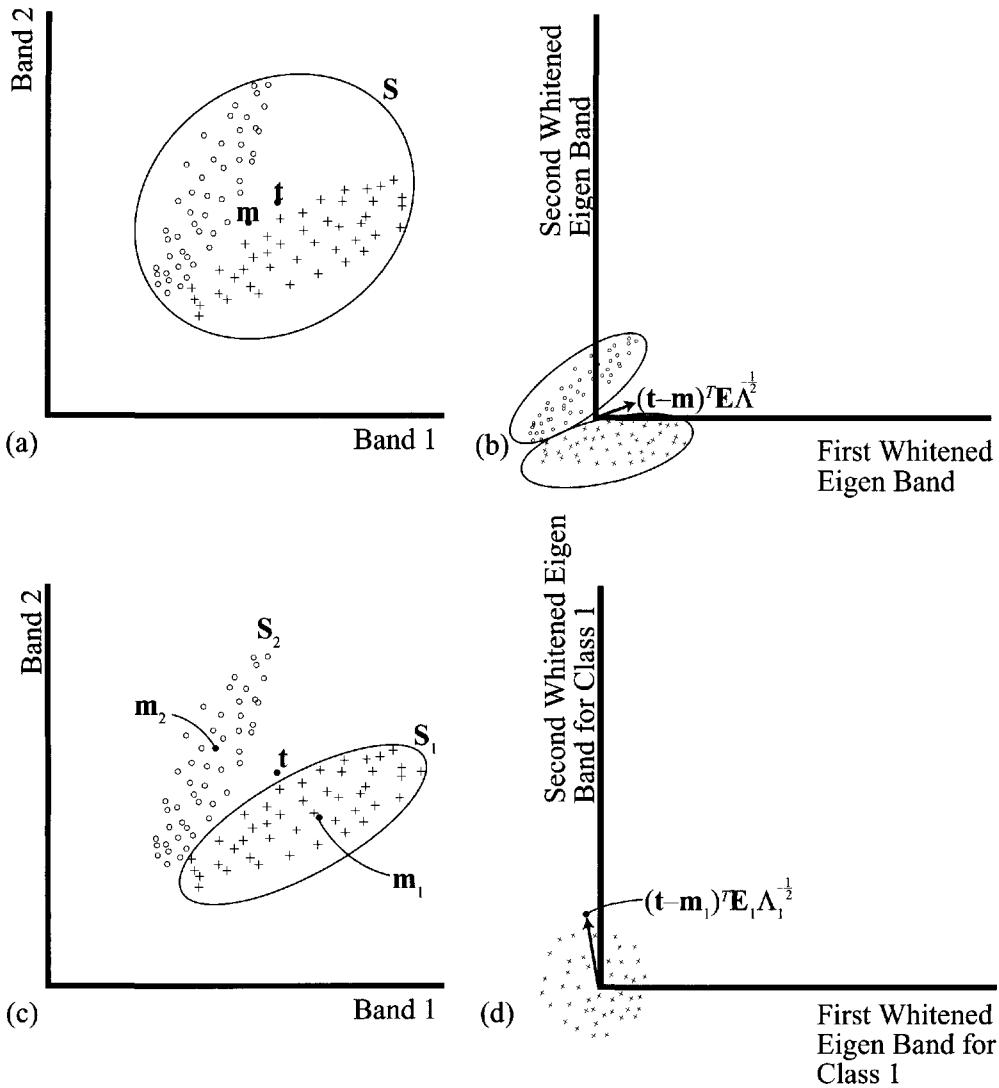


Figure 10.22 Simple two-band illustration of the use of the spectral matched filter with spectral clutter statistics derived from the closest clutter class to the pixel under test. Using the concepts illustrated in Figure 10.21 for visualization of the SMF operator, (a) shows the original data with the overall mean (**m**) and the curve **S** representing an isoprobability surface generated by the overall covariances, (b) shows the data in the clutter whitened space described in association with Figure 10.21 (note that the target is very difficult to distinguish from the scene clutter using the overall clutter statistics), (c) shows the original data including the class means and isoprobability surfaces based on class covariances [Note the target pixel in this case would have been classified into class 1, so we show only the class 1 operator in (d)], and (d) shows the class 1 clutter-whitened space where the target is more separable (note that if the target were compared using a class 2-based SMF, it would be even more separable).

f_c will contain the mixing fractions of the hard end members that combine to form the soft end members. Hard end members will have f_c vectors made up of a one and zeros, while soft end members will have values less than or equal to one that sum to one. By setting a fractional step size Δf, the user controls the number of soft end members. For example, in our two class case, if Δf=1/7, there will be eight cluster centers, two of them representing the hard end members and six of them the soft end members. Stocker and Schaum (1997) point out that the class means and covariances for the soft end member classes can be computed as

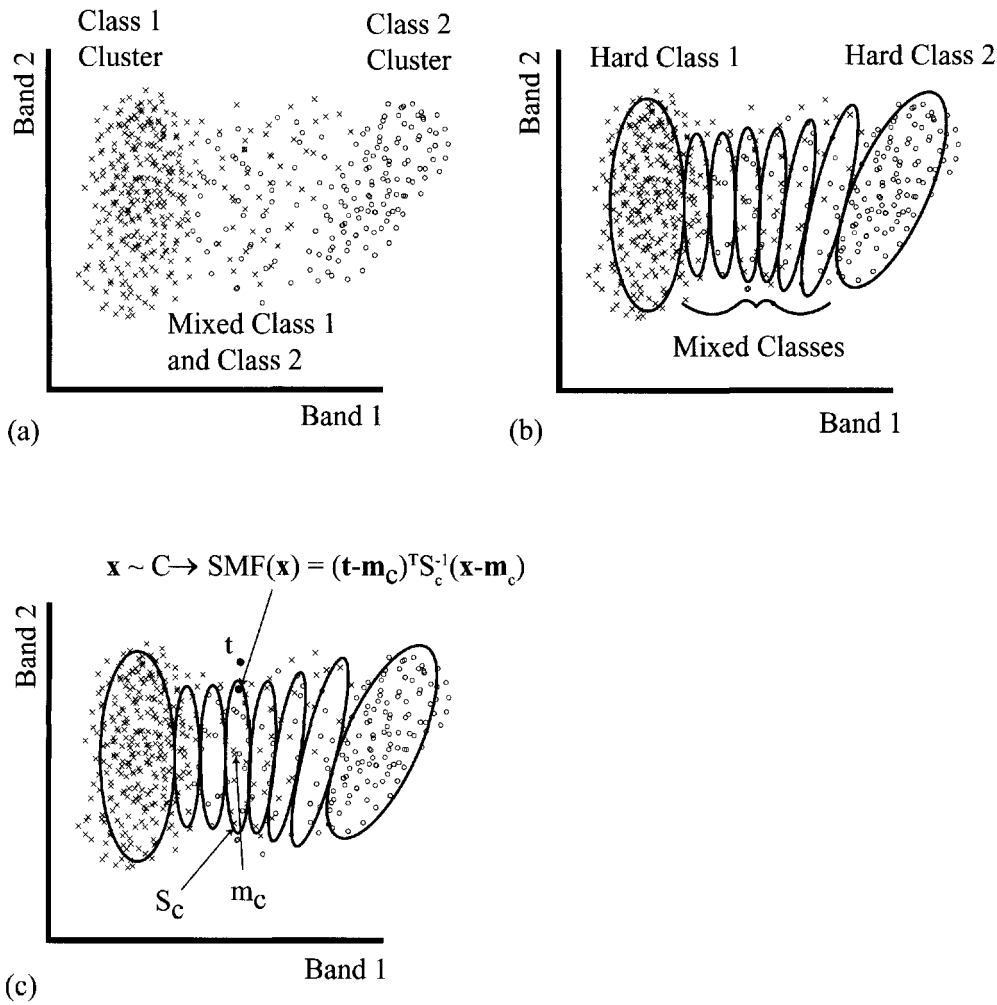


Figure 10.23 Illustration of stochastic mixing model (SMM) concept shown for a simple two-class, two-band case: (a) data shown in two bands of a reduced-dimension space showing two spectral clusters, (b) isoprobability contours illustrating clusters associated with hard end member clusters and mixed classes, and (c) illustration of a target vector and a test pixel that had been assigned to mixed class (c) with mean $\mathbf{m}_c$ and covariance $\mathbf{S}_c$ that can be assessed for target-like character using an SMF operator.

$$\mathbf{m}_c = \sum_{i=1}^{H} f_i \mathbf{m}_i \tag{10.89}$$

$$\mathbf{S}_c = \sum_{i=1}^{H} f_i^2 \mathbf{S}_i \tag{10.90}$$

where $\mathbf{m}_c$ and $\mathbf{S}_c$ are the mean and covariance of the cth soft end member clutter class, H is the number of hard end members, and the f_i values are the scaler fractions that make up the mixing vector $\boldsymbol{f}_c$ that characterizes the cth class. For our two-class problems, the classes can be characterized by the following $\boldsymbol{f}_c$ vectors:

$$\boldsymbol{f}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \boldsymbol{f}_2 = \begin{bmatrix} 6/7 \\ 1/7 \end{bmatrix}, \boldsymbol{f}_3 = \begin{bmatrix} 5/7 \\ 2/7 \end{bmatrix}, \cdots \boldsymbol{f}_8 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{10.91}$$

Note that if the number of hard classes is large or we allow mixtures of more than two hard classes, the number of mixed classes grows rapidly. Thus, the authors suggest that this approach is most appropriate for low-contrast targets (i.e., where simpler approaches fail) and where the clutter can be reduced to a small number of hard end member classes. The problem arises of how to define the hard end member statistics and to assign pixels to the hard and soft classes. Stocker and Schaum (1997) suggest using a slightly modified version of the SEM algorithm described in Section 10.4.1. The modification simply calls for an adjustment to the step where the class statistics are computed. The a priori probabilities are computed normally for all the hard and soft classes based on the class assignments, and the means and covariance are computed in the normal fashion for the hard end member classes. However, the means and covariance for the mixed or soft end member classes are computed using Eqs. (10.89) and (10.90). The a posteriori probability calculations and class assignments proceed as normal, as does the iterative process. The authors suggest that tens of iterations may be necessary using the initialization criteria described in Section 10.4.1 and a stopping criteria using the average log likelihood expressed as

$$L = \frac{1}{N} \sum_{j=1}^{N} \ln \sum_{i=1}^{C} p(i) p(\boldsymbol{x}_j | i) \tag{10.92}$$

where C is the total number of classes (hard and soft), N is the total number of pixels, and L should increase and then stabilize, so the iterative process is stopped when L goes into small fluctuations about a steady-state value. The final class assignments that result from this process provide mixture values for each pixel based on the discrete mixture model for that class defined by the $\boldsymbol{f}_c$ vector. Thus, the SMM provides a statistical approach to describe a structural mixing model. Note, however, that this approach is mixing cluster probability distributions, not extrema, and only discrete steps are allowed, so subtle differences with wholly structured models should be expected. Stocker and Schaum (1997) point out that the utility of the SMM approach is in detection of low-contrast targets and suggest methods to use the SMM clutter models to find both extended (multiple pixel) and subpixel targets.

A simple approach would be to apply an SMF (or other stochastic filter) using the SMM assigned class as the clutter class. This is essentially the approach suggested by Funk et al. (2001) with the SMM used to even more locally define the clutter.

Stocker and Schaum (1997) go a step further, suggesting that we can generate a stochastic mixing model of the target with a background. This requires a statistical estimate of the mean and covariance of each class. Then we can describe the probability that a pixel contains fractional target contribution f using a Gaussian mixing model:

$$p(\mathbf{x}|f) = N[\mathbf{m}(f), \mathbf{S}(f)] \tag{10.93}$$

where

$$\mathbf{m}(f) = f\mathbf{m}_t + (1-f)\mathbf{m}_b \tag{10.94}$$

$$\mathbf{S}(f) = f^2\mathbf{S}_t + (1-f)^2\mathbf{S}_b \tag{10.95}$$

$\mathbf{m}_t$ and $\mathbf{m}_b$ and $\mathbf{S}_t$ and $\mathbf{S}_b$ are the means and covariances of the target and background. The background statistics could be computed globally, in a local spatial window or based on spectral proximity (e.g., SMM). The target mean is typically computed using a scene- or library- derived estimate. In the common case where little additional knowledge is available to estimate the target covariance, we assume the target covariance will mimic the noise covariance. For each pixel, we can then make a best estimate of the abundance f of target using either a simple linear mixing model or the SMM approach discussed above. We can then use the target-background SMM to generate a generalized likelihood ratio test, which Stocker and Schaum (1997) call the *finite target matched filter* (*FTMF*), expressed as

$$FTMF(x) = \frac{p(\mathbf{x}|f)}{p(\mathbf{x}|f=0)} \tag{10.96}$$

where the numerator is a Gaussian estimate of a target background SMM based on Eq. (10.93) using the best estimate (most probable value) of f. The denominator is simply a Gaussian using only the background statistics. The authors suggest that the resulting decision space has the form of a truncated hyperplane that narrows near the target and expands toward the background.

The SMM-based FTMF is similar to the popular *mixture turned matched filter* (*MTMF*) described by Boardman (1998). Boardman (1998) suggests operating in the minimum noise fraction (*MNF*) space (cf. Fig. 10.24) where, using an approach analogous with that used in Eq. (10.50), we can express the spectral matched filter as

$$SMF(\mathbf{x}) = (\mathbf{t}-\mathbf{m})^{\mathrm{T}}\mathbf{\Lambda}^{-1}(\mathbf{x}-\mathbf{m}) = [\mathbf{\Lambda}^{-\frac{1}{2}}(\mathbf{t}-\mathbf{m})]^{\mathrm{T}}[\mathbf{\Lambda}^{-\frac{1}{2}}(\mathbf{x}-\mathbf{m})] \tag{10.97}$$

with an abundance expressed as

$$a = \frac{[\Lambda^{-\frac{1}{2}}(\mathbf{t}-\mathbf{m})]^{\mathrm{T}}[\Lambda^{-\frac{1}{2}}(\mathbf{x}-\mathbf{m})]}{(\mathbf{t}-\mathbf{m})^{\mathrm{T}}\Lambda^{-1}(\mathbf{t}-\mathbf{m})} = \frac{\left[\Lambda^{-\frac{1}{2}}(\mathbf{t}-\mathbf{m})\right]^{T}\left[\Lambda^{-\frac{1}{2}}(\mathbf{x}-\mathbf{m})\right]}{\alpha} \quad (10.98)$$

where **m** is the scene (background) mean, **x** the pixel spectrum of interest, **Λ** is the spectral covariance matrix of the scene (background), **t** is the target mean, and all

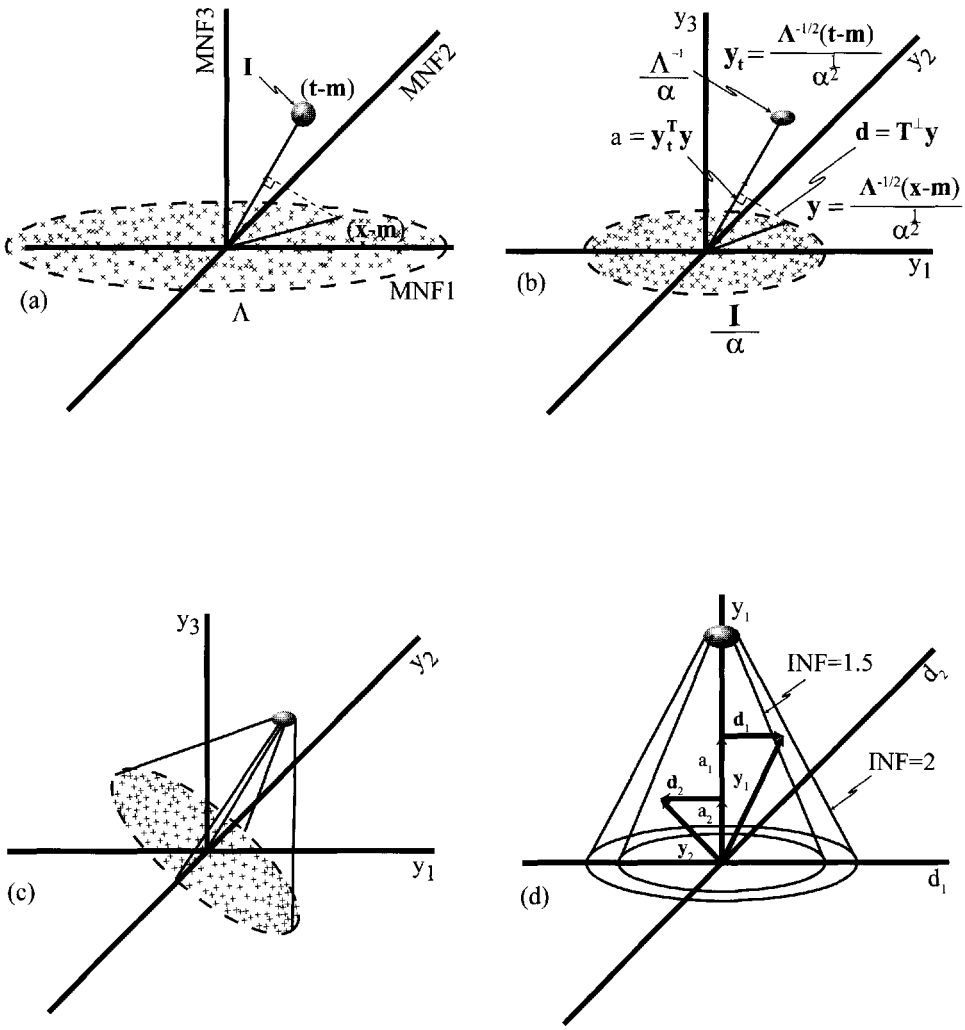


Figure 10.24 Illustration of the mixture tuned matched filter (MTMF) concept: (a) shows a simplified two-band MNF space with the shape of the target and background covariances labeled, (b) shows the data transformed into **y** space including projection of a sample pixel onto the target vector to yield the abundance *a* and the orthogonal projection to yield d, (c) shows the data transformed into **d** space with the nulling operator $\mathbf{T}^{\perp}$ (note the target has been offset for visualization purposes), and (d) shows two sample pixels $\mathbf{y}_1$ and $\mathbf{y}_2$ with different abundances but the same magnitudes for the **d** vectors. However, the infeasibility that is proportional to the number of standard deviations will be less for pixel 2 than for pixel 1.

the values are in MNF space, where we recall from Section 10.2.3 that $\mathbf{\Lambda}$ will be a diagonal matrix of eigenvalues that are equal to the variances in the MNF bands. This process can be thought of as transforming the data into a normalized space described by

$$\mathbf{y} = \frac{\mathbf{\Lambda}^{-\frac{1}{2}}(\mathbf{x}-\mathbf{m})}{\left[(\mathbf{t}-\mathbf{m})^{\mathrm{T}}\mathbf{\Lambda}^{-1}(\mathbf{t}-\mathbf{m})\right]^{\frac{1}{2}}} = \frac{\mathbf{\Lambda}^{-\frac{1}{2}}(\mathbf{x}-\mathbf{m})}{\alpha^{\frac{1}{2}}} \tag{10.99}$$

This normalized space is demeaned, normalized by the expected variance in each band, and further normalized so that the matched filter operated on the target yields a value of one. The target mean vector in this space can be expressed as

$$\mathbf{y}_t = \frac{\mathbf{\Lambda}^{-\frac{1}{2}}(\mathbf{t}-\mathbf{m})}{\left[(\mathbf{t}-\mathbf{m})\mathbf{\Lambda}^{-1}(\mathbf{t}-\mathbf{m})\right]^{\frac{1}{2}}} = \frac{\mathbf{\Lambda}^{-\frac{1}{2}}(\mathbf{t}-\mathbf{m})}{\alpha^{\frac{1}{2}}} \tag{10.100}$$

Thus, the abundance can be expressed as

$$a = [\mathbf{y}_t^{\mathrm{T}}\mathbf{y}] \tag{10.101}$$

representing the normalized matched filter in the transform space [cf. Fig. 10.24(b)]. If a represents how "target-like" a vector is, then projection of the vector $\mathbf{y}$ onto the hyperplane perpendicular to the target should represent how "untarget-like" or how much like the background a vector is [cf. Fig. 10.24(c)]. The nulling operator that projects the target vector $\mathbf{y}_t$ onto the subspace orthogonally to the target can be expressed as

$$\mathbf{T}^{\perp} = [\mathbf{I} - \mathbf{y}_t\mathbf{y}_t^{\#}] \tag{10.102}$$

where $\mathbf{y}_t^{\#}$ is the pseudo-inverse of $\mathbf{y}_t$. Operating on a $\mathbf{y}$ vector with $\mathbf{T}^{\perp}$ yields the vector $\mathbf{d}$ whose magnitude is the component of $\mathbf{y}$ projected onto the subspace orthogonal to $\mathbf{y}_t$ according to

$$\mathbf{d} = \mathbf{T}^{\perp}\mathbf{y} \tag{10.103}$$

$$d = \|\mathbf{d}\| = \left[[\mathbf{T}^{\perp}\mathbf{y}]^{\mathrm{T}}[\mathbf{T}^{\perp}\mathbf{y}]\right]^{\frac{1}{2}} \tag{10.104}$$

which from the Pythagorian theorem is

$$d = \left[\mathbf{y}^{\mathrm{T}}\mathbf{y} - a^2\right]^{\frac{1}{2}} \tag{10.105}$$

For a pixel containing some fractional abundance of target, we can think of $\mathbf{d}$ as the non-target-like behavior vector. This is illustrated in Figure 10.24, where we see that high-abundance targets would be expected to have $\mathbf{d}$ vectors with small magnitudes, and conversely, low-abundance targets with a large background fraction might be expected to have $\mathbf{d}$ vectors with large magnitudes based on the relative variability of targets and backgrounds in the $\mathbf{y}$ space or more rigorously the $\mathbf{d}$ space, where we will refer to the subspace formed by the projection of $\mathbf{y}$ vectors

onto the subspace perpendicular to $\mathbf{y}_t$ as **d** space [cf. Fig. 10.24(d)]. If we take all the target samples in **y** space, including the variability due to noise, and project them into **d** space, they would have a covariance $\mathbf{S}_{dt}$. Similarly, we can express the covariance of the background **y** vectors when projected into **d** space as $\mathbf{S}_{db}$. Using Eq. (10.95) to compute the expected covariance $\mathbf{S}_d(a)$ in **d** space for pixels with fractional abundance a, we obtain

$$\mathbf{S}_d(a) = a^2\mathbf{S}_d(1) + (1-a)^2\mathbf{S}_d(0) = a^2\mathbf{S}_{dt} + (1-a)^2\mathbf{S}_{db} \tag{10.106}$$

Boardman (1998) introduces the statistical distance a pixel is from the target (origin) in this nontarget space **d** as a measure of the *infeasibility* (*INF*) of the pixel exhibiting target-like behavior, which can be expressed as

$$INF(a) = \left[\mathbf{d}^T\mathbf{S}_d^{-1}(a)\mathbf{d}\right]^{\frac{1}{2}} \tag{10.107}$$

The infeasibility can be thought of as how many statistical steps a pixel is from target-like behavior, including the expected variability due to mixing of target and background variability. Figure 10.24(d) introduces the concept of infeasibility cones and shows how two pixels with similar **d** values will generate different infeasibilities depending on their abundance.

Boardman (1998) applies the MTMF to the subpixel target detection problem and shows that it can be used to effectively reduce the number of false positives by rejecting pixels with high infeasibility values (cf. Fig. 10.25). This is shown graphically in Figure 10.25(d) which illustrates a decision space in which threshold abundance (matched filter score) and infeasibility are jointly applied to isolate high-abundance and low-infeasibility pixels as targets.

In order to implement the infeasibility metric for use in the MTMF approach, we need to compute the covariance of the target and background in **d** space. Let's begin with the target. If we assume that the target vector is known (e.g., from a reflectance library), but its variability is otherwise unknown. Then we can assume it will at least have sensor noise introducing variability in the form of the noise covariance estimate ($\mathbf{S}_n$) (cf. Sec. 10.2.2). In the MNF space, we have decorrelated and whitened the noise so we would expect the target covariance to behave in a fashion similar to the noise with a covariance described by the identity matrix (**I**). Recall from Section 10.2.3 that in the second step of the MNF transform, we operate on the data with a principal component transform of the noise-whitened image data. Such a transform will potentially change the target covariance according to

$$\mathbf{S}_{xt} = \mathbf{E}_{\Sigma w}^{\mathrm{T}}\mathbf{I}\mathbf{E}_{\Sigma w} = \mathbf{I} \tag{10.108}$$

where $\mathbf{S}_{xt}$ is the expected covariance of the target in MNF space characterized by image vectors **x**. (Note that in Sec. 10.2.3 we used the symbol **y** to designate the image values in MNF space where now we use an **x**.) $\mathbf{E}_{\Sigma w}$ is the matrix made up of eigenvectors of the noise-whitened image covariance (Σ_w) as columns, and **I** is the estimate of the covariance of the target covariance in the noise-whitened space. However, since the vectors making of $\mathbf{E}_{\Sigma w}$ form an orthonormal set, we see that the target (noise) covariance in the MNF space remains the identity matrix. Also recall from Section 10.2.3 that the image (background) data in the MNF space have been

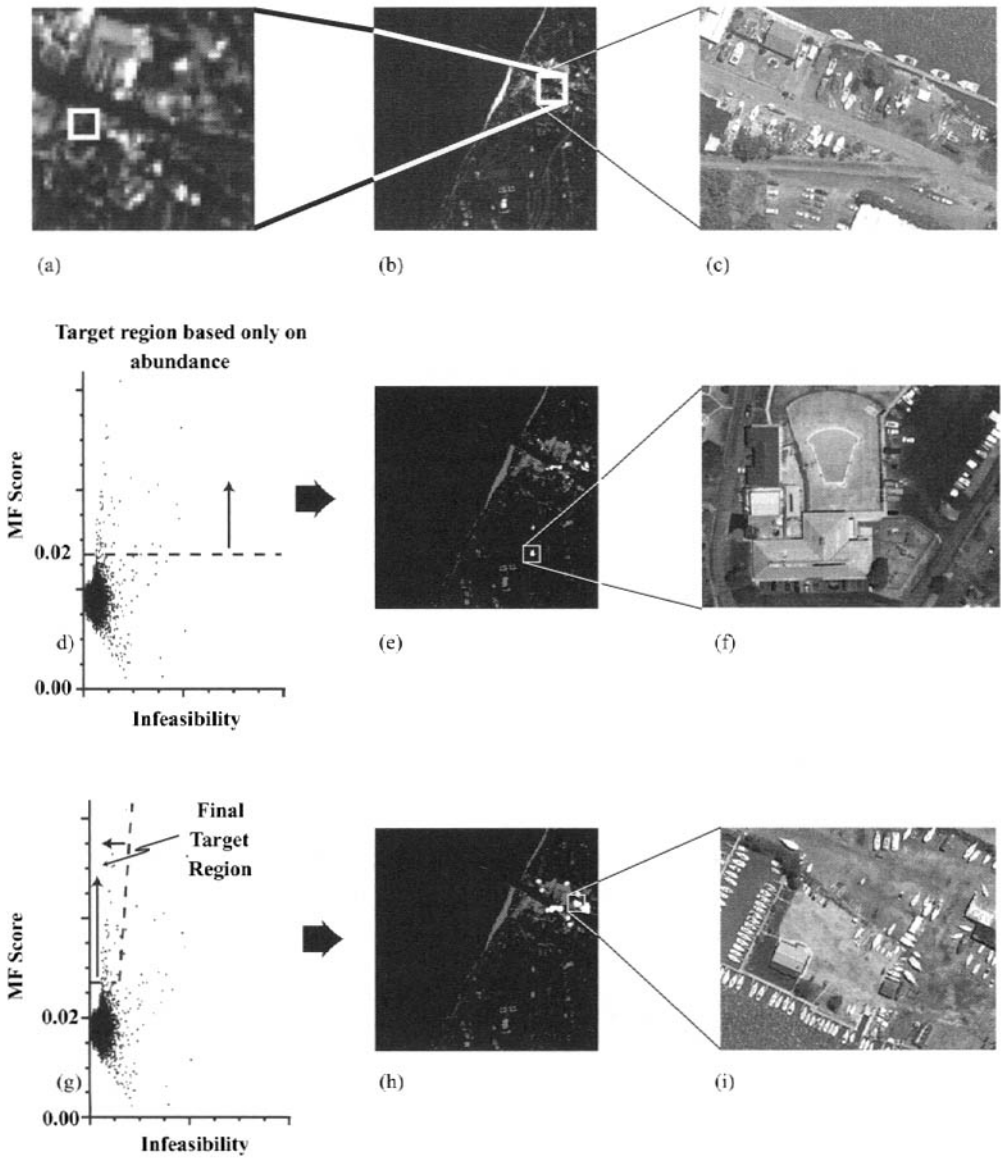


Figure 10.25 Illustration of an application of an MTMF to the detection of blue storm tarps in an AVIRIS image cube of a cluttered scene: (b) a single band of the AVIRIS image with a zoom in (a) and a high-resolution air photo for "truth" in (c). The matched filter results for a target pixel selected from (a) are shown in (e) along with an air photo of false alarms (f) based on matched filter scores (d). (h) Shows the MTMF results for the region shown in (g). An air photo of the selected targets shows the presence of storm tarps. See color plate 10.25.

orthogonalized by the PC transform to yield a diagonalized covariance that can be expressed as

$$\mathbf{S}_{xb} = \mathbf{E}_{\Sigma w}^{T}\boldsymbol{\Sigma}_{w}\mathbf{E}_{\Sigma w} = \mathbf{E}_{\Sigma w}^{T}\left[\left(\mathbf{E}_{N}\boldsymbol{\Lambda}_{N}^{-\frac{1}{2}}\right)^{\mathrm{T}}\mathbf{S}_{r}\left(\mathbf{E}_{N}\boldsymbol{\Lambda}_{N}^{-\frac{1}{2}}\right)\right]\mathbf{E}_{\Sigma w} = \boldsymbol{\Lambda} \tag{10.109}$$

where $\mathbf{E}_N$ is the matrix made up of eigenvectors of the noise covariance matrix as columns, $\boldsymbol{\Lambda}_N$ is the diagonal matrix of eigenvalues of the noise covariance matrix, $\mathbf{S}_r$ is the image covariance in the original space (e.g., reflectance), and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues of the noise-whitened image covariance matrix (Σ_w).

Transforming the data from MNF space (**x**) to the (**y**) space illustrated in Figure 10.24 and described by Eq. (10.99) will result in target and background covariances expressed as

$$\mathbf{S}_{yt} = \left(\frac{\boldsymbol{\Lambda}^{-\frac{1}{2}}}{\alpha^{\frac{1}{2}}}\right)^{\mathrm{T}}(\mathbf{S}_{xt})\left(\frac{\boldsymbol{\Lambda}^{-\frac{1}{2}}}{\alpha^{\frac{1}{2}}}\right) = \left(\frac{\boldsymbol{\Lambda}^{-\frac{1}{2}}}{\alpha^{\frac{1}{2}}}\right)^{\mathrm{T}}\mathbf{I}\left(\frac{\boldsymbol{\Lambda}^{-\frac{1}{2}}}{\alpha^{\frac{1}{2}}}\right) = \frac{\boldsymbol{\Lambda}^{-1}}{\alpha} \tag{10.110}$$

and

$$\mathbf{S}_{yb} = \left(\frac{\boldsymbol{\Lambda}^{-\frac{1}{2}}}{\alpha^{\frac{1}{2}}}\right)^{\mathrm{T}}\mathbf{S}_{xb}\left(\frac{\boldsymbol{\Lambda}^{-\frac{1}{2}}}{\alpha^{\frac{1}{2}}}\right) = \left(\frac{\boldsymbol{\Lambda}^{-\frac{1}{2}}}{\alpha^{\frac{1}{2}}}\right)^{\mathrm{T}}\boldsymbol{\Lambda}\left(\frac{\boldsymbol{\Lambda}^{-\frac{1}{2}}}{\alpha^{\frac{1}{2}}}\right) = \frac{\mathbf{I}}{\alpha} \tag{10.111}$$

respectively. Finally, projection of the **y** vectors using the nulling operator $\mathbf{T}^{\perp}$ into the **d** space according to Eq. (10.103) will result in the required covariance matrices for the target and background expressed as

$$\mathbf{S}_{dt} = (\mathbf{T}^{\perp})^{\mathrm{T}}\mathbf{S}_{yt}(\mathbf{T}^{\perp}) = \frac{\left(\mathbf{T}^{\perp}\right)^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\left(\mathbf{T}^{\perp}\right)}{\alpha} \tag{10.112}$$

and

$$\mathbf{S}_{db} = (\mathbf{T}^{\perp})^{\mathrm{T}}\mathbf{S}_{yb}(\mathbf{T}^{\perp}) = (\mathbf{T}^{\perp})^{\mathrm{T}}\frac{\mathbf{I}}{\alpha}(\mathbf{T}^{\perp}) = \frac{\left(\mathbf{T}^{\perp}\right)^{\mathrm{T}}\left(\mathbf{T}^{\perp}\right)}{\alpha} \tag{10.113}$$

respectively. These are the covariance matrices needed for the MTMF calculation of infeasibility expressed in Eq. (10.106) and (10.107) and illustrated in Figure 10.25.

In closing this chapter, we should recall that most of the algorithms described here can be applied in a variety of spaces, including digital count, radiance, reflectance, or reduced-dimensional transform spaces derived from any of these. The important thing to recognize is that the target and background descriptions applied in each case must be known or transformed into a common space where the algorithm is applied. It is also useful to recognize that many slight variations or hybridizations of these algorithms have been or could be developed, so users should make sure they understand the subtleties of the algorithms they are using. This is particularly important because many of these algorithms are optimal given

a certain set of assumptions. In the common case where our data deviate from the assumed behavior, an alternate or modified algorithm may perform better. This is the bane of students (i.e., there are lots more algorithms out there to learn) and the boon of government contractors (i.e., there are lots more algorithm development contracts to sell).

10.7 REFERENCES

Beaven, S., Hoff, L.E., & Winter, E.M. (1999). Comparison of SEM and linear unmixing approaches for classification of spectral data. *Proceedings SPIE*, Vol. 3753, pp. 300-307.

Boardman, J.W. (1998) "Leveraging the high dimensionality of AVIRIS data for improved sub-pixel target unmixing and rejection of false positives: mixture tuned matched filtering." Summaries of the Seventh Annual JPL Airborne Geoscience Workshop, Pasadena, CA.

Boardman, J.W., Druse, F.A., & Green, R.O. (1995). "Mapping target signatures via partial unmixing of AVIRIS data." Fifth Annual JPL Airborne Earth Science Workshop, Vol. 1, AVIRIS Workshop, pp. 23-26.

Clark, R.N., Sutley, S.J., Dalton, J.B., McDougal, R.R., Gent, C.A., Swayze, G.A., Livo, K.E., & Kokaly, R.F. (2003). Imaging spectroscopy: earth and planetary remote sensing with the USGS Tetracorder and expert systems. *J. Geophysical Research*, Vol. 108, No. E12, pp. 5-1 – 5-44.

Conte, E., Lops, M., & Ricci, G. (1995). Asymptotically optimum radar detection in compound-Gaussian clutter. *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 31(2), pp. 617-625.

Craig, M.D. (1994). Minimum-volume transforms for remotely sensed data. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 32, pp. 542-552.

Draper, N.R., & Smith, H. (1981). *Applied Regression Analysis*. 2nd ed., John Wiley & Sons, NY.

Farrand, W.H., & Harsanyi, J.C. (1997). Mapping the distribution of mine tailings in the Coeur d'Alene River Valley, Idaho, through the use of a constrained energy minimization technique. *Remote Sensing of Environment,* Vol. 59, No. 1, pp.64-76.

Funk, C.C., Theiler, J., Roberts, D.A., & Borel, C.C. (2001). Clustering to improve matched filter detection of weak gas plumes in hyperspectral thermal imagery. *IEEE Transactions on Geoscience and Remote Sensing,* Vol. 39, Issue 7, pp. 1410-1420.

Gao, B.-C. (1996). NDWI — a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, Vol. 58, No. 3, pp. 257-267.

Green, A.A., Berman, M., Switzer, P., & Craig, M.D. (1988). A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 26, No. 1, pp. 65-74.

Gross, H.N., & Schott, J.R. (1996). Application of spatial resolution enhancement and spectral mixture analysis to hyperspectral images. *Proceedings SPIE*, Vol. 2821-4, pp. 30-41.

Gross, H.N., & Schott, J.R. (1998). Application of spectral mixture analysis and image fusion techniques for image sharpening. *Remote Sensing of Environment*, Vol. 63, No. 2, pp. 85-94.

Harsanyi, J.C., & Chang, C. (1994). Hyperspectral image classification and dimensionality reduction: an orthogonal subspace projection approach. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 32, No. 4, pp. 779-785.

Hunt, B.R., & Cannon, T.M. (1976). Nonstationary assumptions of Gaussian models of images. *IEEE Transactions Systems, Man and Cybernetics*, Vol. SMC-6, No.12, pp. 876-882.

Ifarraguerri, A., & Chang, C.I. (1999). Multispectral and hyperspectral image analysis with convex cones. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 37, pp. 756-770.

Jimenez, L.O. & Landgrebe, D.A. (1998). Supervised classification in high dimensional space: geometrical, statistical and asymptotical properties of multivariate data. *IEEE Transactions on Systems, Man and Cybernetics, Part C,* Vol. 28, No. 1, pp. 39-54.

Kelly, E.J. (1986). An adaptive detection algorithm. *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 22, pp. 115-127.

Keshava, N., Kerekes, J.P., Manolakis, D.G., & Shaw, G.A. (2000). Algorithm taxonomy for hyperspectral unmixing. *Proceedings SPIE*, Vol. 4049, pp. 42-63.

Konno, D. (1999). "Development and testing of improved spectral unmixing techniques." M.S. Thesis, Center for Imaging Science, Rochester Institute of Technology.

Kruse, F.A., Lefkoff, A.B., Boardman, J.W., Heidebrecht, K.B., Shapiro, A.T., Barloon, P.J., & Goetz, A.F.H. (1993). The spectral image processing system (SIPS)-interactive visualization and analysis of imaging spectrometer data. *Remote Sensing of Environment*, Vol. 44, pp. 145-163.

Lee, J.B., Woodyatt, A.S., & Berman, M. (1990). Enhancement of high spectral resolution remote-sensing data by a noise-adjusted principal components transform. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 28, No. 1, pp. 255-304.

Lee, K. (2003). "A subpixel scale target detection algorithm for hyperspectral imagery." Ph.D. Dissertation, Center for Imaging Science, Rochester Institute of Technology.

Manolakis, D.G., Shaw, G.A., & Keshava, N. (2000). Comparative analysis of hyperspectral adaptive matched filter detectors. *Proceedings SPIE*, Vol. 4049, pp. 2-17.

Manolakis, D.G., Siracusa, C., Marden, D., & Shaw, G.A. (2001). Hyperspectral adaptive matched-filter detectors: practical performance comparison. *Proceedings SPIE*, Vol. 4381, pp. 18-33.

Margalit, A., Reed, I.S., & Gagliardi, R.M. (1985). Adaptive optical target detection using correlated images. *IEEE Transactions Aerospace Electronic Systems*, Vol. AES-21, No. 3, pp. 394-405.

Masson, P., & Pieczynski, W. (1993). SEM algorithm on unsupervised statistical seg-

mentation of satellite image. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 31, No. 3, pp. 618-633.

Newland, D. (1999). "Evaluation of stepwise spectral unmixing with HYDICE data." B.S. Thesis, Center for Imaging Science, Rochester Institute of Technology.

Reed, I.S., & Yu, X. (1990). CFAR detection of the optical pattern with unknown spectral distribution. *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 38, No. 10, pp. 1760-1770.

Roberts, D.A., Smith, M.O., & Adams, J.B. (1993). Green vegetation, non-photosynthetic vegetation and soils in AVIRIS data. *Remote Sensing of Environment*, Vol. 44, pp. 255-269.

Roberts, D.A., Gardner, M., Church, R., Ustin, S., Scheer, G., & Green, R.O. (1998). Mapping chapparal in the Santa Monica mountains using multiple end member spectral mixture models. *Remote Sensing of Environment*, Vol.65, pp. 267-279.

Robey, F.C., Fuhrmann, D.R., Kelly, E.J., & Nitzberg, R. (1992) A CFAR adaptive matched filter detector. *IEEE Transactions on Aerospace and Electronic systems*, Vol. 28, pp. 208-218.

Scharf, L.L., & McWhorter, L.T. (1996). "Adaptive matched subspace detectors and adaptive coherence." Proceedings 30th Asilomar Conference on Signals and Systems, pp. 114-117.

Schaum, A.P. (2001). Spectral subspace matched filtering. *Proceedings SPIE*, Vol. 4381, pp. 1-17.

Schaum, A.P., & Stocker, A.D. (2002). Joint hyperspectral subspace detection derived from a Bayesian likelihood ratio. *Proceedings SPIE*, Vol. 4725, pp. 225-233.

Schott, J.R., Lee, K., Raqueño, R., Hoffmann, G., & Healey, G. (2003). "A subpixel target detection technique based on the invariance approach." Proceedings of the AVIRIS Workshop, NASA, JPL.

Shen, S. (2002). Information-theory-based band selection and utility evaluation for reflective spectral systems. *Proceedings SPIE*, Vol. 4725, pp. 18-29.

Slater, D., & Healey, G. (1998). Analyzing the spectral dimensionality of outdoor visible and near-infrared illumination functions. *Journal of the Optical Society of America A*, Vol. 15, No. 11, pp. 2913-2920.

Smith, M.O., Ustin, S.L., Adams, J.B., & Gillespie, A.R. (1990a). Vegetation in Deserts: I. a regional measure of abundance for multispectral images. *Remote Sensing of Environment*, Vol. 31, pp. 1-26.

Smith, M.O., Ustin, S.L., Adams, J.G., & Gillespie, A.R. (1990b). Vegetation in deserts: II. Environmental influences on regional abundance. *Remote Sensing of Environment*, Vol. 31, pp. 27-52.

Stocker, A.D. & Schaum, A.P. (1997). Application of stochastic mixing models to hyperspectral detection problems. *Proceedings SPIE*, Vol. 3071, pp. 47-60.

Stocker, A.D., Reed, I.S., & Yu, X. (1990). Multidimensional signal processing for electro-optical target detection. *Proceedings SPIE*, Vol. 1305, pp. 218-231.

West, J.E., Messinger, D.W., Ientilucci, E.M., Kerekes, J.P., Schott, J.R. (2005). Matched filter stochastic background characterization for hyperspectral target detection. *Proceedings SPIE*, Vol. 5806, pp. 1-12.

Winter, M.F. (1999). N-FINDR: an algorithm for fast autonomous spectral end-member determination in hyperspectral data. *Proceedings SPIE*, Vol. 3753, pp. 266-275.

This page intentionally left blank

CHAPTER 11
USE OF PHYSICS-BASED MODELS TO SUPPORT SPECTRAL IMAGE ANALYSIS ALGORITHMS

Most of the algorithms described in the previous chapters fall into the general category of data-driven algorithms. By this we mean that the parameters used in the algorithms are largely extracted from the image data themselves. This has the advantage that little or no ancillary data are required. However, data-driven approaches can often be enhanced by taking advantage of knowledge we may have about the expected behavior of the spatial or spectral properties of the world. This knowledge is often described by physical models of the target, background, or image formation process. Hence, we will collectively refer to these often disparate algorithms as physics-based as opposed to data-driven algorithms. However, the user should recognize that this is a gray boundary since nearly all physics-based models have data-driven components and many predominately data-driven approaches include some ancillary inputs (e.g., target reflectance) that may be physics based.

The use of physics-based models is particularly applicable to spectral data analysis. This is because in many cases the analytical process involves, somehow, solving for, or constraining, the key parameters that drive the physical model describing a spectral phenomenon. Since most of the relevant models include several key parameters, it takes at least that many observations (i.e., "several") to isolate or solve for these physical parameters. In Chapter 10, we saw that spectral image data typically include at least a few and potentially many independent dimensions, making it a good candidate for use with physics-based models. As a result, essentially all of the algorithms included in this chapter are aimed at analysis of spectral imagery.

We have also, for convenience, included in this chapter thermal spectral image analysis algorithms because they essentially all use Planck blackbody physics as part of the analytical approach.

11.1 SPECTRAL THERMAL INFRARED ANALYSIS METHODS

This section deals with analysis of spectral thermal infrared data where the emphasis is on studying the spectral emissivity, as well as the temperature of the surface. The first step in this process is typically atmospheric compensation, which might employ any of the single- or dual-band methods described in Section 7.3. However, applying these methods on a band-by-band basis can be cumbersome and, more important, often leads to slight differences band-to-band that manifest as spectral emissivity artifacts and/or band-to-band temperature variations. To overcome this limitation, Kahle and Alley (1992) suggests adjusting the inputs to a radiative transfer (RT) code such as MODTRAN to try to reduce the variation in retrieved temperature for objects with known emissivity. For example, if water or dense vegetation are present in a scene, their spectral emissivities are well characterized, and we would expect the retrieved temperatures for a pixel to be the same in all spectral bands if the atmosphere is properly characterized. By adjusting the atmospheric temperature, water vapor, and possibly ozone profiles, improved uniformity in retrieved temperature can be achieved. The results of this process yield an imagewide or regional solution (i.e., it must be spatially computed or corrected for view angle changes, elevation changes, or significant atmospheric changes) for spectral transmission and upwelled and downwelled radiance. The transmission [$\tau(\lambda)$] and upwelled radiance ($L_{u\lambda}$) can then be used to solve for the surface-leaving radiance [$L_\lambda(0)$] according to

$$L_\lambda(0) = \frac{L_\lambda(H) - L_{u\lambda}}{\tau(\lambda)} \tag{11.1}$$

where $L_\lambda(H)$ is the observed radiance at altitude H, the spectral terms are all the effective values in the spectral band, and we have suppressed the use of the view-angle-dependent notation from Chapter 7 for clarity. The surface-leaving radiance can in turn be expressed as

$$L_\lambda(0) = \varepsilon(\lambda) L_{T\lambda} + (1 - \varepsilon(\lambda)) L_{d\lambda} \tag{11.2}$$

Equation (11.2) introduces the classic temperature emissivity problem. Assuming $L_{d\lambda}$ in all spectral bands of interest is available from MODTRAN, $L_\lambda(0)$ values are available for each of ℓ bands in my sensor and we have $\ell + 1$ unknowns corresponding to the ℓ unknown emissivities and the temperature (T), we have too many unknowns relative to our ability to write independent equations. Thus, we cannot directly solve for the unknowns using standard simultaneous equation approaches. For those pixels in a scene where we have well-known values of emissivity (typically few), we can generate ℓ values for T in each band. Since we know

there is only one valid kinetic temperature T, this leads to the efforts suggested above to reduce the variability in T.

Young et al. (2002) point out that even with a very good estimate of the atmospheric parameters by an RT code, the slight spectral mismatch between the RT code and the instrument calibration will yield significant residual artifacts in the retrieved emissivity spectra once a temperature is fixed [i.e., Eq. (11.2) is expressed in terms of emissivity and solved for in the form of ℓ independent equations in ℓ unknowns]. They suggest using a combination of an *in-scene atmospheric compensation* (*ISAC*) method with a single-band RT solution. The ISAC method first reduces the number of pixels to be included in the analysis. This is done by converting the radiance in each band to apparent temperature based on the spectral response function for each band and the Planck equation [cf. Eq. (4.68)]. It then identifies the spectral band in which the largest number of pixels has the highest apparent temperature compared to their apparent temperature in other bands. The pixels that have this property in the selected band are then segmented into a subset for further processing based on the logic that these pixels are most likely to represent a set of pixels whose emissivity is close to unity in the selected band. Furthermore, because atmospheric effects tend to reduce apparent temperature (at least when the apparent surface temperature is more than the effective temperature of the atmosphere), the selected band should also be the clearest. Thus, the selection process has isolated the blackest pixels in the clearest band. These pixels are tentatively assigned the apparent temperature at the sensor as the apparent surface temperature, and that temperature is used as the tentative estimate of the kinetic temperature for that pixel in subsequent steps. The radiance of the subsetted pixels in each band is then plotted against the effective in-band Planckian radiance for an object with the temperature estimated for that pixel. Combining Eqs. (11.1) and (11.2) yields

$$L_\lambda(H) = \hat{\tau}(\lambda)\left[\varepsilon(\lambda)L_{\hat{T}\lambda} + (1-\varepsilon(\lambda))L_{d\lambda}\right] + \hat{L}_{u\lambda} \tag{11.3}$$

where $\hat{T}$ is the tentative surface temperature and $\hat{\tau}(\lambda)$ and $\hat{L}_{u\lambda}$ are the unscaled transmission and upwelled radiance for the spectral band λ that are derived by a fit to the plotted data. We can expect the results of this plot to form a line with slope $\hat{\tau}(\lambda)$ and intercept $\hat{L}_{u\lambda}$ for pixels whose emissivity is close to unity in the band under consideration. Pixels with $\varepsilon(\lambda)$ less than unity should fall below this line. Thus, the unscaled transmission and upwelled radiance should be fit to those points that form the upper surface of a plot such as shown in Figure 11.1(a) [cf. Young et al. (2002) for the details of the fitting method]. Young et al. (2002) report that when these uncorrected atmospheric parameters are used to retrieve the surface radiance from objects of near unit emissivity (e.g., water and dense vegetation), they are nearly completely free of the spectral artifacts typical of the results using just an RT process [cf. Fig. 11.1(b)].

Thus, the unscaled transmission and upwelled radiance have effectively removed the spectral variability due to the atmosphere and incorporated any spectral artifacts due to spectral mismatch between instrument calibration and RT models into the unscaled values. The downside to this process is that the $\hat{T}, \hat{\tau}(\lambda)$ and $\hat{L}_{u\lambda}$

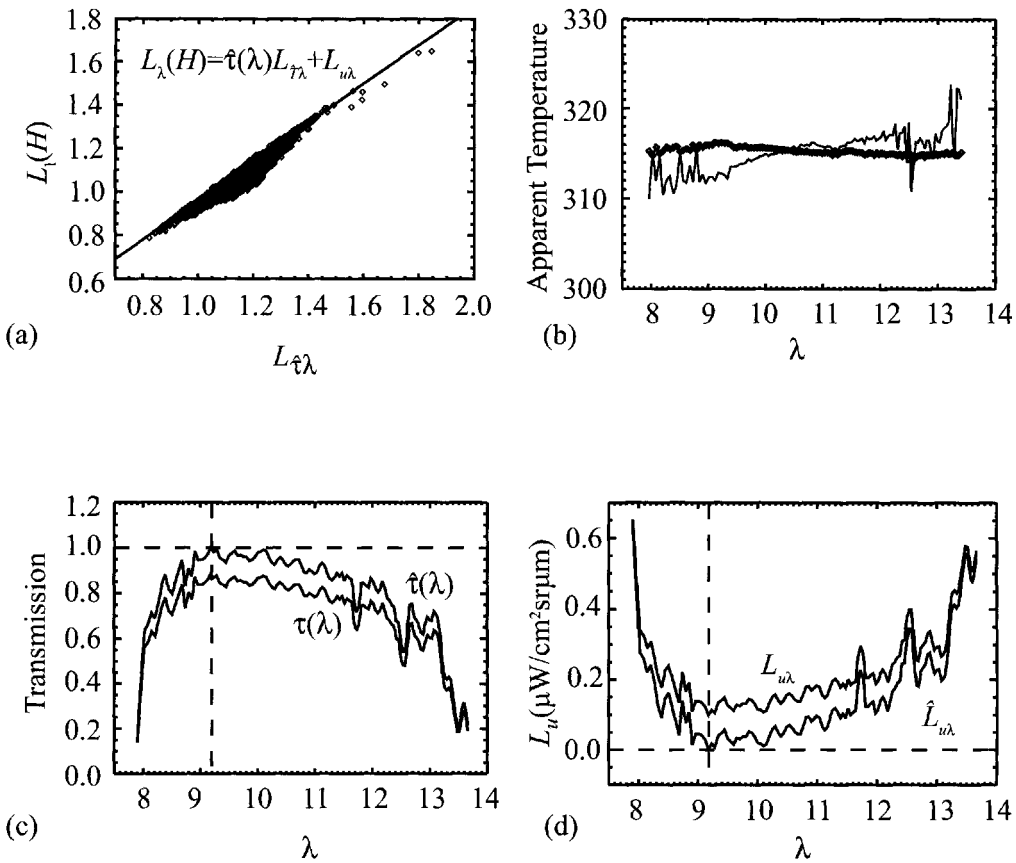


Figure 11.1 Steps in the combined ISAC-radiative transfer atmospheric compensation approach: (a) plot of observed radiance versus in-band Planckian radiance for the tentative or scaled temperature including a best linear fit through the upper surface of the data, (b) plot of surface radiance (expressed as apparent temperature) versus wavelength for data recovered independently using the ISAC scaled parameters (diamonds) and an RT code (note that Data are forced to match at λ = 10.3), (c) scaled and unscaled transmission as a function of radiance, and (d) scaled and unscaled upwelled radiance as a function of wavelength (note that the reference band is indicated).

values are all unscaled and not actual values. This can be seen most clearly if we look at the highest apparent temperature band where we have forced $\hat{\tau}(\lambda)$ to be one, $\hat{L}_{u\lambda}$ to be zero, and $L_\lambda(H)=L_\lambda(0)=L_{\hat{\tau}\lambda}$. None of this is quite true, with the result that the unscaled values are spectrally self-consistent but numerically incorrect. Young et al. (2002) suggest a method to correct the unscaled values to good numerical estimates of the true values while preserving the spectral self-consistency of the unscaled values. This is accomplished by forcing the unscaled values to match a reliable estimate of $\tau(\lambda)$ and $L_u\lambda$ in one band and then spectrally correcting all the other unscaled values to this absolute or scaled space using the requirement to maintain a constant retrieved apparent surface temperature for a blackbody radiator. In a reference band (0) where transmission $\tau(0)$ and upwelled radiance (L_{u0}) are known, we can express the radiance for a blackbody as

$$L_0(H) = L_{T_1 0}\tau(0) + L_{u0} \tag{11.4}$$

which we have already expressed in terms of unscaled parameters as

$$L_0(H) = L_{\hat{T}_1}\hat{\tau}(0) + \hat{L}_{u0} \tag{11.5}$$

Thus, we can solve for $\hat{T}_1$ as:

$$\hat{T}_1 = \text{Planck}^{-1}\left[\frac{L_{T_1 0}\tau(0) + L_{u0} - \hat{L}_{u0}}{\hat{\tau}(0)}\right] \tag{11.6}$$

where we have written Eqs. (11.4) – (11.6) in terms of an explicit temperature T_1 and unscaled temperature $\hat{T}_1$ and Planck^{-1} is the inverse Planck function (i.e., the function that inverts radiance in a bandpass to apparent temperature). By requiring $\hat{T}$ to be independent of the spectral band, we can write an expression in terms of the unknown true or expected transmission and upwelled radiance in any band (λ) and the unscaled values according to

$$\text{Planck}^{-1}\left[\frac{L_{T_1 0}\tau(0) + L_{u0} - \hat{L}_{u0}}{\hat{\tau}(0)}\right] = \hat{T}_1 = \text{Planck}^{-1}\left[\frac{L_{T_1 \lambda}\tau(\lambda) + L_{u\lambda} - \hat{L}_{u\lambda}}{\hat{\tau}(\lambda)}\right] \tag{11.7}$$

If we express

$$L_\lambda(H, T_1) = L_{\hat{T}_1 \lambda}\hat{\tau}(\lambda) + \hat{L}_{u\lambda} = L_{T_1 \lambda}\tau(\lambda) + L_{u\lambda} \tag{11.8}$$

then $L_{\hat{T}_1 \lambda}$ is the radiance a blackbody radiator would have in the bandpass represented by λ for an object with a temperature $\hat{T}_{10}$ found by analyzing the left-hand side of Eq. (11.7) (i.e., it is the unscaled temperature corresponding to the actual temperature T_1 found using reference band (0) parameters) and $L_\lambda(H,T_1)$ is the observed radiance in the band represented by λ for a blackbody with temperature T_1. If we compute $L_\lambda(H,T_1)$ and $L_\lambda(H,T_2)$ for two temperatures in each band represented by λ, as well as the blackbody radiances for those temperatures ($L_{\hat{T}_1 \lambda}, L_{\hat{T}_2 \lambda}$), we can scale Eq. (11.8) for the unknown values $\tau(\lambda)$ and $L_{u\lambda}$ in each band according to

$$\tau(\lambda) = \frac{L_\lambda(H, T_2) - L_\lambda(H, T_1)}{L_{T_2 \lambda} - L_{T_1 \lambda}} \tag{11.9}$$

and

$$L_{u\lambda} = \frac{L_{T_2 \lambda} L_\lambda(H, T_1) - L_{T_1 \lambda} L_\lambda(H, T_2)}{L_{T_2 \lambda} - L_{T_1 \lambda}} \tag{11.10}$$

In practice, T_2 and T_1 are selected to span the expected thermal range of the scene, and $\tau(0)$ and L_{u0} are typically generated using MODTRAN. Figure 11.1(c) and (d) show unscaled and expected solutions for τ and L_u as a function of wave-

length. Note that by bridging through the unscaled temperature, all the spectral character of the two unscaled values is preserved in the final values. Thus, we preserve the well-behaved spectral character of the unscaled surface-leaving radiance values while generating more accurate absolute results. Note that the maximum apparent temperature band used in the ISAC step and the reference band designated (0) in Eq. (11.7) need not be the same. It is also important to recognize that the combined ISAC process normalized using the single band RT model only yields the spectral surface-leaving radiance [Eq. (11.1)], and we still need to perform the temperature emissivity separation step. Furthermore, the $L_{d\lambda}$ term is not available from ISAC and must presumably come from an RT process. However, by referring to Table 4.1, we see that the reflected skylight term is typically quite small, and spectral artifacts in $L_{d\lambda}$ are likely to have limited impact on the retrieved emissivity spectra (at least when ε is large). Gu et al. (2000) suggest an alternative approach that also uses a combination of empirical measurements from the scene and radiative transfer theory to solve for $\tau(\lambda)$ and $L_{u\lambda}$.

Once we have computed a surface-leaving radiance using some form of atmospheric compensation, a means of separating the temperature and emissivity components must be implemented. As suggested earlier, this is an ill-posed problem, and one must either make some assumptions about the data or use some form of estimation technique to effectively reduce the number of free parameters by one. Kahle and Alley (1992) describe a method they attribute to Vincent et al. (1975) that takes advantage of the observation that many materials have high emissivity values near 11.75 μm. By fixing the emissivity for all pixels at 11.75 μm at a high constant value (e.g., 0.95), the surface temperature can be estimated by solving the surface-leaving radiance equation for $L_{T\lambda}$ and inverting the Planck equation to yield T according to

$$L_{T\lambda} = \frac{L_{11.75}(0) - L_{d11.75}}{\varepsilon(11.75)} + L_{d11.75} \tag{11.11}$$

or

$$L_{T\lambda} \approx \frac{L_{11.75}(0)}{\varepsilon(11.75)} \tag{11.12}$$

where Eq. (11.11) is used when an estimate of $L_{d\lambda}$ is available and Eq. (11.12) yields a reasonable approximation when ε is large (i.e.,$\varepsilon \approx 1$). Using the value of T found in this fashion, the surface-leaving radiance in all other bands can be inverted to spectral emissivity values. This yields a reasonable shape for emissivity spectra; however, it has the disadvantage that the effective number of bands is reduced by one (i.e., the band assigned a fixed emissivity carries no information), and the absolute emissivity and temperature values will be in error due to the error in the estimated emissivity at the reference wavelength (e.g., 11.75 μm in this example). Finally, this method can induce systematic noise into the results since any noise in the reference band will be propagated into all the spectral emissivity results. An alternate approach suggested by Kahle and Alley (1992) is to simply compute the apparent temperature [Eq. (11.11 or 11.12)] in each band assuming a high constant

emissivity in all bands (e.g., 0.98 or 1). The pixel is then assigned the maximum apparent temperature, and the emissivity in all the other bands is computed based on that temperature value. The emissivity spectra that results from this process will have the same spectral shape as those from the fixed reference band emissivity approach; however, their relative magnitude can vary pixel to pixel. For many natural surfaces, there is a high likelihood that over a wide spectral range with narrow spectral bands, the effective in-band emissivity will be close to one in at least one band. However, for low emissivity surfaces, both of these approaches can lead to sizable errors.

Borel (2003) suggests a way to combine the ISAC method with an assumption based on the relative smoothness of emissivity spectra compared to atmospheric spectra to achieve both atmospheric compensation and temperature emissivity separation (cf. Fig. 11.2). The method, referred to as *automatic retrieval of temperature and emissivity using spectral smoothness* (*ARTEMISS*), begins by estimating the transmission and upwelled radiance spectra using the ISAC approach described above. Recognizing the scaling issues associated with the ISAC values, Borel (2003) uses large numbers of $\tau(\lambda)$, $L_{u\lambda}$ and $L_{d\lambda}$ values generated by MODTRAN and convolved to match the spectral bands of the sensor data being analyzed. By running MODTRAN (cf. Chap. 7) for many atmospheres, a lookup table (LUT) can be generated that represents the range of atmospheric temperatures, column water vapor amounts, and ozone concentrations that might exist. This LUT includes $\tau(\lambda)$, $L_{d\lambda}$ and $L_{u\lambda}$ spectra for each possible atmospheric condition. Borel (2003) suggests searching the LUT by comparing ISAC-generated transmission spectra with the MODTRAN transmission spectra (e.g., by linear regression) to generate a smaller set of candidate atmospheres. Next, a final atmosphere is selected by running a smooth emissivity retrieval algorithm on a subset of pixels. This algorithm begins by assigning an emissivity of 0.95 in a wavelength region where emissivities are typically high (e.g., 11.75 μm) and solving for the apparent temperature (T) from the thermal radiance using

$$L_{T0} = \frac{L_0 - L_{u0} - (1-\varepsilon(0))L_{d0}\tau(0)}{\tau(0)\varepsilon(0)} \tag{11.13}$$

where the 0 refers to the reference spectral band (e.g., 11.75 μm). Based on this estimate of T, the emissivity in each spectral band can be computed according to

$$\varepsilon(\lambda) = \frac{L_\lambda - L_{u\lambda} - \tau(\lambda)L_{d\lambda}}{L_{T\lambda}\tau(\lambda) - L_{d\lambda}\tau(\lambda)} \tag{11.14}$$

The emissivity spectrum computed in this fashion is compared to a smoothed version of itself under the assumption that many emissivity spectra are relatively smooth over a small window. Borel (2003) suggests a local standard deviation metric ($\boldsymbol{\sigma}_\varepsilon$) based on the difference between the emissivity in each band and the average of that value and the emissivity in the adjacent spectral bands. Because we have probably selected the wrong temperature on the first iteration of this process, there will be excess spectral structure in the retrieved emissivity spectrum as indicated by high $\boldsymbol{\sigma}_\varepsilon$ values. By varying T about the initial estimate, we can find

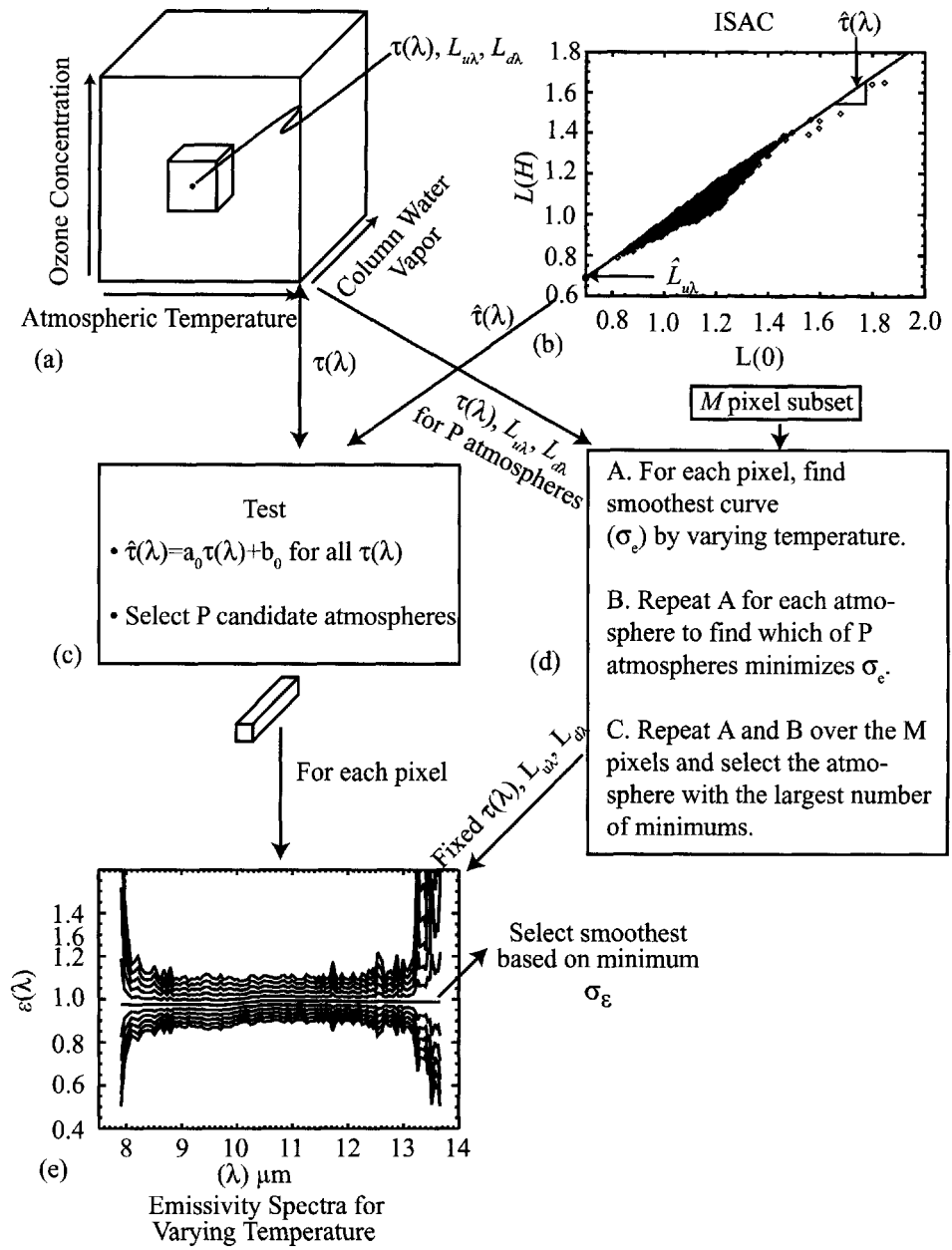


Figure 11.2 Steps in the ARTEMISS temperature-emissivity retrieval process: (a) run MODTRAN to generate a LUT of transmission, upwelled radiance, and downwelled radiance spectra for a range of possible atmospheres; (b) run ISAC to generate scaled atmospheric transmission spectrum; (c) select *P* candidate atmospheres by comparing LUT transmission spectra $\tau(\lambda)$ to the ISAC spectrum $\hat{\tau}(\lambda)$, (d) on a subset (*M*) of image pixels, iteratively adjust temperature until the smoothness metric σ_ε is minimized for each of *P* atmospheres, and then select which of the *P* atmospheres had the most votes cast by the subset of *M* pixels; and (e) use the selected atmosphere and the temperature adjustment and smoothness metric for all pixels to select the final temperatures and compute emissivity spectra.

a minimum value for σ_ε. This process can be repeated for each of the candidate atmospheres selected by the ISAC regression analysis. As each pixel is tested, the atmosphere that produces the lowest σ_ε value based on the temperature iteration process receives a vote. The atmosphere receiving the most votes on the initial data subset is then used for the entire image region being processed. All the data are then processed using the temperature iteration with smooth curve criteria to select the final temperatures and corresponding emissivity spectra.

The ARTEMISS process achieves both atmospheric correction and temperature emissivity separation using a combination of physical models and principles and data-driven approaches. However, as Borel (2003) points out, it is sensitive to noise, and multiple temperature/atmosphere combinations can yield smooth emissivity spectra. Thus, the user is cautioned that conditioning steps such as using ISAC to initialize the candidate atmospheres and testing over a range of pixel types to select the final atmosphere can reduce errors. It is also important to recognize that the algorithm as presented assumes that the same atmosphere is impacting all pixels (i.e., that view angle, elevation, and atmospheric stability are not significantly influencing the atmospheric variables within the image region being processed).

Using the techniques presented above, thermal spectral data can be converted to emissivity spectra that can be analyzed in essentially the same fashion as reflectance spectra using all of the methods presented in Chapter 10. In addition, a surface temperature estimate is available that may provide additional discrimination or target status information.

Hook et al. (1992) suggest an alternative to the temperature emissivity separation problem by defining a new spectral variable called the *alpha residual* that can be derived from the surface-leaving radiance. The alpha residual spectrum is to first order independent of temperature, and an emissivity spectrum can also be converted to an alpha residual spectrum so that library spectra can be compared to image-derived spectra in the alpha residual space. The alpha residual approach is based on the Wien approximation to the Planck blackbody equation, which is valid to within about 1% for objects at 300 K at 10 μm. If we assume the reflected downwelled term is negligible compared to the emitted term, then the surface-leaving radiance for atmospherically compensated thermal data can be approximated as

$$L_\lambda(0) = \varepsilon L_{T\lambda} \cong \varepsilon \frac{2hc^2}{\lambda^5} e^{-hc/\lambda kT} = \varepsilon \frac{C_1}{\lambda^5} e^{-C_2/\lambda T} \tag{11.15}$$

where $C_1 = 2hc^2$ and $C_2 = hc/k$. For narrow spectral bands, the spectral radiance given by the Wien approximation should be a good estimate of the mean effective spectral radiance in the bandpass. For this case, taking the log of the surface-leaving radiance and multiplying each band (i) by the mean wavelength in the band yields

$$X_i = \lambda_i ln(L_{\lambda i}(0)) = \lambda_i ln\varepsilon_i + \lambda_i lnC_1 - 5\lambda_i ln\lambda_i - \frac{C_2}{T} \tag{11.16}$$

where the last term is conveniently independent of wavelength. Taking the average of all the X_i values for a given pixel and subtracting it from the individual X_i band values removes the temperature effects yielding the alpha residual spectral values given by

$$\alpha_i = X_i - \bar{X} = \lambda_i ln\varepsilon_i + \lambda_i lnC_1 - 5\lambda_i ln\lambda_i - \frac{1}{N}\sum_{i=1}^{N}\lambda_i ln\varepsilon_i - \frac{1}{N}\sum_{i=1}^{N}\lambda_i lnC_1 + \frac{1}{N}\sum_{i=1}^{N}5\lambda_i ln\lambda_i \quad (11.17)$$

The alpha residual spectrum is easily computed from a radiance spectrum. As seen in the right-hand side of Eq. (11.17), the alpha residual is only a function of the emissivity spectrum, the spectral band centers, and physical contents. Thus, we can compute alpha residual spectra for each pixel in an atmospherically compensated spectral image and for any emissivity spectrum of interest. To reduce errors introduced by using bandpass values and the Wien approximation, the alpha residual spectrum for a laboratory emissivity can be multiplied by the Planck function based on a crude temperature estimate and convolved with the sensor response to yield an effective in-band spectral radiance that can then be operated on just like the surface-leaving radiance to yield an alpha residual spectrum. Using this approach, an individual target spectrum, or whole libraries of spectra, can be converted to alpha residual space, and all the spectral image analysis tools presented in Chapter 10 can be applied to the alpha residual image spectra. There are some important limitations of the alpha residual approach. First, it does not yield a temperature estimate. Second, as shown in Figure 11.3, the mean emissivity level is lost. Finally by neglecting the effects of reflected downwelled radiance, the alpha residual spectra derived from the scene data for low-emissivity (high-reflectivity) targets are apt to include spectral artifacts induced by the reflected downwelled radiance that will not be in library spectra. This can potentially confuse spectral analysis algorithms. On the other hand, as shown in Figure 11.3, the alpha residuals approach is very simple to apply and preserves much of the spectral character for materials that have spectral structure in the thermal infrared.

Before leaving this general topic of temperature emissivity separation, we should note that there are so many methods available in part due to their relative ease of use and the applicability to the types of data and scenarios to be analyzed. For example, if your data match the high emissivity-assumptions, some of the simpler methods needing only a few bands may be applicable. On the other hand, if you have many spectral bands and relevant targets with low emissivities, some of the more involved methods may be required. Kaiser (1999) compared the performance of many of the methods presented here, as well as several others, including a method incorporating empirical data about the spectral shape of observed emissivity spectra (Gillespie et al. 1998). This study, based on analysis of SEBASS thermal infrared imaging spectrometer data atmospherically corrected with the ISAC algorithm, indicated that the relatively straightforward methods presented here were able to reproduce absorption features in the retrieved emissivity spectra. However, while general spectral character was generally mapped, the mean level

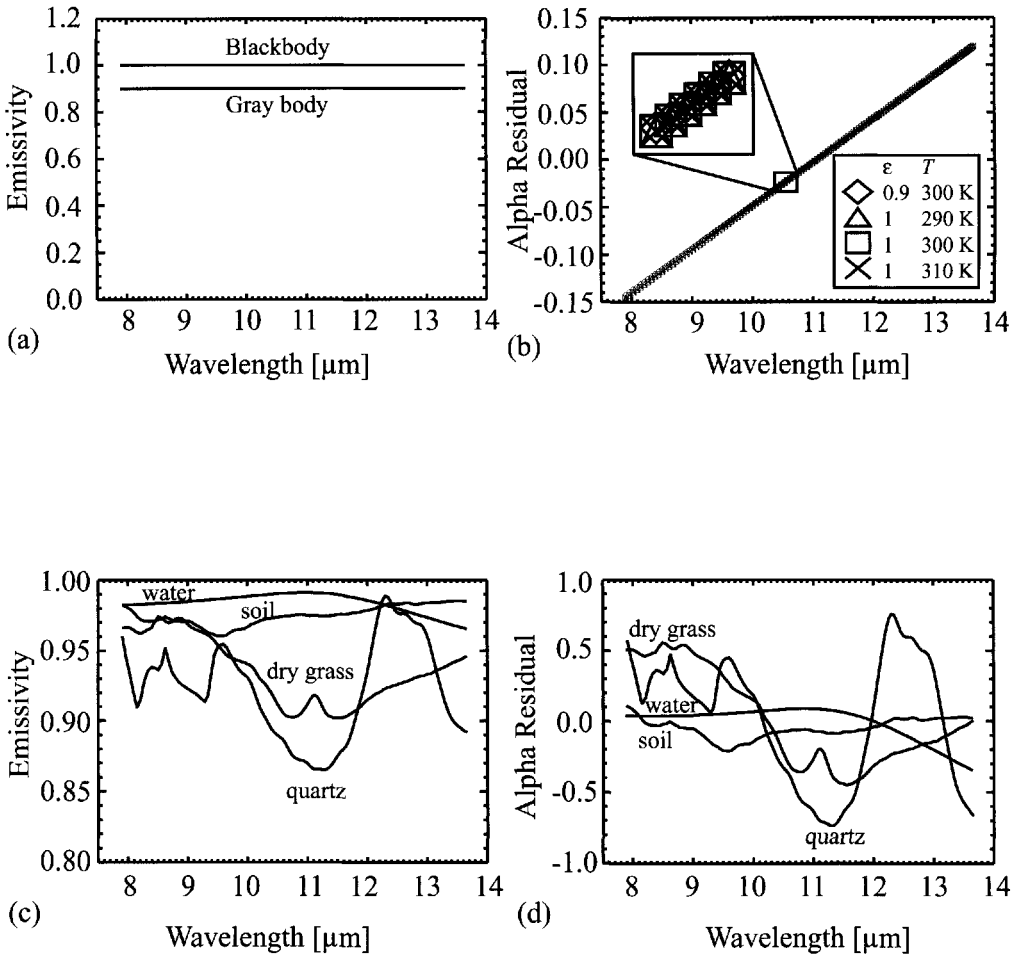
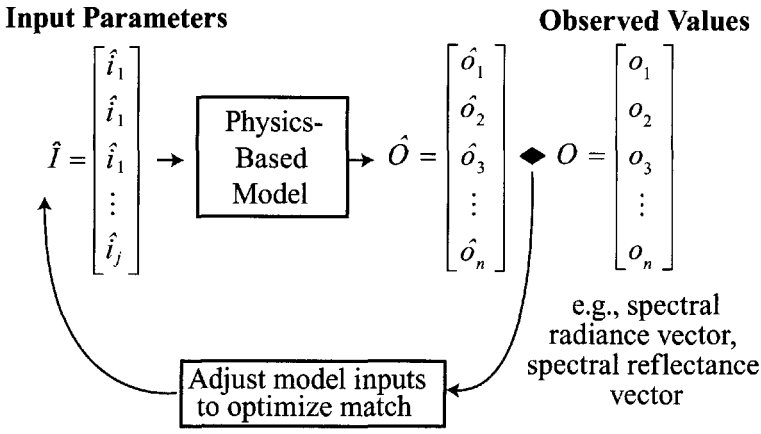


Figure 11.3 Comparison of emissivity and alpha residual spectra: (a) emissivity spectra for a blackbody and a gray body, (b) the alpha residual spectra computed from the radiance associated with a blackbody at three different temperatures and the gray body at 300 K show that thermal effects are effectively suppressed but mean emissivity information is also lost, (c) and (d) show corresponding emissivity and alpha residual spectra computed from radiance spectra for various material in the ASTER database when observed at 300 K.

was often not well represented, especially for low-emissivity materials. This result was true even for more involved algorithms.

11.2 MODEL MATCHING USING RADIATIVE TRANSFER MODELS

One of the earliest and most widely adopted of the physics-based algorithmic approaches involves the use of radiative transfer (RT) models to predict the observed spectrum using some set of input parameters that are based on initial assumptions about the pixel of interest. The predicted spectrum is compared either over a local spectral range or in total to the observed spectrum. The inputs are then adjusted



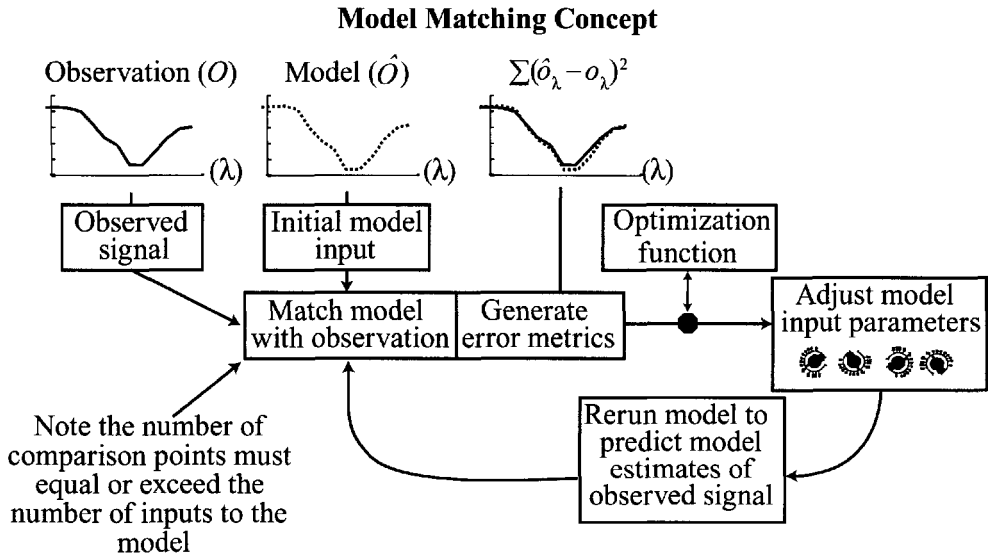


Figure 11.4 Illustration of spectral model matching concept.

until a match is declared and the inputs to the model that achieve the best match are asserted to be valid for that pixel (cf. Fig. 11.4).

11.2.1 Model Matching Applied to Atmospheric Compensation

The *nonlinear least squared spectral fit* (*NLLSSF*) described by Green et al. (1993) and summarized in Section 7.6.1 is an early example of using this approach to solve for atmospheric parameters in the VNIR-SWIR spectral region. The MODTRAN RT code [cf. Berk et al. (1989)] is used along with a parametric description of spectral reflectance to predict sensor-reaching radiance in selected spectral windows. The process conceptually iterates until a best match is found.

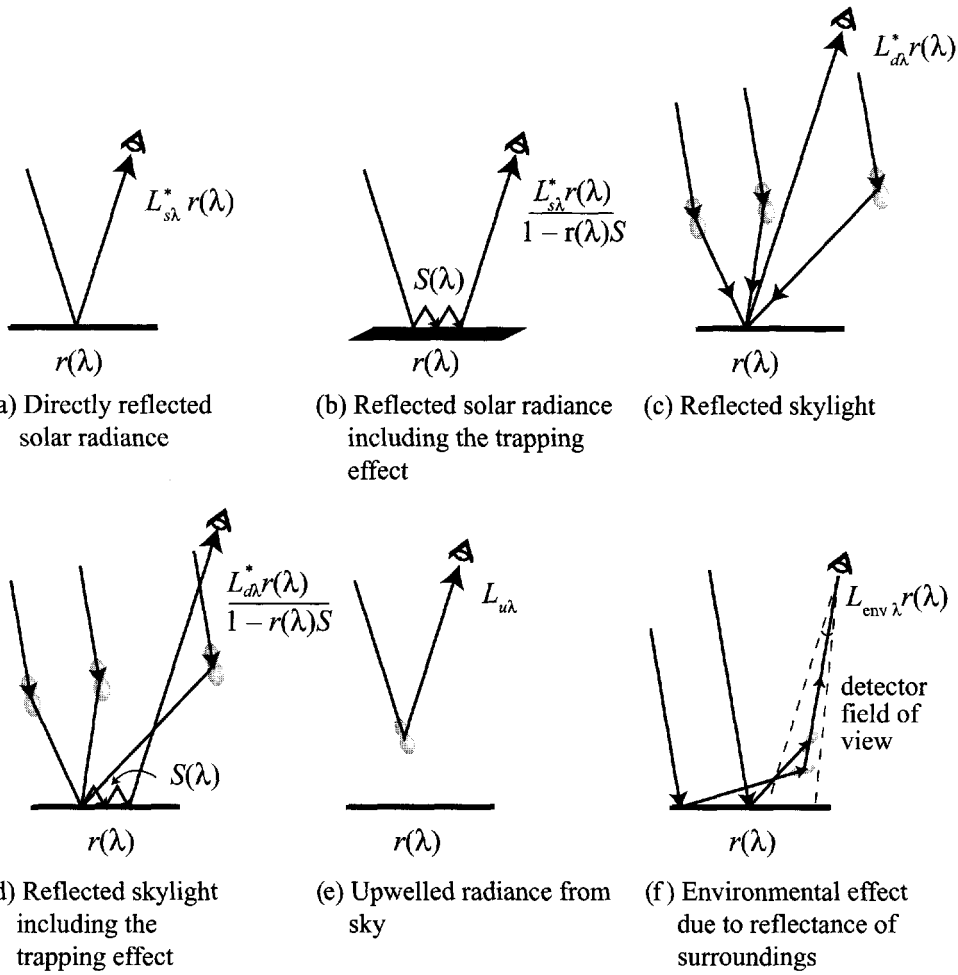
Sanders et al. (2001) describes an extension to the NLLSSF approach that further takes advantage of the MODTRAN RT code to more fully incorporate adjacency effects into the atmospheric compensation process. Note that some of these improvements, particularly the use of the atmospheric point spread function, are also captured in the FLAASH algorithm incorporated in the ENVI software package [cf. Matthew et al. (2003)]. Sanders et al. (2001) suggest that the observed spectral radiance for a target with reflectance $r(\lambda)$ having effectively infinite spatial extent can be expressed as (cf. Fig. 11.5):

$$L_\lambda = \frac{L^*_{S\lambda} r(\lambda)}{1 - r(\lambda)S(\lambda)} + \frac{L^*_{d\lambda} r(\lambda)}{1 - r(\lambda)S(\lambda)} + L_{env\lambda} r(\lambda) + L_{u\lambda} \tag{11.18}$$

where $L^*_{S\lambda}$ is the radiance reaching the sensor for a target with a reflectance factor of one for direct solar reflectance neglecting multiple scattering (trapping effects). The $[1 - r(\lambda)S(\lambda)]$ term accounts for the multiple scattering or trapping effect of the atmosphere, $L^*_{d\lambda} = L_{d\lambda}\tau_2(\lambda)$ is the downwelled radiance reaching the sensor for a target with a reflectance factor of one neglecting trapping effects, $L_{u\lambda}$ is the upwelled radiance, L_{env} is the radiance reaching the sensor reflected from the Earth surrounding the target having a reflectance factor of one that is scattered into the sensor's line of site (i.e., it is due solely to adjacency and does not include any interaction with the target pixel), and $S(\lambda)$ is the spherical scattering albedo of the atmosphere. It is important to recognize that the magnitude of $L_{env\lambda}$ is a function of the spectral illumination and the magnitude and phase of the scattering in the lower atmosphere. Thus, it will be rapidly attenuated with wavelength as the scattering terms fall off. Sanders et al. (2001) indicate that the MODTRAN RT code can be used to generate all the terms in Eq. (11.18) except for $r(\lambda)$. They suggest that a model matching approach such as NLLSSF can then be used to generate local (pixel-specific) or regional (many-pixel or whole-scene) values for the critical atmospheric parameters (i.e., pressure depth, column water vapor, and aerosol amount). At this point, Eq. (11.18) can be inverted to yield initial surface reflectance estimates. This is essentially the NLLSSF approach proposed by Green et al. (1993). However, this approach is only fully valid for cases where the reflectance of the surround is the same as the reflectance of the target. Incorporating the reflectance of the surround or background (r_b) as a distinct value leads to a revised expression for the sensor reaching radiance of

$$L_\lambda = \frac{L^*_{S\lambda} r(\lambda)}{1 - r_b(\lambda)S(\lambda)} + \frac{L^*_{d\lambda} r(\lambda)}{1 - r_b(\lambda)S(\lambda)} + L_{env} r_b(\lambda) + L_u(\lambda) \tag{11.19}$$

By studying Figure 11.5(f), it becomes clear that due to the forward scattering nature of the atmosphere, the reflectance of pixels immediately adjacent to the target pixel should have a stronger influence on the value of $r_b(\lambda)$ than pixels farther away. Sanders et al. (2001) use MODTRAN estimates of the scattering phase functions as a function of elevation and wavelength to generate an effective point spread function (PSF) for the scattering of flux by the atmosphere into the sensor line of sight (cf. Fig. 11.6). This normalized PSF is convolved with the initially



(a) Directly reflected solar radiance

(b) Reflected solar radiance including the trapping effect

(c) Reflected skylight

(d) Reflected skylight including the trapping effect

(e) Upwelled radiance from sky

(f) Environmental effect due to reflectance of surroundings

Figure 11.5 Illustration of the radiometric terms used in Eq. (11.18). Note the scattering interactions in the atmosphere are shown with a scattering phase function, and the multiple scattering by the atmosphere is indicated by reflection of the rays by the spherical albedo $S(\lambda)$ term.

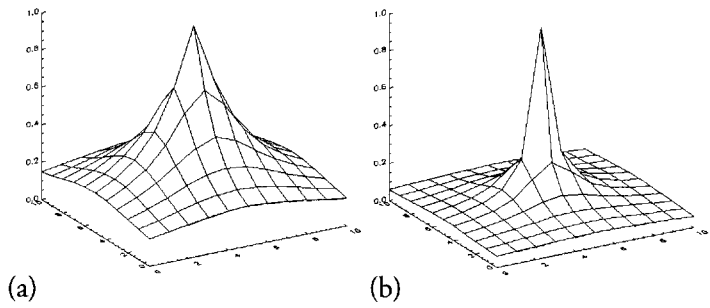
retrieved reflectance images on a band-by-band basis to compute the effective reflectance of the background. In a second pass through the image, Eq. (11.19) is solved for reflectance using the background reflectance values from the convolution process. Note that for uniform regions there will be no change in the retrieved reflectance [since $r_b(\lambda) = r(\lambda)$], and Eq. (11.19) reduces to Eq. (11.18), which was used in the first pass. However, as shown in Figure 11.6, the reflectance of small targets can be significantly changed and the retrieved reflectance improved using this two-pass approach.

11.2.2 Model Matching Applied to Water-Quality Parameter Retrieval

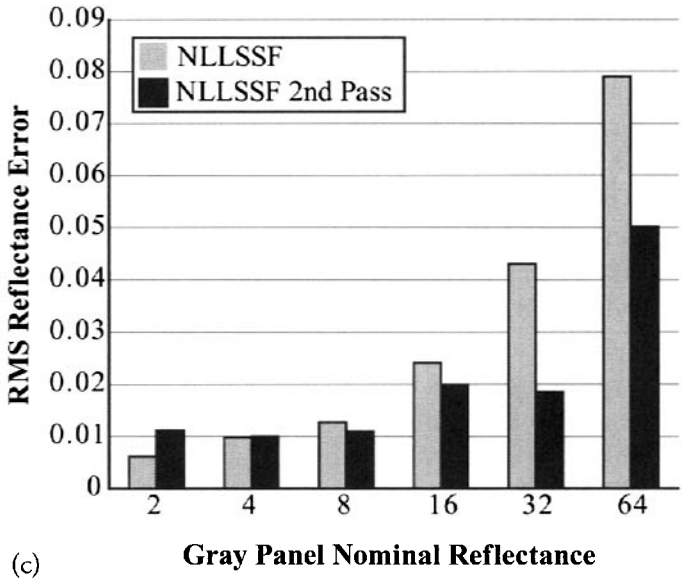
The model matching approach has also been applied to the retrieval of the concentration of water-quality parameters using VNIR imaging spectrometer data. In

this case, the radiative transfer model is an in-water RT model that predicts the reflectance of the water as a function of the illumination/viewing conditions, the water roughness, and the concentration of the three primary coloring agents in the water. The model matching approach described by Raqueño et al. (2000) uses the *Hydrolight* model described by Mobley (1994). The Hydrolight model performs RT calculations for a layered water model that incorporates the absorption and scattering properties of the water itself, as well as the multiple constituents considered including chlorophyll, total suspended material, and colored dissolved organic material. The Hydrolight RT calculations include treatment of the incident light field, the in-water absorption, and scattering and produce estimates of the radiance exiting the water, which can be expressed as the remotely sensed reflectance or the bidirectional reflectance factor (BDRF) for the water column including the air–water interface. The transmission through the air–water interface is computed as a function of the statistical distribution of the orientations of the facetized wind-roughened water surface. Based solely on the wind speed, the roughness of the surface can be described based on the observations and calculations of Cox and Munk (1956). From this roughness description and the absorption coefficient of water, the fresnel reflection equations can be used to compute the mean surface reflection and transmission (for a patch that is large compared to the local statistics).

In order to understand the inputs to the in-water RT model, we need to briefly consider the water-quality application. Pure water is a poor scatterer but scatters more at shorter wavelengths and is highly absorptive at longer wavelengths, leading to its overall low reflectance, blue color, and near zero reflectance in the near infrared. Phytoplankton in water is typically characterized by the concentration of chlorophyll (C), which is the dominant pigment in most plant algae. In practice, a number of pigments are tied to the nominal chlorophyll concentration based on the algae species present in a water body or the sample used to characterize the optical properties. The phytoplankton cells scatter broadly across the visible and near infrared, and the chlorophyll pigments absorb strongly in the blue and red, with the net result that water containing only phytoplankton will take on a greenish shade and will still have near zero reflectance in the near infrared. Organic material, as it decays, releases humic and tannic acids into the water, which are collectively characterized by the concentration of colored dissolved organic material (CDOM). CDOM, because it is dissolved in the water, is typically characterized only by its absorption properties (i.e., it is assumed to induce no incremental scattering), which are strongest in the blue, giving water with high CDOM concentrations a yellowish brown shade. Swamps and bogs have high concentrations of CDOM, leading to the very dark appearance from overhead and the yellow shade of water samples. The open ocean is usually characterized by low concentrations of phytoplankton and very low concentrations of CDOM. Coastal waters and most fresh waters may have substantially higher concentrations of these constituents, as well as significant concentrations of suspended materials (SM). This suspended material is a catchall for inorganic suspended sediments (e.g., soil particles), as well as organic material suspended in the water column. Typically, these are large particles that will settle out of the water column given time and low turbulence levels. They tend to be strong scatterers across the visible and into the near infrared with no



(a) (b)



(c)

Figure 11.6 The atmospheric PSF due to environmental scattering derived from the MODTRAN code is shown for two spectral bands of AVIRIS in terms of AVIRIS 20 m pixels corresponding to (a) the 402 nm band and (b) the 2100 nm band. Note that the PSF is much broader at shorter wavelengths. (c) The RMS error between retrieved reflectance and ground truth reflectance for six reflectance panels. The analysis used the NLLSSF approach with HYDICE data at a DOE atmospheric radiation measurement ARM site. Note that the panels were approximately gray on a grass background and were several pixels on a side. The RMS value is the average over all VNIR-SWIR spectral bands.

strong spectral shape to their absorption. As a result, they tend to raise the overall reflectance of the water, and when combined with significant CDOM levels, the water will take on a brown (muddy) shade. The overall increase in concentration of constituents in coastal waters and the presence of SM tends to raise the reflectance of coastal waters and fresh waters in the visible, as well as causing the near infrared reflectance to deviate from zero. This is particularly important for atmospheric compensation because the open ocean community tends to rely on the near zero reflectance of water in the near infrared to characterize the aerosol levels, which are the critical component of atmospheric compensation over water [cf. Gordon et

al. (1997)]. Figure 11.7 shows typical absorption spectra of the constituents used in an in-water RT model.

The absorption spectra and scattering spectra for each constituent are combined with concentration levels [ ] to generate the optical properties of each layer in the water according to:

$$a(\lambda) = a_W(\lambda) + [CDOM]a_{CDOM}(\lambda) + [SM]a_{SM}(\lambda) + [C]a_C(\lambda) \quad (11.20)$$

and

$$b(\lambda) = b_W(\lambda) + [SM]b_{SM}(\lambda) + [C]b_C(\lambda) \quad (11.21)$$

where $a(\lambda)$ and $b(\lambda)$ are the absorption and scattering cross sections, respectively. Based on the optical properties of the constituents (including angular scattering phase functions), the concentration of each constituent, the illumination conditions, and the surface optical properties as characterized by wind speed, the Hydrolight model solves for the spectral reflectance of the water column. The results of Hydrolight runs for a particular scenario are shown in Figure 11.8. By running the RT model for many combinations of the three in water constituents using optical properties appropriate for the water system under study, a three-dimensional lookup table (LUT) can be built where each cell in the LUT contains a spectral reflectance spectrum (cf. Fig. 11.9). By setting the illumination, viewing conditions, water surface conditions, and sensor spectral response for each band, the LUT can be made specific to the imaging spectrometer data to be analyzed.

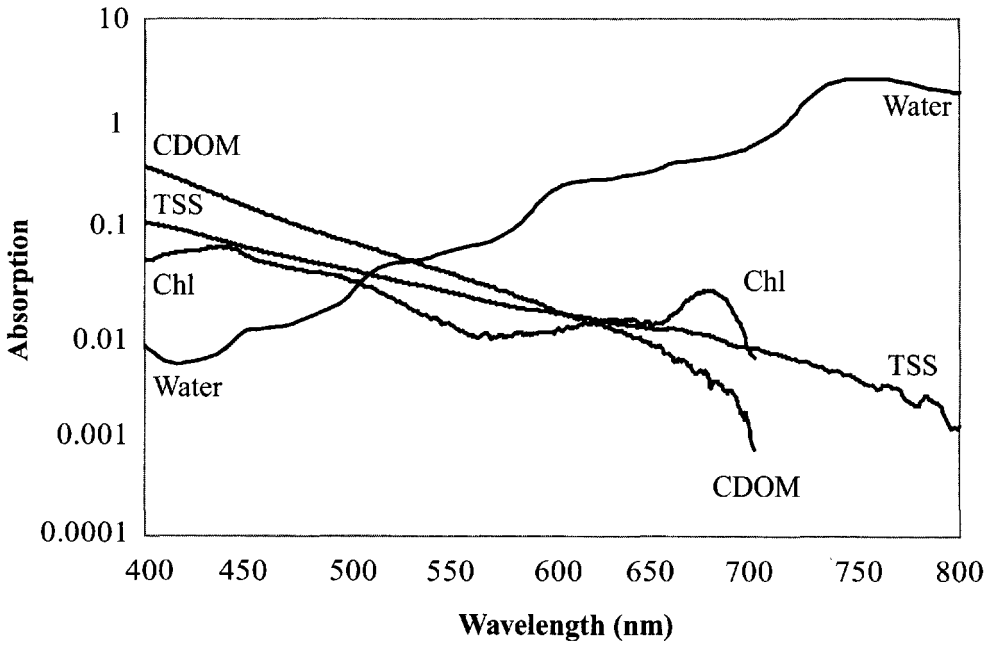


Figure 11.7 Absorption spectra of constituents used in water radiative transfer models.

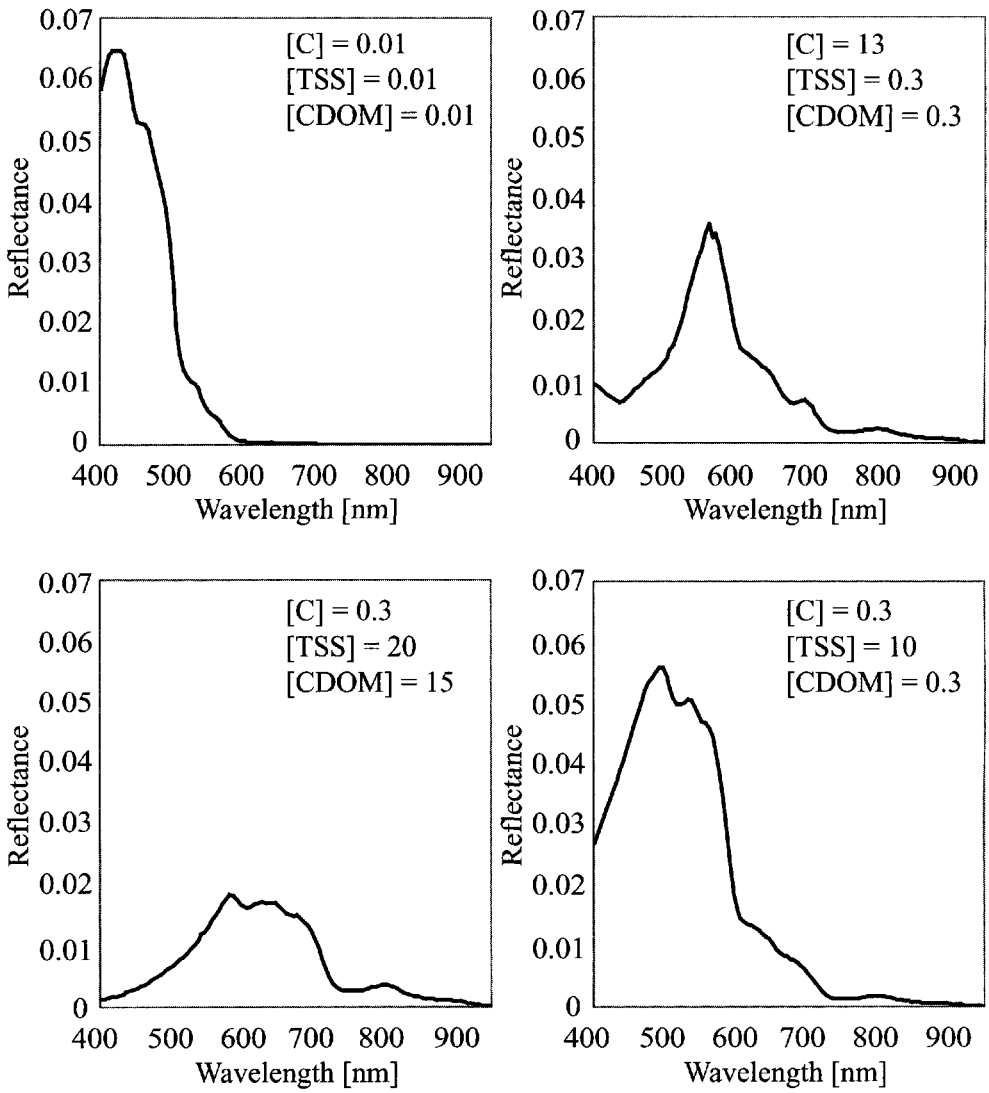


Figure 11.8 Reflectance spectra for water generated using Hydrolight.

In order to take advantage of the LUT using the model matching approach described above, the imaging spectrometer radiance cube must be inverted to a reflectance cube. This can be a very difficult task because most of the spectral methods described in Section 7.6 rely on the energy in the reflected signal, which can be very low over water. This is compounded by the need for a very good correction since the reflectance signal from water is typically only a few reflectance units. A residual error of one reflectance unit might represent a 30% error in retrieved reflectance. In the open ocean, we might take advantage of the near zero reflectance of water in the NIR as suggested by Gordon et al. (1997). Raqueño et al. (2000) suggest an alternative approach for coastal or freshwater cases where the water quality in a localized area can be estimated. For the AVIRIS image shown in Figure 11.10, the concentrations of constituents in the open water was known to be quite low and could be estimated based on seasonal conditions or, in this case, samples coincident with the overflight. These concentrations were used to estimate

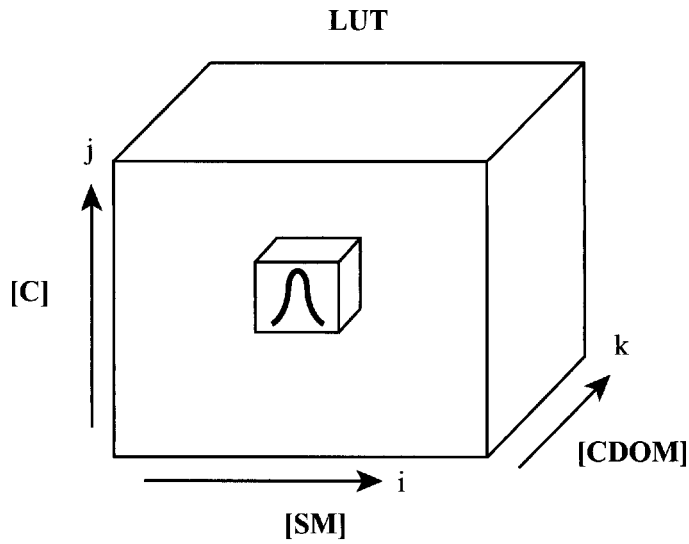
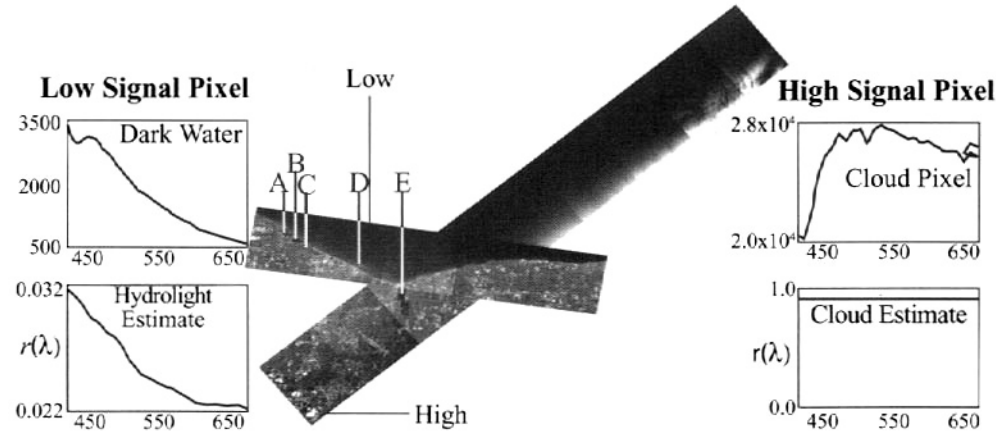


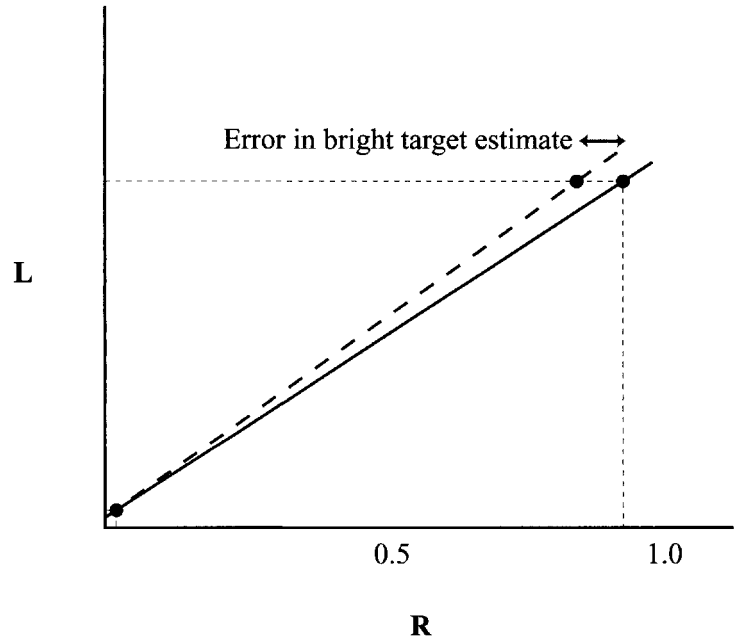
Figure 11.9 Multiparameter LUT where each cell is a spectral reflectance vector, examples of which are shown in Figure 11.8.

the reflectance of a dark water pixel using the Hydrolight model. The reflectance of a bright pixel in the scene was also estimated. In this case, a cloud pixel was estimated to have a uniform reflectance of approximately 0.9. These reflectance spectra can then be used to perform an empirical line method (ELM) inversion to reflectance for each spectral band. As shown in Figure 11.10, a significant error in the estimated reflectance of the bright target will generate very small errors in the retrieved water reflectance because the water pixels are so dark. This atmospheric inversion method not only inverts the image data into reflectance space but also forces consistency between the in-water RT model and the retrieved results for the pixel used as the dark object in the ELM. This can help to remove any residual sensor, atmosphere or modeling errors.

Once the image is inverted to reflectance, the spectrum for each pixel is matched against the LUT using an RMS error metric as the matching criteria. The search process is optimized using the Amoeba algorithm suggested by Green et al. (1993) and described by Press (1986). Because the LUT space is well represented as piecewise linear, the search process can include three-dimensionally linearly interpolated spectra yielding denser sampling of the parameter concentrations than available directly from the LUT entries. Figure 11.11 shows concentration maps derived from the model matching process for the three in water constituents. These maps show spatially important features of this analysis process. First, the patterns are spatially consistent and slowly varying as expected for water parameters. This is most prominent in the various bays along the shore where variations between the bays are pronounced with less variation within the bay. Second, the ability of the method to differentiate between sources of brightness variation is seen in the [SM] map where the Genesee River sediment plume into Lake Ontario is mapped as a



(a)



(b)

Figure 11.10 Illustration showing (a) two AVIRIS flight lines over Lake Ontario shoreline along with sample points and spectra used in ELM-based atmospheric compensation and (b) an example ELM fit in one band is shown as the solid line. The dashed lines show how little error is introduced for low-reflectance targets by a large error in the estimate of the bright target's reflectance.

high concentration of [SM] but does not show up in the [C] and [CDOM] maps. A more quantitative comparison is shown in Figure 11.12, where the retrieved concentration is shown with ground truth values for several sample locations mapped in Figure 11.10. The results show relatively low residual error compared to the natural variation in the constituents. This means the relative trophic status of these waters could potentially be mapped and monitored using this approach.

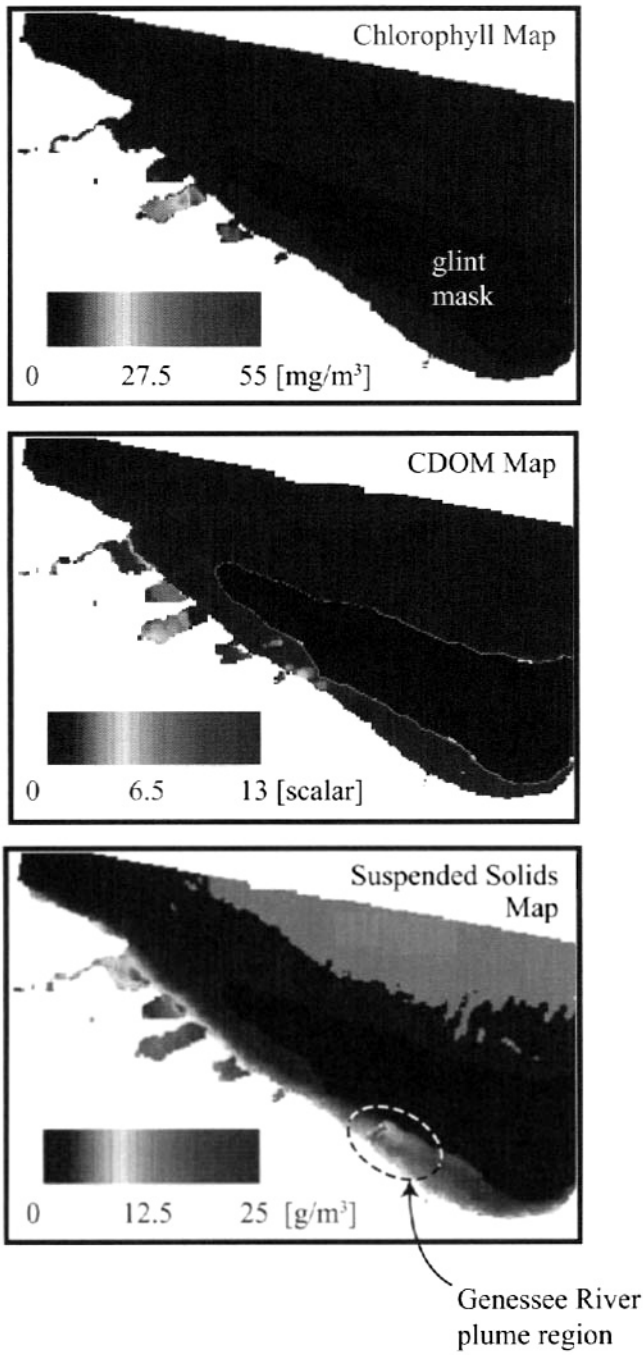


Figure 11.11 Concentration maps of water-quality constituents derived using the model matching method applied to AVIRIS images of the Lake Ontario shoreline. Note the region labeled glint was not analyzed due to solar glint from the surface that can be avoided by proper timing of the image acquisition. See color plate 11.11.

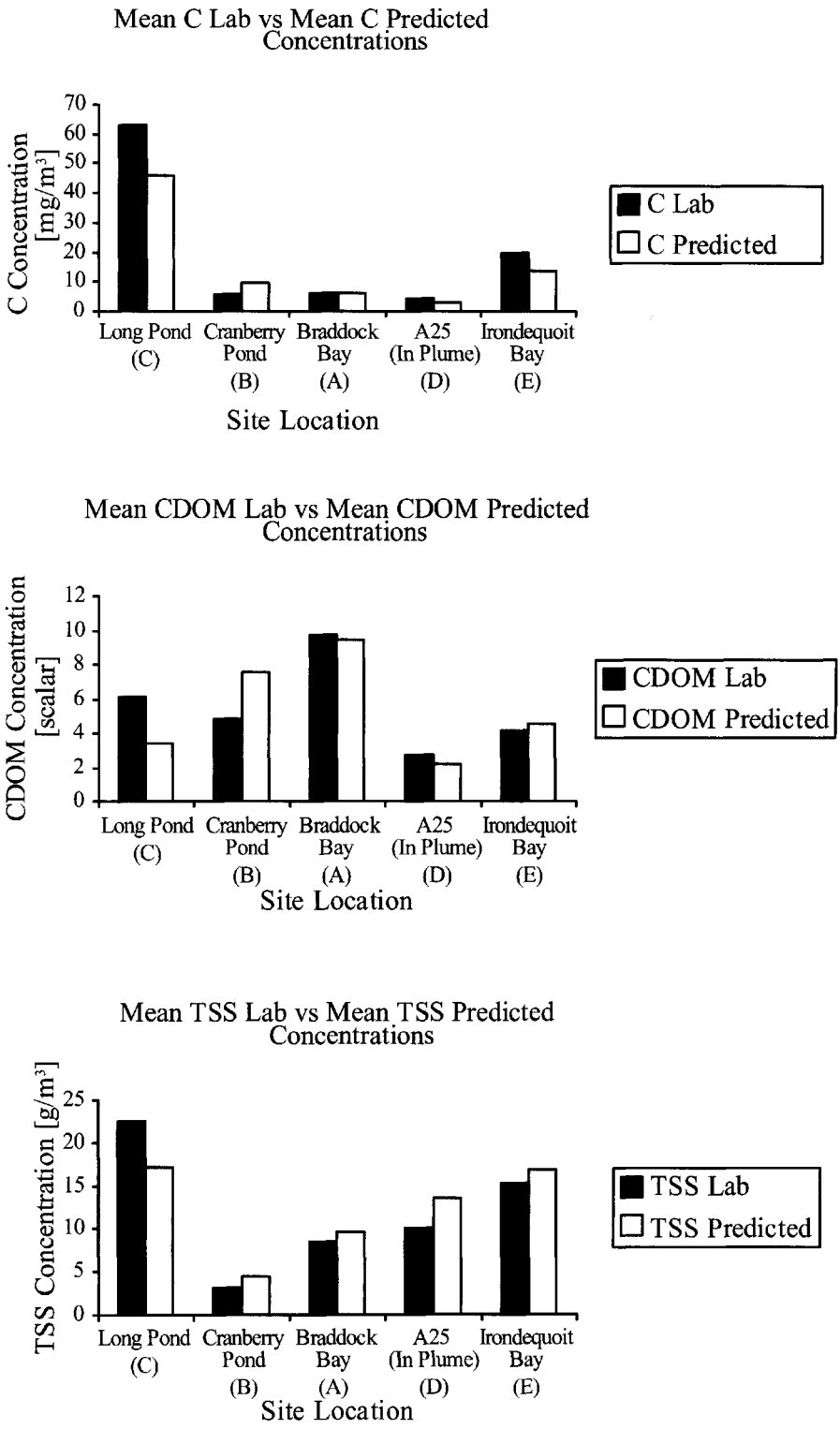


Figure 11.12 Comparison of concentrations derived from the model matching process to laboratory values measured from samples taken at the time of the AVIRIS overflight for the locations labeled in Figure 11.10.

11.3 STATISTICAL INVERSION OF PHYSICAL MODELS

A limitation of many physics-based models used in remote sensing is that they are fundamentally forward propagation models. This means they are designed to take in descriptions of the world and to produce estimates of how the world will appear to various sensing devices. In practice, what we often want to do is to run these models backward and invert observations acquired with sensors into descriptions of the world (cf. Fig. 11.13). However, many of the models include nonlinear steps that do not readily invert. This is particularly characteristic of the radiative transfer codes such as the MODTRAN code described in Chapter 7. Hernandez-Baquero and Schott (2000a) suggest using linear predictive statistics to capture the information in a radiative transfer code in an invertible fashion. In particular, they use *canonical correlation regression* (*CCR*) analysis to estimate MODTRAN inputs

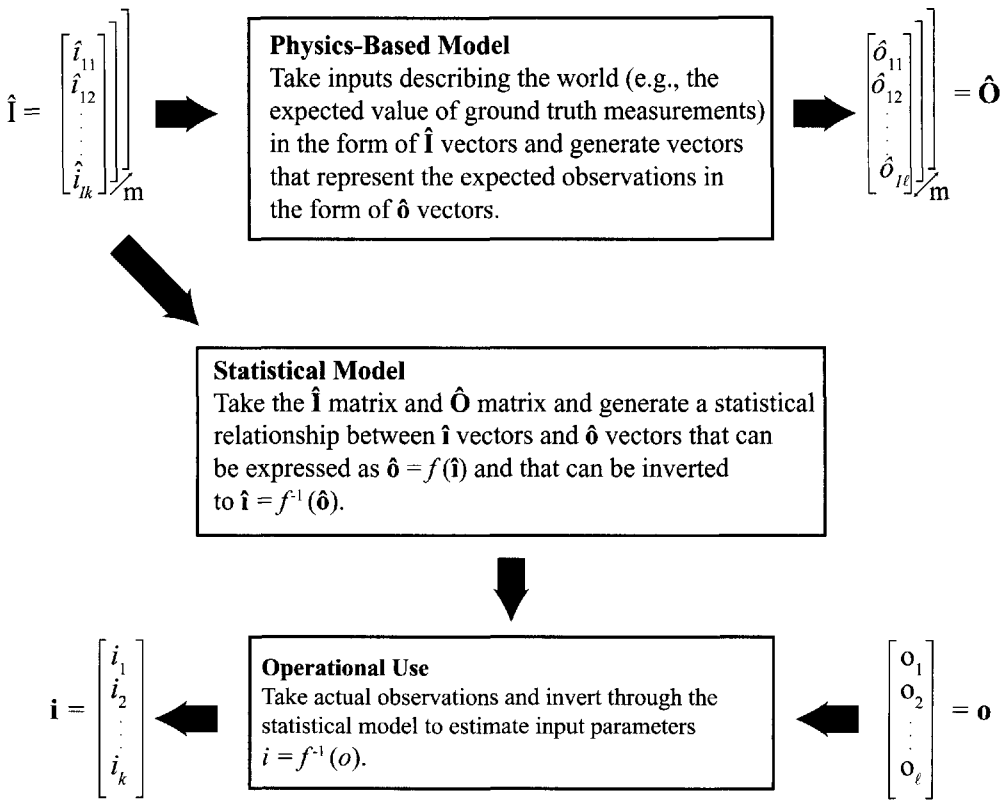


Figure 11.13 Illustration of the concept of using physical models to train a statistical model that can be inverted in cases where the physical model cannot. In general, the inputs (**î**) are physical descriptions of the world we might expect to know from ground truth and the outputs (**ô**) are the observations we would expect to make based on the input vector. The outputs will typically be in terms of reflectance or radiance spectral vectors that are directly or indirectly related to a remotely sensed observation. Once the statistical model is built, we use actual observations (**o**) in the place of expected values to predict input values (**i**).

or atmospheric parameters from MODTRAN-predicted spectral radiance values. The goal is to invert observed spectral radiance values to MODTRAN input values that describe the Earth (e.g., surface temperature) or to atmospheric parameters (e.g., transmission and path radiance) that can be used to perform atmospheric compensation. The principles behind CCR, as applied to this problem, are treated in Hernandez-Baquero and Schott (2000b), and the details of the CCR statistics are covered in Jackson (1991). Our treatment will be at the applications level since a full mathematical derivation of the CCR process is outside our scope. CCR is designed to generate a linear predictive model that maximizes the correlation between the variability of one set of parameters and a second set based on the joint covariance structure between the two sets (i.e., the covariance of both data sets is included in the process). CCR is focused on reproducing the variability of the input variables about their mean based on the variability of the output variables about their mean. As such, it is important when training the statistical model to ensure that the distribution used in the training process is indicative of the distribution we expect to observe.

To illustrate the approach, Hernandez-Baquero and Schott (2000a) took three sets of radiosonde data and built three separate CCR models to predict MODTRAN's behavior for the three data sets. The three data sets consisted of sample radiosondes representing the U.S. West Coast, the U.S. East Coast, and a more global data set. The MODTRAN inputs for each run were characterized by the temperature at each of n atmospheric layers, the amount of water vapor at each of n layers, and the surface temperature. The rest of the MODTRAN parameters were fixed at values most closely representing the imaging/atmospheric conditions associated with the actual observation conditions (e.g., latitude, longitude, ground elevation, sensor elevation and view angle, and aerosol type). This led to a $k = 2n + 1$ element input vector. For the initial runs, the emissivity input to MODTRAN was assumed to be that of a blackbody, and the target temperature was set at the temperature of the lowest radiosonde and at values 2, 4, and 6 K above and below this value. For later runs, spectral emissivity curves for ocean, desert, and farmland were included as inputs to MODTRAN impacting the observed radiance vectors but not the recorded input vectors. The MODTRAN output spectral radiance vectors were convolved and sampled to represent either 10 bands in the LWIR corresponding to the MASTER airborne sensor or 128 LWIR bands corresponding to the SEBASS sensor. Figure 11.14 illustrates the data set used as input to MODTRAN, as well as the MODTRAN-generated data.

Let $\hat{\mathbf{I}}$ be the m by k matrix composed of MODTRAN input data as row vectors ($\hat{\mathbf{i}}$) and $\hat{\mathbf{O}}$ be the m by ℓ matrix of corresponding MODTRAN radiance vectors convolved and resampled to form the ℓ dimensional row vectors $\hat{\mathbf{o}}$ representing the sensor reaching radiance. Then the inputs to the CCR process are the matrices $\mathbf{Y}$ and $\mathbf{X}$ formed of $\hat{\mathbf{I}}$ and $\hat{\mathbf{O}}$, respectively, according to

$$\mathbf{Y}^{\mathrm{T}} = \hat{\mathbf{I}}^{\mathrm{T}} - <\hat{\mathbf{i}}>^{\mathrm{T}} \tag{11.22}$$

and

$$\mathbf{X}^{\mathbf{T}} = \hat{\mathbf{O}}^{\mathbf{T}} - <\hat{\mathbf{o}}>^{\mathbf{T}} \tag{11.23}$$

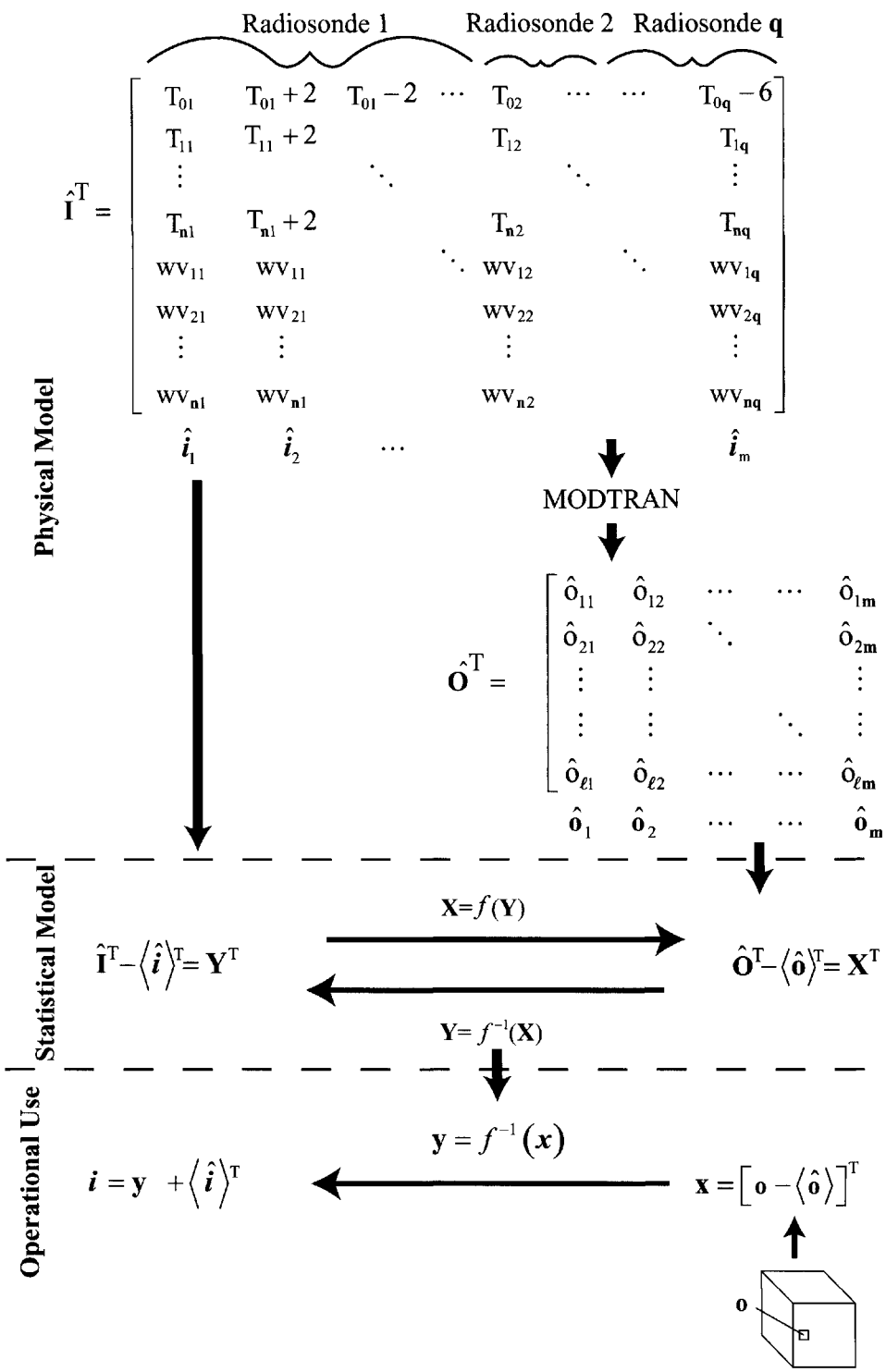


Figure 11.14 Illustration of the CCR approach applied to MODTRAN LWIR data. The matrix of MODTRAN input data are used to generate the matrix of output data. Each column of $\hat{\mathbf{I}}$ characterizes the input data used to generate the corresponding column of $\hat{\mathbf{O}}$. The matrices of demeaned vectors **y** and **x** form the inputs to the CCR process (see Fig. 11.15), which generates the function which can be used to estimate MODTRAN input vectors **I** from observed image vectors **O**.

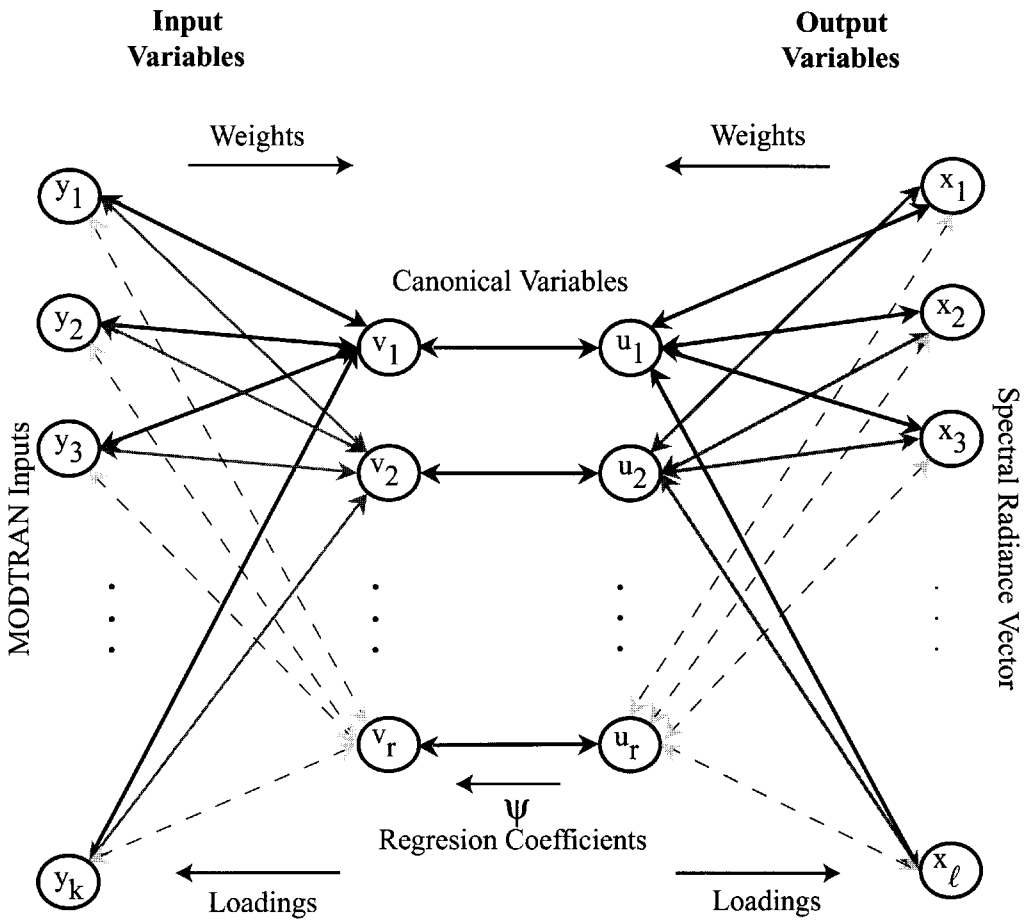


Figure 11.15 Illustration of the relationship of the variables used in canonical correlation regression analysis. Note that this is a fully reversible process and that the weights and loading can provide insight into the relative importance of the variables.

where $\langle \hat{\mathbf{i}} \rangle$ and $\langle \hat{\mathbf{o}} \rangle$ are the mean values of the rows of $\hat{\mathbf{I}}$ and $\hat{\mathbf{O}}$, respectively. The covariances of $\mathbf{Y}$ and $\mathbf{X}$ can then be expressed as

$$\mathbf{S}_{YY} = \frac{1}{m}\mathbf{Y}^{T}\mathbf{Y} \tag{11.24}$$

and

$$\mathbf{S}_{XX} = \frac{1}{m}\mathbf{X}^{T}\mathbf{X} \tag{11.25}$$

Referring to Figure 11.15, we can form two sets of canonical variables v and u where each v variable is a linear combination of the y variables (k in number) that make up the row vectors ($\mathbf{y}$) in the $\mathbf{Y}$ matrix, and each u variable is a linear combination of the x canonical variables (ℓ in number) that make up the row vectors

(**x**) in the **X** matrix. The v variables are the projection of the **y** vectors onto the eigenvectors making up the matrix **B,** which diagonalizes the composite matrix $\mathbf{C}_Y$ according to

$$\mathbf{C}_Y\mathbf{B}=\mathbf{S}_{YY}^{-1}\mathbf{S}_{YX}\mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}\mathbf{B}=\boldsymbol{\psi}\mathbf{B} \tag{11.26}$$

where **B** is the matrix composed of the eigenvectors of $\mathbf{C}_Y$, $\boldsymbol{\psi}$ is the diagonal matrix of eigenvalues whose square roots are referred to as canonical correlations and S_{YX} and S_{XY} are the covariance matrices of **Y** with **X** and **X** with **Y** respectively, i.e.,

$$\mathbf{S}_{YX}=\frac{1}{m}\mathbf{Y}^{\mathrm{T}}\mathbf{X} \tag{11.27}$$

and

$$\mathbf{S}_{XY}=\frac{1}{m}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \tag{11.28}$$

Similarly the u variables are formed by projecting the **x** vectors onto the eigenvectors making up the matrix **A**, which diagonalizes the composite matrix $\mathbf{C}_X$ according to:

$$\mathbf{C}_X\mathbf{A}=\mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}\mathbf{S}_{YY}^{-1}\mathbf{S}_{YX}\mathbf{A}=\psi\mathbf{A} \tag{11.29}$$

Thus

$$u_i=\mathbf{x}\boldsymbol{a}_i \tag{11.30}$$

and

$$v_i=\mathbf{y}\mathbf{b}_i \tag{11.31}$$

or

$$\mathbf{u}=\mathbf{x}\mathbf{A} \tag{11.32}$$

and

$$\mathbf{v}=\mathbf{y}\mathbf{B} \tag{11.33}$$

or

$$\mathbf{U}=\mathbf{X}\mathbf{A} \tag{11.34}$$

and

$$\mathbf{V}=\mathbf{Y}\mathbf{B} \tag{11.35}$$

where u_i is the ith output canonical variable generated by projecting the output vectors onto the ith eigenvector making up the matrix **A** (note that the u variables will be decorrelated from each other by the eigen analysis), v_i is the ith input canonical variable generated by projecting the input vectors onto the ith eigenvector making up the matrix **B** (note that the v variables are also decorrelated from each other by

the eigen analysis). We can form row vectors **u** and **v** composed of all the output and input canonical variables respectively using Eq. (11.32) and (11.33), respectively. In practice, we find that the covariance matrices and the composite matrices are often rank deficient, and we must use singular value decomposition (SVD) (cf. Chap. 9) to perform the analysis, with only the significant canonical correlations included. As a result, the **u** and **v** vectors will be of the same dimension since **A** and **B** share common eigenvalues. Finally, the matrices **U** and **V** are simply the matrices made by transforming the entire output data set **X** and input data set **Y** into the corresponding canonical space. Because the canonical variables are independent of each other, the best linear estimator of **V**, expressed as $\hat{\mathbf{V}}$, can be solved for by linear least-squares theory using the pseudo-inverse according to

$$\hat{\mathbf{V}}=\mathbf{U}\boldsymbol{\beta}_{cc} \tag{11.36}$$

where

$$\boldsymbol{\beta}_{cc}=\mathbf{U}^{\#}\mathbf{V}=\left[\left(\mathbf{U}^{\mathrm{T}}\mathbf{U}\right)^{-1}\mathbf{U}^{\mathrm{T}}\right]\mathbf{U}\boldsymbol{\beta}_{cc}=\boldsymbol{\beta}_{cc} \tag{11.37}$$

is the matrix of linear predictors (regression coefficients) that predict **v** canonical vectors given **u** canonical vectors (cf. Fig. 11.15) and is equal to ψ. In practice, because the canonical variables are ranked in terms of the amount of covariance they characterize, only a small subset of canonical variables need to be retained to invert to the input vector **y**. This entire process illustrated in Figure 11.14 can be expressed mathematically as

$$\begin{aligned}
\boldsymbol{i}=\mathbf{y}^{\mathrm{T}}+<\hat{\boldsymbol{i}}>&=[\mathrm{f}^{-1}(\mathbf{x})]^{\mathrm{T}}+<\hat{\boldsymbol{i}}>\\
&=[\mathbf{v}\mathbf{B}^{\mathrm{T}}]^{\mathrm{T}}+<\hat{\boldsymbol{i}}>\\
&=\left[\left(\mathbf{u}\boldsymbol{\psi}\right)\mathbf{B}^{\mathrm{T}}\right]^{\mathrm{T}}+<\hat{\boldsymbol{i}}>\\
&=\left[\left((\mathbf{x}\mathbf{A})\boldsymbol{\psi}\right)\mathbf{B}^{\mathrm{T}}\right]^{\mathrm{T}}+<\hat{\boldsymbol{i}}>\\
&=\left[\left(((\mathbf{o}-<\hat{\mathbf{o}}>)^{\mathrm{T}}\mathbf{A})\boldsymbol{\psi}\right)\mathbf{B}^{\mathrm{T}}\right]^{\mathrm{T}}+<\hat{\boldsymbol{i}}>
\end{aligned} \tag{11.38}$$

where f^{-1} is the inverse of the forward model, **o** is the observed image radiance vector, **i** is the predicted input vector associated with **o**, and we have invoked the fact that $\mathbf{B}^{\mathrm{T}}=\mathbf{B}^{-1}$ to invert **v** vectors to **y** vectors [cf. Eq. (11.33)].

To test the basic concept that MODTRAN-generated data sets could be estimated using a linear predictive statistical model, a 10-band LWIR data set acquired with the ModIS-Aster (MASTER) airborne sensor was studied. Two data sets were generated. In the first a set of sensor-reaching radiance values were predicted (simulated) using MODTRAN corresponding to water temperatures (measured as part of the study) using a best estimate of the prevailing atmospheric conditions. In the second data set, actual measured sensor-reaching radiance vectors were used for the same targets as shown in Figure 11.16. The results are shown in Table 11.1 for the three radiosonde data sets for which models were built. The simulated data

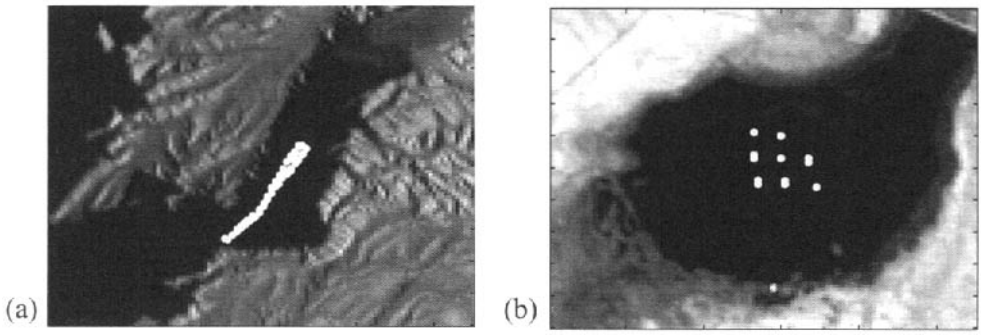


Figure 11.16 MASTER LWIR images showing thermal sampling locations in (a) Lake Mead and (b) Cold Springs Reservoir.

show the errors in retrieved surface temperature just in the modeling process to be quite low for this case where the model treated all targets as blackbodies [which water closely approximates ($\varepsilon \approx 0.986$)]. The actual errors in retrieved temperature are also quite low. The actual errors include errors due to sensor and ground truth instrument calibration and instrument noise, as well as modeling errors. They show reasonable results for all data sets; however, the West Coast radiosonde data set corresponding to the collection location conditions shows slightly better overall performance. Note that the CCR process also produces temperature and water vapor profiles as part of the inversion process. Thus, it can be used as a crude atmospheric sounding process.

In a second experiment, the input data were modified to include three targets with varying emissivity (water, desert, and farm), to determine if the CCR process could generate a set of inversion statistics that focused on the surface temperature and atmospheric effects and ignored the emissivity variations. The results shown in Table 11.2 were run for simulated data representing both 10 bands of MASTER data and 128 bands of SEBASS data for the three radiosonde data sets. The actual retrieved water temperature errors for these CCR models are also shown for the MASTER data set. These results show that the direct retrieval errors are rather large for the MASTER resolution, but at SEBASS resolution the CCR models appear to compensate for much of the emissivity variation. The error in the actual re-

Table 11.1 Error in Retrieved Temperature (K) Attributed to the Modeling Process and Observed Error for MASTER Data from Two Water Sites Analyzed Using Three Different Radiosonde Data Sets and Blackbody Assumptions

Radiosonde Set	Simulated MASTER Data		Actual MASTER Data	
	Lake Mead	Cold Spring	Lake Mead	Cold Spring
West Coast (low water vapor range)	0.11	0.07	0.37	0.71
East Coast	0.15	0.16	0.71	1.16
Global	0.08	0.27	0.33	0.88

Table 11.2 Errors in Retrieved Temperature (K) Attributed to the Modeling Process for Varying Emissivity Targets and Actual Errors for Water Targets Using the Models Developed for Variable Emissivity Targets

Radiosonde Set	Simulated MASTER		Simulated SEBASS		Actual MASTER	
	Lake Mead	Cold Springs	Lake Mead	Cold Springs	Lake Mead	Cold Springs
West Coast	1.13	1.45	0.60	0.47	1.87	3.50
East Coast	1.19	1.91	0.53	0.55	1.75	1.95
Global	1.99	2.59	1.24	1.23	2.7	2.05

trievals reflects the additional error due to sensor noise and instrument calibration. Note that the increased error in retrieval temperature in the actual data between Table 11.1 and 11.2 reflects a more universal model where target emissivity is not assumed to be nearly black.

These results indicate that a nonlinear RT model like MODTRAN may, under a limited set of conditions, be approximated by a linear invertible statistical model that can be used to solve in this case for surface temperatures. Hernandez-Baquero and Schott (2000a) point out how the CCR process can also be used to invert observed radiance to spectral estimates of atmospheric transmission and upwelled and downwelled radiance that can be used to retrieve both surface temperature and spectral emissivity. A limitation of the CCR approach is that it emphasizes a model built on the variation in the input and output data about the ensemble average. The user must therefore take considerable care in building a model data set that reflects the distribution of output data expected in the actual observation data set.

11.4 INCORPORATION OF PHYSICS-BASED MODELS INTO SPECTRAL ALGORITHM TRAINING

One of the persistent problems facing the remote sensing algorithm developer is the disconcerting fact that a target's appearance (i.e., spectral reflectance) can take on many manifestations. A given spectral reflectance vector can generate many spectral radiance vectors depending on atmospheric and illumination effects. In fact, even after standard atmospheric correction of a scene to apparent reflectance, local illumination effects will continue to cause significant changes in the apparent reflectance spectrum. Healey and Slater (1999) suggest an approach to this problem that employs physics-based models to describe the target using a set of basis vectors that incorporate atmospheric and illumination effects. This results in a target detection algorithm that does not vary with changes in the atmospheric and illumination effects (i.e., it is invariant with respect to these variables). For this "*invariant*" approach to be viable, we must be able to predict the range of spectral values the target might take on in spectral space and then be able to span that range with a small set of basis vectors. In essence, this means that the target only occupies a relatively small subspace of the entire spectral space that, hopefully, does not overlap extensively with the background space. Healey and Slater (1999)

suggest that the target's appearance in spectral radiance space can be predicted by generating spectral radiance vectors using MODTRAN for a single reflectance vector and a range of atmospheric, illumination, and sensor geometry conditions [cf. Fig. 11.17(a)]. Included in these calculations are variations in the sunlight-to-skylight ratio to include objects in shadow. For any given reflectance spectrum, they generated 17,920 viable sensor reaching radiance spectra that were convolved and resampled to 210 bands corresponding to the Hydice sensor. These radiance spectra can potentially be described by a small set of basis vectors according to

$$\boldsymbol{L}_i = \boldsymbol{M}\boldsymbol{a}_i + \boldsymbol{\varepsilon}_i = \sum_{j=1}^{N} \boldsymbol{m}_j a_{ij} + \boldsymbol{\varepsilon}_i \tag{11.39}$$

where $\boldsymbol{L}_i$ is the ith spectral radiance vector, $\mathbf{M}$ is the matrix made up of basis vectors as columns, $\mathbf{a}_i$ is the column vector of scalar weights to be applied to each basis vector to estimate the ith radiance vector, $\boldsymbol{\varepsilon}_i$ is the residual error, N is the number of basis vectors, $\mathbf{m}_j$ is the jth basis vector, and a_{ij} is the scalar weight applied to the jth basis vector to estimate the ith spectral radiance vector.

Given an ensemble of 17,920 vectors for a given reflectance spectrum, Slater and Healey (1998) use SVD (cf. Sec. 9.1.3) to generate a set of N basis vectors. They repeated this process for a range of target reflectance spectra (498 different targets) and varied the number of basis vectors. They found, using the residual error as a metric, that nine or fewer basis vectors were typically sufficient to span the target spectral radiance space. The actual target space in any given image is likely to be even smaller than this since the range of conditions included in the MODTRAN runs was much larger than the range expected in a single image. Since this number (≤ 9) is much smaller than the dimensionality of the image, we may expect that the target subspace is small compared to the image/background subspace, making target detection reasonable even with the many ways a target may be manifest. An example of this is shown in Figure 11.17(b). To make the illustration manageable, we show only a three-dimensional radiance space. If within this space, two basis vectors can be used to describe the target space, then linear combinations of those two basis vectors will occupy a two-dimensional subspace. The likelihood of target detection is improved because this subspace is a small fraction of the whole. However, the relative location of the target and background subspaces is still the critical factor in the performance of a target detection algorithm.

Healey and Slater (1999) suggest a target detection algorithm based on thresholding the residual error term in Eq. (11.39). The residual is calculated after attempting to generate a best-fit description of each pixel to the target model described by Eq. (11.39). This means we must first find the best estimate of the vector $\mathbf{a}_i$ for each spectral radiance vector $\mathbf{L}_i$ in the scene. For a given target reflectance vector, we can solve using SVD for a set of basis vectors to span the radiance space described by MODTRAN for that target [cf. Fig. 11.17(a)]. The vector of weights that best describe an observed radiance can then be solved for using the pseudo-inverse according to:

$$\boldsymbol{M}^{\#}\boldsymbol{L}_i = \boldsymbol{M}^{\#}\boldsymbol{M}\boldsymbol{a}_i = \hat{\boldsymbol{a}}_i = (\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M})^{-1}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{L}_i = \boldsymbol{I}^{-1}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{L}_i = \boldsymbol{M}^{\mathrm{T}}\boldsymbol{L}_i \cong \boldsymbol{a}_i \tag{11.40}$$

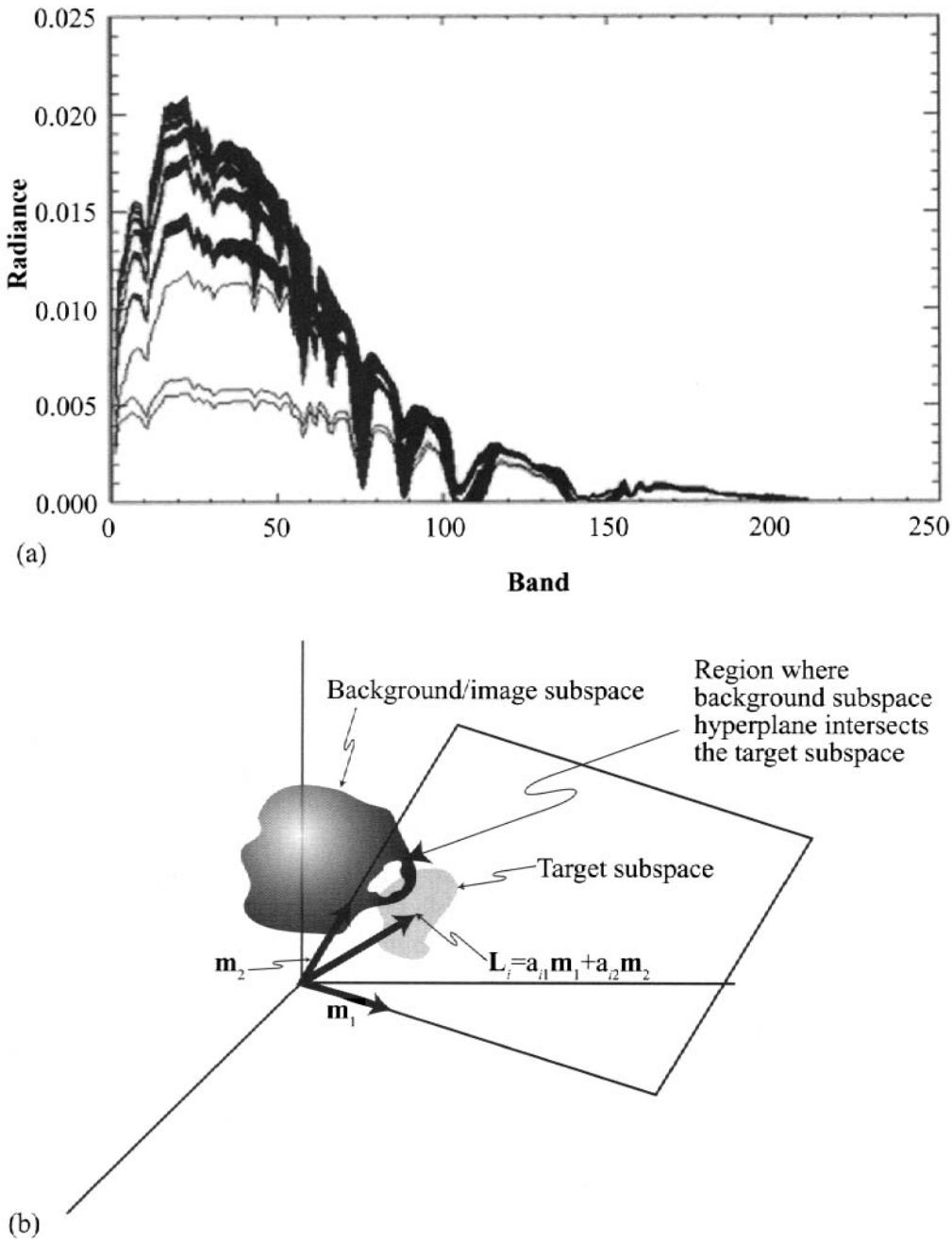


Figure 11.17 Illustration of concepts used in the Invariant algorithm: (a) Family of spectral radiance curves generated using MODTRAN applied to a single reflectance spectrum. (b) Simplified 3-D illustration of radiance space with the background data occupying some subset of the entire space and the target occupying a 2-D subspace of the whole spanned by basis vectors $\mathbf{m}_1$ and $\mathbf{m}_2$.

or

$$\hat{a}_{ij} = \boldsymbol{m}_j^{\mathrm{T}} \boldsymbol{L}_i \tag{11.41}$$

where we have taken advantage of the fact that the basis vectors making up the matrix **M** are found using SVD and are therefore orthonormal such that $\mathbf{M}^T\mathbf{M}$ is the identity matrix. The magnitude of the residual error vector can then be expressed as

$$\|\varepsilon_i\| = \|\boldsymbol{L}_i - \boldsymbol{M}\hat{\boldsymbol{a}}_i\| = \|\boldsymbol{L}_i - \boldsymbol{M}(\boldsymbol{M}^{\mathrm{T}}\boldsymbol{L}_i)\| \tag{11.42}$$

If the pixel is target-like, it should be well described by the target basis vectors making up **M**, and the residual error should be quite small. On the other hand, a background pixel could be poorly described by the model and result in a large error. Thus, we can threshold a residual error image to perform target detection. Healey and Slater (1999) applied this approach to HYDICE data and were able to locate targets in various illumination conditions (full sun, partial sun, full shadow) with very low false-alarm rates. By comparison, the SAM algorithm (cf. Sec. 10.32), when trained on the sunlit targets, had still not found some of the shadowed targets after large numbers of false alarms. The value of this approach is not only in its performance but also in the limited amount of user information required to apply it. All the user needs to specify is the target reflectance spectrum and the sensor spectral response (assuming an instrument calibrated to spectral radiance). The algorithm can then generate the target-sensor-specific basis vectors and then form the residual error images This approach not only eliminates the need for atmospheric compensation but also offers the potential of accounting for atmospheric and illumination variations not compensated for by traditional methods.

A serious limitation to this basic approach is that it is designed to find fully resolved target pixels. A mixed pixel will be a poor match to the target basis vector model described by Eq. (11.39) and will produce high residual error. Thai and Healey (2002) and Schott et al. (2003) suggest extensions to the invariant approach to deal with the subpixel case. We will present the approach of Schott et al. (2003) since it utilizes the max-D approach already described in Chapter 10 to deal with the case of mixed target pixels contaminating the background basis vectors.

Schott et al. (2003) suggest using the standard invariant method described above to generate a set of target basis vectors, except they use the max-D method (described in Section 10.3.1) rather than the SVD to generate basis vectors that more closely resemble native end members. They also suggest only varying MODTRAN over the range of atmospheric and illumination conditions expected in the scene when forming the ensemble of possible target radiance vectors. This is designed to make the target subspace as large as necessary to encompass the expected appearance of the target, but not larger. This reduces the likelihood of overlap between the target and background subspaces. They also suggest adding sensor noise to the MODTRAN-predicted target radiance spectra to more appropriately simulate what the targets will look like in the image space. The max-D algorithm is separately run on the scene to generate a set of candidate background end mem-

bers. Schott et al. (2003) point out that these candidate end members may unwittingly include target or mixed target background end members (cf. Fig. 11.18). To overcome this limitation, the target basis vectors are included with the scene vectors when max-D is run on the scene data. One or more of these "pure" target vectors is typically selected as a scene end member. However, as shown in Figure 11.18, any less pure or mixed end members in the scene will tend not to be selected. Any target end members (or any scene end member closely matching them) selected from the concatenated scene-target data can then be removed, leaving a set of background-only end members. Note that all the data are typically transformed into a reduced dimensional space (e.g., MNF transform from Sec. 10.2.3) prior to any analysis. Thus, all the vectors in the following treatment should be assumed to be in the transform space.

The residual error associated with the background model for the *i*th pixel can then be expressed as

$$\boldsymbol{\varepsilon}_{Bi} = \boldsymbol{L}_i - \boldsymbol{B}\boldsymbol{b}_i = \boldsymbol{L}_i - \boldsymbol{B}\boldsymbol{B}^{\#}\boldsymbol{L}_i = [\boldsymbol{I} - \boldsymbol{B}\boldsymbol{B}^{\#}]\boldsymbol{L}_i = \boldsymbol{p}_B^{\perp}\boldsymbol{L}_i \tag{11.43}$$

where $\mathbf{B}$ is the matrix having the background basis vectors as columns, $\mathbf{b}_i$ is the vector of best fit weights to be applied to the basis vectors to model the *i*th radiance vector, $\mathbf{B}^{\#}$ is the pseudo-inverse of $\mathbf{B}$, and $\mathbf{p}_B^{\perp}$ is the nulling operator (cf. Sec. 10.3.2) that projects $\mathbf{L}_i$, the transformed image vector, onto the subpixel orthogonal to the vectors making up $\mathbf{B}$.

Similarly, the residual error associated with a mixed target background model for the *i*th pixel can be expressed as

$$\boldsymbol{\varepsilon}_{Hi} = \boldsymbol{L}_i - (\boldsymbol{B}\boldsymbol{b}_i + \boldsymbol{T}\boldsymbol{t}_i) = \boldsymbol{L}_i - \boldsymbol{H}\boldsymbol{h}_i = \boldsymbol{L}_i - \boldsymbol{H}\boldsymbol{H}^{\#}\boldsymbol{L}_i = [\boldsymbol{I} - \boldsymbol{H}\boldsymbol{H}^{\#}]\boldsymbol{L}_i = \boldsymbol{p}_H^{\perp}\boldsymbol{L}_i \tag{11.44}$$

where $\mathbf{T}$ is the matrix having the target basis vectors as columns, $\mathbf{t}_i$ is the vector of target weights associated with the *i*th pixel, $\boldsymbol{H}$ is the matrix formed by concatenating the $\boldsymbol{B}$ and $\boldsymbol{T}$ matrices (i.e., it has background and target basis vectors as columns), and $\boldsymbol{h}$ is the vector of best-fit weights to be applied to model the *i*th radiance vector as a mixture of target and background.

We can use the generalized likelihood ratio test (GLRT) introduced in Section 10.4.2.2 to formulate and compare two hypotheses concerning the status of the pixel represented by $\mathbf{L}_i$. The GLRT is an extension of classic Neyman-Pearson detection theory that replaces the unknown parameters in a probability distribution function with their maximum likelihood estimate (MLE) [cf. Lehman (1959)]. In our case, we seek to compare the hypothesis that the pixel includes a target (H_1) to the hypothesis that it only contains background (H_0). If we assume that the error estimation approach produces residual errors which are Gaussian, independent and identically distributed (iid), we can express these hypotheses as follows:

$$H_0 : L \sim N(\boldsymbol{B}\boldsymbol{b}, \sigma_0^2 \boldsymbol{I}) \tag{11.45}$$

and

$$H_1 : L \sim N(\boldsymbol{H}\boldsymbol{h}, \sigma_1^2 \boldsymbol{I}) \tag{11.46}$$

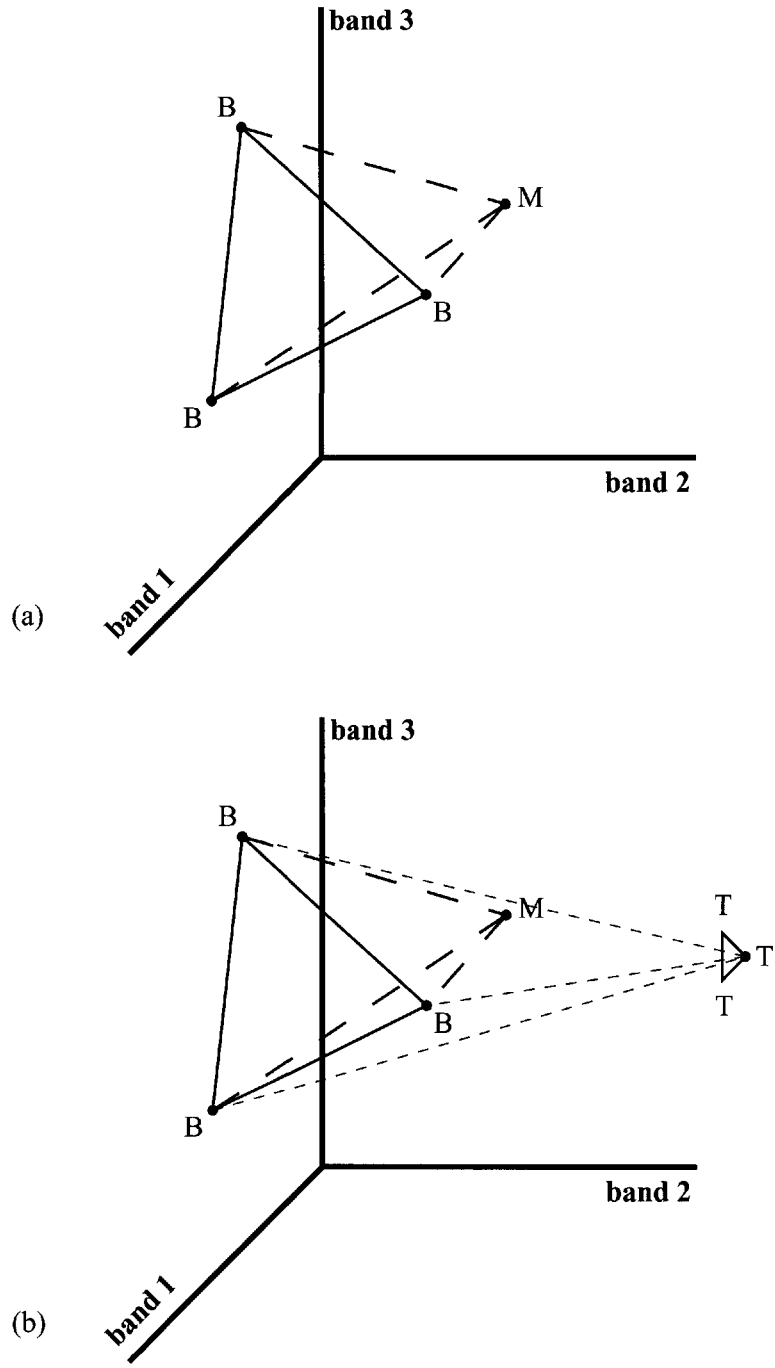


Figure 11.18 Illustration of the results of applying the max-D algorithm to (a) scene data and (b) scene data concatenated with the target basis vectors derived from a max-D analysis of the MODTRAN-generated ensemble of possible target spectral radiance vectors—B is a background end member, T is a target end member, and M is a mixed target-background end member.

where σ_0 is the standard deviation per band of the error in estimating L using the background assumption and σ_1 is the standard deviation per band of the error in estimating **L** using the mixed target-background assumption. For the null hypothesis, this can be read that for this particular pixel, the maximum likelihood estimate for **L** is a multivariate normal distribution with a mean of $\mathbf{Bb} \cong \mathbf{BB}^{\#}\mathbf{L}$ and the Gaussian iid error about that mean has a standard deviation of σ_0 in each band.

The likelihood functions for the distribution can be expressed as

$$L(\boldsymbol{b},\sigma_0^2;\boldsymbol{L}) = (2\pi\sigma_0^2)^{-k/2} e^{-\frac{\|\boldsymbol{\varepsilon}_B\|^2}{2\sigma_0^2}} \tag{11.47}$$

and

$$L(\boldsymbol{h},\sigma_1^2;\boldsymbol{L}) = (2\pi\sigma_1^2)^{-k/2} e^{-\frac{\|\boldsymbol{\varepsilon}_H\|^2}{2\sigma_1^2}} \tag{11.48}$$

where k is the number of spectral channels (e.g., MNF bands) included in the analysis.

The generalized likelihood ratio can be expressed in detector form as

$$GLRT = \frac{L(\boldsymbol{h},\sigma_1^2;\boldsymbol{L})}{L(\boldsymbol{b},\sigma_0^2;\boldsymbol{L})} > \gamma \tag{11.49}$$

where γ is the detection threshold for deciding that the pixel contains a target. Since σ_0 and σ_1 are unknown, they are estimated using their MLE values according to

$$\sigma_0^2 = \frac{1}{k}\|\boldsymbol{\varepsilon}_B\|^2 = \frac{1}{k}\boldsymbol{\varepsilon}_B{}^{\mathrm{T}}\boldsymbol{\varepsilon}_B \tag{11.50}$$

and

$$\sigma_1^2 = \frac{1}{k}\|\boldsymbol{\varepsilon}_H\|^2 = \frac{1}{k}\boldsymbol{\varepsilon}_H{}^{\mathrm{T}}\boldsymbol{\varepsilon}_H \tag{11.51}$$

Equation 11.49 then becomes

$$GLRT = \frac{\left[2\pi k(\boldsymbol{\varepsilon}_H^{\mathrm{T}}\boldsymbol{\varepsilon}_H)\right]^{\frac{-k}{2}} e^{\frac{-k\boldsymbol{\varepsilon}_H{}^{\mathrm{T}}\boldsymbol{\varepsilon}_H}{2\boldsymbol{\varepsilon}_H^{\mathrm{T}}\boldsymbol{\varepsilon}_H}}}{\left[2\pi k(\boldsymbol{\varepsilon}_B^{\mathrm{T}}\boldsymbol{\varepsilon}_B)\right]^{\frac{-k}{2}} e^{\frac{-k\boldsymbol{\varepsilon}_B^{\mathrm{T}}\boldsymbol{\varepsilon}_B}{2\boldsymbol{\varepsilon}_B^{\mathrm{T}}\boldsymbol{\varepsilon}_B}}} = \left(\frac{\boldsymbol{\varepsilon}_B^{\mathrm{T}}\boldsymbol{\varepsilon}_B}{\boldsymbol{\varepsilon}_H^{\mathrm{T}}\boldsymbol{\varepsilon}_H}\right)^{\frac{k}{2}} > \gamma \tag{11.52}$$

Since taking the $k/2$ root of Eq. (11.52) will not change the rank ordering of results, an implementable version of the *GLRT* can be expressed as

$$GLR = \left(\frac{\varepsilon_B^2}{\varepsilon_H^2}\right) = \frac{\left[\boldsymbol{L}-\boldsymbol{Bb}\right]^2}{\left[\boldsymbol{L}-\boldsymbol{Hh}\right]^2} = \frac{\left[\boldsymbol{L}-\boldsymbol{BB}^{\#}\boldsymbol{L}\right]^2}{\left[\boldsymbol{L}-\boldsymbol{HH}^{\#}\boldsymbol{L}\right]^2} = \frac{\left[\left(\boldsymbol{I}-\boldsymbol{BB}^{\#}\right)\boldsymbol{L}\right]^2}{\left[\left(\boldsymbol{I}-\boldsymbol{HH}^{\#}\right)\boldsymbol{L}\right]^2} = \left[\frac{\boldsymbol{P}_B^{\perp}\boldsymbol{L}}{\boldsymbol{P}_H^{\perp}\boldsymbol{L}}\right] > \hat{\gamma} \tag{11.53}$$

where $\hat{\gamma}$ is a scalar detection threshold that is set interactively by the user for the particular GLR image. Inspection of the GLR operator indicates that the numerator

will be small and the denominator larger when a background model is a good fit. Conversely, the GLR will be large when the target model is a better fit.

Figure 11.19 shows an example of the subpixel invariant technique applied to a Hydice image containing test panels. The image was pixel averaged and resampled to a lower resolution where the test target was only present in subpixel abundance in four pixels. The GLR operator results show very good performance in this low-clutter example. Figure 11.20 shows a more stressing example where the subpixel invariant algorithm was applied to an AVIRIS image of Lake Ontario shoreline, including a portion of Rochester, NY. This is a much busier scene, and the target spectrum was a reddish brown basketball court paint spectrum acquired in the field approximately one year after the AVIRIS image was acquired. The results in Figure 11.20(b) show two regions of strong return. The first is the basketball court where the field spectrum [Fig. 11.20(c)] were obtained, and the second is from the perimeter of a tennis court that air photo and site inspection revealed had been painted with the same paint.

The results show very good target detection and false-alarm suppression and are particularly encouraging because they do not require atmospheric correction and are designed to deal with targets under a range of illumination conditions. The real utility of these physics-based algorithms must still be defined as they are tested on increasingly difficult targets. What is most encouraging about these approaches is that by putting more of the physics into the algorithm, the burden on the user can be significantly reduced. In this case, the only user inputs are the reflectance spectrum of the target of interest (assuming images calibrated to spectral radiance are available) and rough estimates of sensor viewing and atmospheric conditions. O'Donnell et al. (2004) extended this physics-based modeling with the invariant approach a step further to study factory stack plumes. They generate an ensemble of target spectral radiance vectors in the LWIR using a governing equation of the form

$$L_\lambda = \left[\varepsilon_s(\lambda) L_{T_s\lambda} \tau_p(\lambda) + \varepsilon_p L_{T_p\lambda}\right] \tau(\lambda) + L_{u\lambda} \tag{11.54}$$

where ε_s is the background emissivity $L_{T_s\lambda}$ is the Planckian radiance from a surface with temperature T, ε_p is the mean effective emissivity of the gas plume with a mean effective temperature T_p producing Planckian spectral radiance $L_{T_p\lambda}$, $\tau_p(\lambda)$ is the transmission of the plume, $\tau(\lambda)$ is the atmospheric transmission, and $L_{u\lambda}$ is the atmospheric upwelled radiance. Recognize that for an absorbing plume, the transmission can be described as

$$\tau_p(\lambda) = 1 - \varepsilon_p(\lambda) \tag{11.55}$$

and the emissivity can be expressed as

$$\varepsilon_p(\lambda) = ck(\lambda, T_g) \tag{11.56}$$

where c is the concentration path length and k is the absorption spectrum at a particular temperature T_g. By varying the critical parameters impacting the sensor reaching radiance expressed in Eq. (11.54) using database values for the gases, temperature ranges for the plume relative to the background, and MODTRAN for

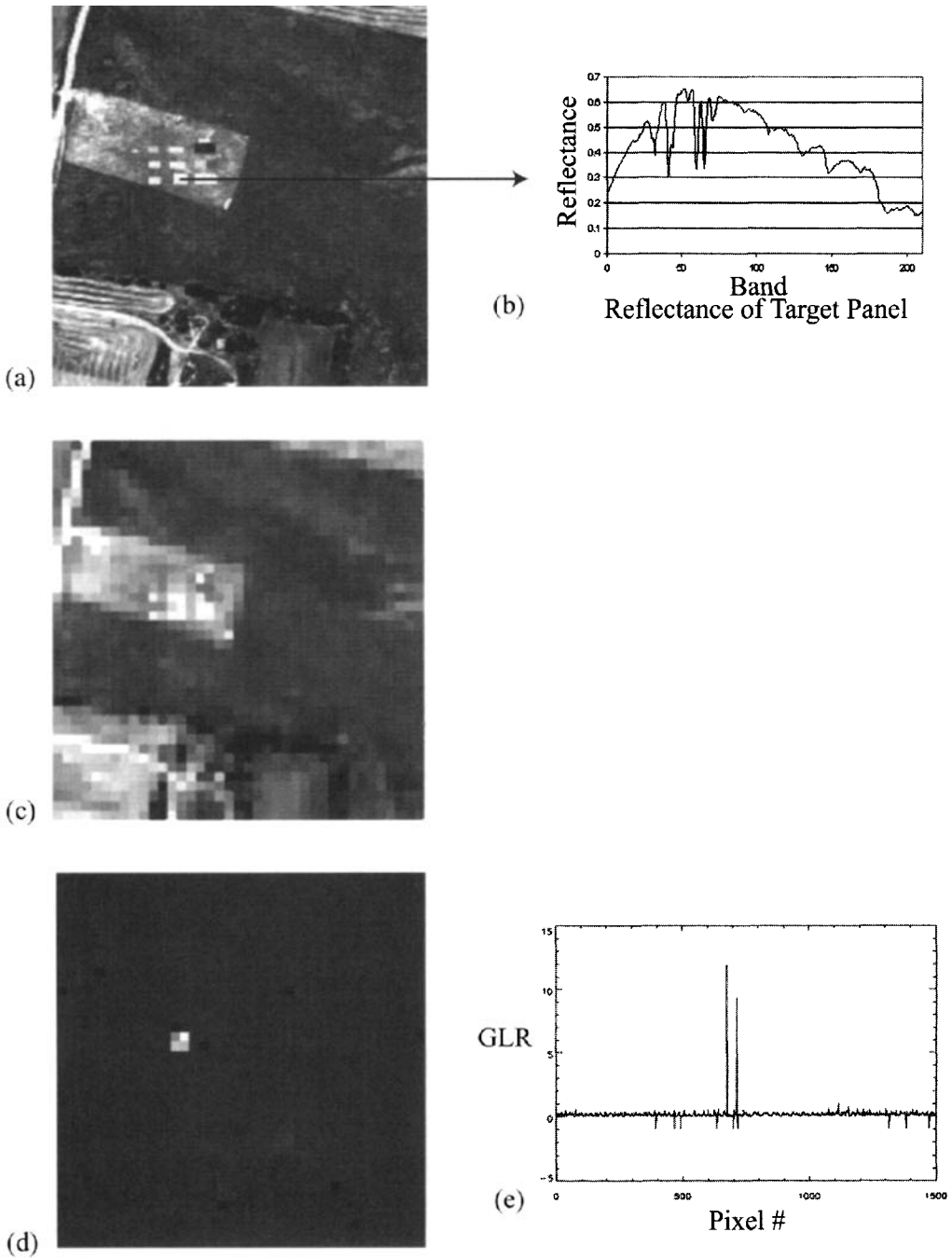


Figure 11.19 Sample results from the subpixel invariant approach applied to a Hydice image of test panels: (a) a single band from the original image, (b) reflectance spectrum of test panel, (c) blurred and resampled image, (d) gray level image of the GLR results showing the four subpixel targets, and (e) a scan profile of the GLR results showing how well separated the targets are from the background.

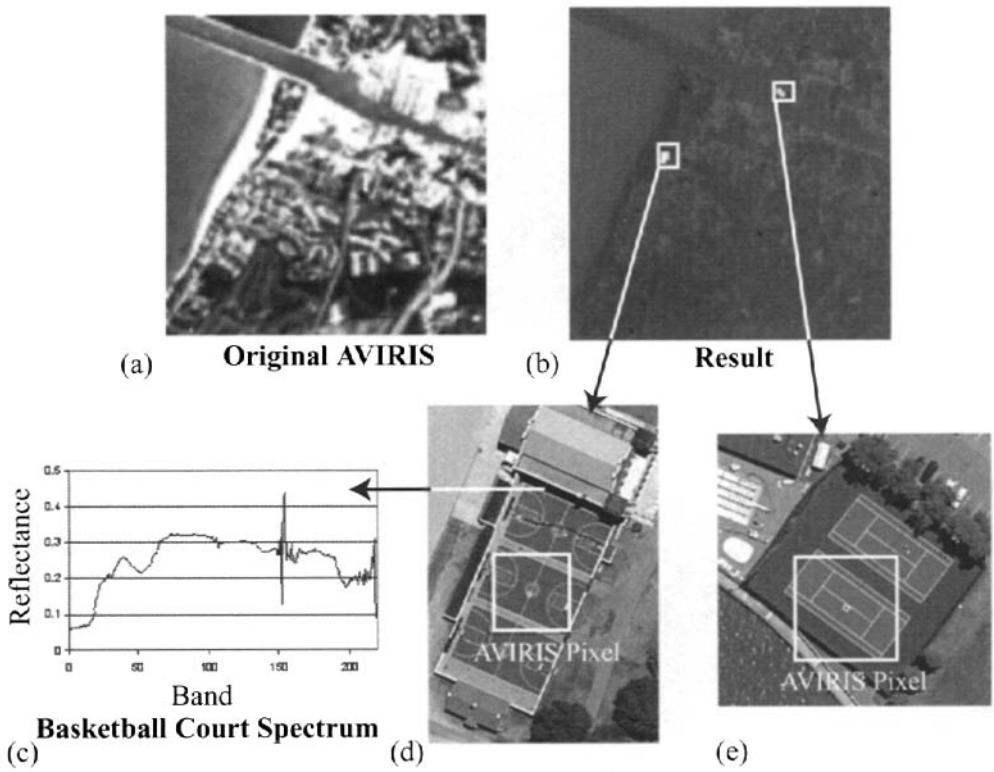
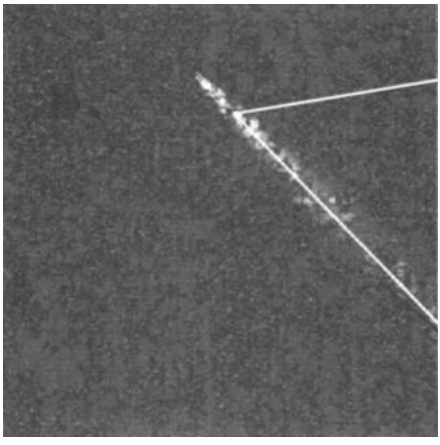


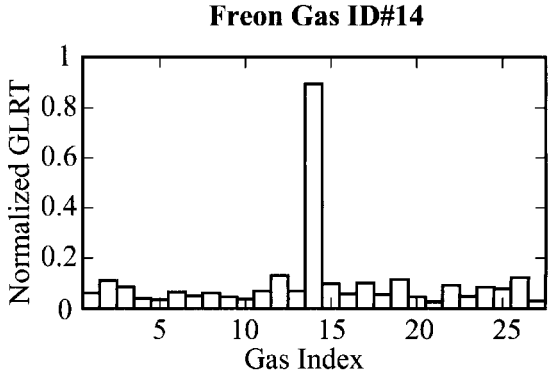
Figure 11.20 Subpixel invariant approach applied to a cluttered scene: (a) a single band from the original image, (b) results of the detection operator, (c) spectrum of the target, (d) high-resolution air photo of the target region showing approximate size of an AVIRIS pixel, and (e) air photo of the second "target" located. See color plate 11.20.

the radiation propagation, a range of estimates of sensor-reaching radiance values can be estimated for each gas of interest. From these values, the max-D approach can be used to generate a set of target basis vectors. Similarly, max-D can be applied to a region of the scene assumed to contain no plumes to generate a set of background basis vectors. Then the GLR operator of Eq. (11.53) can be used to perform detection of the target gas. O'Donnell et al. (2004) suggest an approach where a separate GLR operator is constructed for each gas of interest (they used the top 26 gases in the EPA hazardous gas list). A GLR image is then generated for each gas to form a detection cube. Figure 11.21 shows an example of this process applied to a synthetic image containing a Freon and an NH_3 plume. Figure 11.21(b) shows the target detection image for Freon, along with an example detection vector, for pixels in the detection cube where the band index in the detection vector refers to each of the 26 gases tested. Note that ammonia (NH_3), which is present in the atmosphere, has a poorer background suppression level, but the much denser ammonia plume is still detectible.

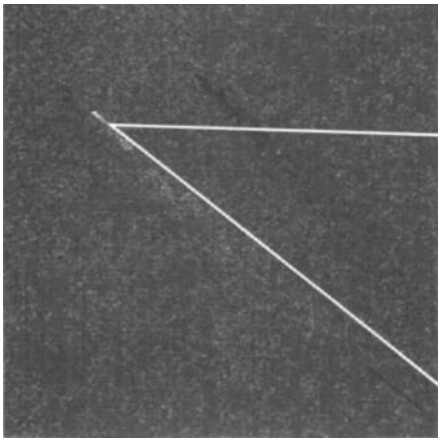
Ientilucci and Schott (2005) describe how the physics-based modeling approach, discussed in this section, can be combined with a geometric version of the infeasibility concept, discussed in Section 10.6, to yield improved target detection through false-alarm suppression.

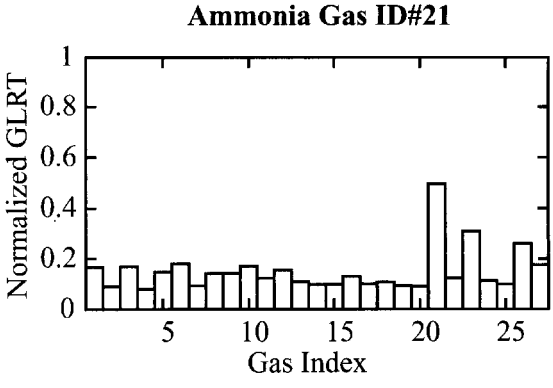
(a)





(b)





(c)

Figure 11.21 Illustration of the physics-based invariant model applied to plume detection: (a) a 10.37 μm band from a DIRSIG synthetic spectral radiance image containing two plumes that can be faintly seen, (b) the Freon detection results and an average of plume pixels showing the strong return in the Freon band (14), and (c) the ammonia (band 21) results and an average of ammonia plume pixel detection spectra.

11.5 INCORPORATION OF PHYSICS-BASED SPATIAL SPECTRAL MODELS INTO ALGORITHM TRAINING

A logical extension of the physics-based modeling approach in support of algorithm development is to use synthetic scene generation tools to predict the signature of targets. This approach has been used to a significant degree for predicting signatures for use in synthetic aperture radar (SAR) automatic target recognition (ATR) algorithms [cf. Douglas et al. (2004)]. At its simplest, this involves making three-dimensional computer-aided design (CAD) models of targets (e.g., vehicles) and then predicting radar cross sections of what the target would "look" like in various orientations. These predictions are then used to train ATR algorithms about how a target signature might appear in actual SAR imagery. Liu and Healey (2003) suggest that a conceptually similar approach might be used for detection of subpixel targets with significant 3-D structure using imaging spectrometer data.

Liu and Healey (2003) begin by generating synthetic hyperspectral images of 3-D targets under a range of poses, backgrounds, and illumination/atmospheric conditions. The images are synthesized using the *Digital Imaging and Remote Sensing Image Generation (DIRSIG)* model described in Chapter 14. The resulting images represent the spectral radiance "image" reaching the sensor, which is then spatially and spectrally convolved to generate spectral vectors simulating what the target or target/background vectors would look like to a particular sensor. These spectral vectors look like a target subspace that can be spanned by a low-dimensional set of basis vectors found by an SVD analysis using the same approach as used by Slater and Healey (1998) in the invariant method (cf. Sec. 11.4). The target subspace can then be described by the matrix **T** made up of the basis vectors as columns. Similarly, a set of background basis vectors can be found using image data from a region known to have no targets. Liu and Healey (2003) suggest that in cases where a target-free region cannot be identified, a preprocessing step to remove target-like pixels may be employed. This would involve projecting image pixels onto the target subspace (**T**) and removing any pixels with large magnitudes according to

$$\|\boldsymbol{L}_T\|^2 = (\boldsymbol{P}_T\boldsymbol{L})^T(\boldsymbol{P}_T\boldsymbol{L}) \tag{11.57}$$

where

$$\boldsymbol{P}_T = \boldsymbol{T}\left(\boldsymbol{T}^T\boldsymbol{T}\right)^{-1}\boldsymbol{T} \tag{11.58}$$

is the projection operator that projects the image vector **L** onto the subspace spanned by vectors making up **T** and $\|\mathbf{L}_T\|^2$ is the squared magnitude of the projection of **L** onto **T**. A set of basis vectors spanning the target-free background subspace can then be found and assembled into a matrix **B**. Finally, a combined target and background subspace **H** can be generated by combining **T** and **B**. At this point, a generalized likelihood ratio test can be employed to detect targets as described in the previous section. This approach essentially represents a subpixel invariant method that incorporates the variability induced by the 3-D shadowing and orien-

tation effects of the target along with atmospheric and illumination effects in the creation of the target subspace. As a result, the detection algorithm should be less sensitive (invariant) to these variations.

In testing the performance of this 3-D invariant method by generation of ROC curves for targets in simulated data, Liu and Healey (2003) compared the GLR test metric with an orthogonal projection ratio (OPR) they describe as

$$OPR = \frac{\|\boldsymbol{L}_H\|^2 - \|\boldsymbol{L}_B\|^2}{\|\boldsymbol{L}_H\|^2 - \|\boldsymbol{L}_T\|^2} \tag{11.59}$$

where L_H, L_B, and L_T are the projections of the image vector **L** onto the combined target-background, background, and target subspaces, respectively [cf. Eq. (11.57)]. This operator is designed to remove the background component of the projection from the numerator and the target component from the denominator to improve overall detection performance. Their results for detection of targets in synthetic imagery showed promising performance for the overall 3-D invariant method with slightly better performance for the OPR detector. Liu and Healey (2004) have extended this basic approach using a nonparametric generalized likelihood ratio.

In closing this section, it is important to recognize that the potential for development and training of algorithms using physics-based modeling approaches has just begun to be explored. The opportunities for increased performance and analysis in this new area at first appear quite promising. However, it is important to recognize two potential limitations of this approach. The first is that the physics-based models must adequately represent the target and/or clutter phenomenology we need to deal with (i.e., we need good models), and the second is that the real and the modeled data sets must end up in common units. Mismatches between real and modeled data can be introduced by inadequate models, inadequate knowledge, or databases to support the inputs to the models and/or calibration errors in the sensor. As a result, while we believe physics-based modeling in support of algorithm development holds much promise, there is much work to be done in developing better modeling tools and defining algorithmic approaches that are less sensitive to some of the sources of mismatch between real and modeled data.

11.6 REFERENCES

Berk, A., Bernstein, L.S., & Robertson, D.C. (1989). "MODTRAN: a moderate resolution model for LOWTRAN 7." GL-TR-89-0122 Spectral Sciences, Burlington, MA.

Borel, C.C. (2003). "ARTEMISS – an algorithm to retrieve temperature and emissivity from hyper-spectral thermal image data." 28th Annual GOMACTech Conference, Hyperspectral Imaging Session, Tampa, FL.

Cox, C., & Munk, W. (1956). Slopes of the sea surface deduced from photographs of sun glitter. *Bulletin Scripps Institute of Oceanography of the University of California*, Vol. 6, No. 6, pp. 401-488.

Douglas, J., Burke, M., & Ettinger, G.J. (2004). High-resolution SAR ATR performance analysis. *Proceedings SPIE*, Vol. 5427, pp. 293-401.

Gillespie, A.R., Matsunaga, T., Rokugawa, S., & Hook, S.J. (1998). Temperature and emissivity separation from advanced spaceborne thermal emission and reflection radiometer (ASTER) images. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 36, pp. 1113-1126.

Gordon, H.R., Tao, D., & Zhang, T. (1997). Remote sensing of ocean color and aerosol properties: resolving the issue of aerosol absorption. *Applied Optics*, Vol. 36, No. 33, pp. 8670-8685.

Green, R.O., Conel, J.E., & Roberts, D.A. (1993). Estimation of aerosol optical depth, pressure elevation, water vapor and calculation of apparent surface reflectance from radiance measured by the Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) using a radiative transfer code. *Proceedings SPIE*, Vol. 1937, pp. 2-11.

Gu, D., Gillespie, A.R., Kahle, A.B., & Palluconi, F.D. (2000). Autonomous atmospheric compensation (AAC) of high resolution hyperspectral thermal infrared remote-sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 38, No. 6, pp. 2557-2570.

Healey, G., & Slater, D. (1999). Models and methods for automated material identification in hyperspectral imagery acquired under unknown illumination and atmospheric conditions. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 37, No. 6, pp. 2706-2717.

Hernandez-Baquero, E.D., & Schott, J.R. (2000a). Atmospheric compensation for surface temperature and emissivity separation. *Proceedings SPIE*, Vol. 4049, pp. 400-410.

Hernandez-Baquero, E.D., & Schott, J.R. (2000b). "Atmospheric and surface parameter retrievals from multispectral thermal imagery via reduced rank multivariate regression." Proceedings of the International Geoscience and Remote Sensing Symposium (IGARSS).

Hook, S.J., Gabell, A.R., Green, A.A., & Kealy, P.S. (1992). A comparison of techniques for extracting emissivity information from thermal infrared data for geologic studies. *Remote Sensing of Environment*, Vol. 42, pp. 123-135.

Ientilucci, E.J., & Schott, J.R. (2005). Target detection in a structured background environment using an infeasability metric in an invariant space. *Proceedings SPIE*, Vol. 5806, pp 491-502.

Jackson, J.E. (1991). *A Users Guide to Principal Components*. Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, NY.

Kahle, A.B., & Alley, R.E. (1992). Separation of temperature and emittance in remotely sensed radiance measurements. *Remote Sensing of Environment*, Vol. 42, pp. 107-111.

Kaiser, R.D. (1999). Quantitative comparison of temperature/emissivity algorithm performance using SEBASS data. *Proceedings SPIE*, Vol. 3717, pp. 47-57.

Lehman, E.L. (1959). *Testing Statistical Hypotheses*. John Wiley & Sons, NY.

Liu, Y., & Healey, G.E. (2003). "Recognizing 3D objects in cluttered backgrounds." *Proceedings SPIE*, Soc. Vol. 5093, pp. 91-96.

Liu, Y., & Healey, G.E. (2004). Using nonparameteric distribution estimates for sub pixel detection of 3-D objects. *Proceedings SPIE*, Vol. 5425, pp. 91-96.

Matthew, M.W., Adler-Golden, S.M., Berk, A., Felde, G.W., Anderson, G.P., Gorodetzky, D., Paswaters, S.E., & Shippert, M. (2003). Atmospheric correction of spectral imagery: evaluation of the FLAASH algorithm with AVIRIS data. *Proceedings SPIE*, Vol. 5093, pp. 474-482.

Mobley, C.D. (1994). *Light and Water: Radiative Transfer in Natural Waters*. Academic Press, Boston, MA.

O'Donnell, E.M., Messinger, D.W., Salvaggio, C., & Schott, J.R. (2004). Identification and detection of gaseous effluents from hyperspectral imagery using invariant algorithms. *Proceedings SPIE*, Vol. 5425, pp 573-582.

Press, W.H. (1986). *The Art of Scientific Computing: Numerical Recipes*. Cambridge University Press, Cambridge.

Raqueño, R., Simmons, R., Fairbanks, R., & Schott, J.R. (2000). "A model-based approach to hyperspectral analysis of water constituents." Proceedings AVIRIS Earth Science and Application Workshop.

Sanders, L.C., Schott, J.R., & Raqueño, R. (2001). A VNIR/SWIR atmospheric correction algorithm for hyperspectral imagery with adjacency effect. *Remote Sensing of Environment,* Vol. 78, Issue 3, pp. 252-263.

Schott, J.R., Lee, K., Raqueño, R., Hoffmann, G., & Healey, G. (2003) "A subpixel target detection technique based on the invariance approach." Proceedings AVIRIS Workshop, NASA-JPL.

Slater, D., & Healey, G. (1998). Analyzing the spectral dimensionality of outdoor visible and near infrared illumination functions. *Journal of Optical Society of America*, Vol. 15, pp. 2913-2920.

Thai, B., & Healey, G. (2002). Invariant subpixel material detection in hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 40, Issue. 3, pp. 599-608.

Vincent, R.K., Rowan, L.C., Gillespie, R.E., & Knapp, C. (1975). Thermal-infrared spectra and chemical analysis of twenty-six igneous rock samples. *Remote Sensing of Environment*, Vol. 4, pp. 199-209.

Young, S.J., Johnson, B.R., & Hackwell, J.A. (2002). An in-scene method for atmospheric compensation of thermal hyperspectral data. *Journal of Geophysical Research*. Vol. 107, No. D24, 4774.

CHAPTER 12
IMAGE/DATA COMBINATION AND INFORMATION DISSEMINATION

One of the most important and often overlooked links in the image chain is the distribution of the information to the user. In many cases, the information is provided in a form where only an imaging expert can use it (i.e., the user sees meaningless blobs) or without sufficient support data to justify meaningful decisions (e.g., a class map with no information on the class accuracies). Entire fields of science are evolving around the topics of information display and scientific visualization. Remote sensing scientists need to take advantage of advances in these fields to assist in the distribution of remotely sensed information. The entire field of *geographic information systems* (*GIS*) has evolved to address methods to merge multiple sources of spatially distributed data. This is a large field that merges the mapping sciences, remote sensing science, and a range of data processing and analysis techniques. We will touch here on only a few points related to basic image display issues and methods for merging various forms of spatial information both for greater information extraction and improved image display.

12.1 IMAGE DISPLAY

In many cases, our output product will be the raw image, or more likely a processed version of the raw image, with the final interpretation done by the user. We must assume that, in general, the user is an applications specialist, not a remote sensing or imaging expert. If the data are improperly displayed, the user may well assume the information is not available from the image data, rather than question the display or processing procedures. It is incumbent on the remote sensing expert to

close the final links in the image chain by making sure that the information in the image is properly presented to the user.

In general, there are two media for image display. Soft copy refers to images displayed on video or computer monitors. Hard copy refers to images printed on photographic transparency or print media (e.g., photographic media). We need to recognize the capabilities and limitations of the display devices, as well as the capabilities and limitations of the human visual system, which is the next link in the image chain. Before we discuss the display media, we will review a few critical features of the *visual response system* [for a more thorough treatment of visual systems in this context, refer to Levine, (1985)].

First (referring to Fig. 12.1), recall that the eye is made up of a simple lens and two arrays of detector elements on the focal plane (retina). Concentrated in a region (fovea) near the visual axis is an array of detectors (cone cells) with each element sensitive to either red, green, or blue flux in a manner conceptually similar to the color CCD cameras illustrated in Figure 6.18(a). The visual response system interpolates the spectral samples into what appears to be a continuous color image. The remainder of the retina is covered with monochrome receptors (rods) that provide low-light response (scotopic response) and peripheral vision. In general, we are concerned with focused observations with adequate lighting (photopic response) where the cells in the foveal region are the dominant receptors.

Visual response is expressed in terms of perceived brightness. The visual response function is extremely adaptive, allowing us to see over many orders of

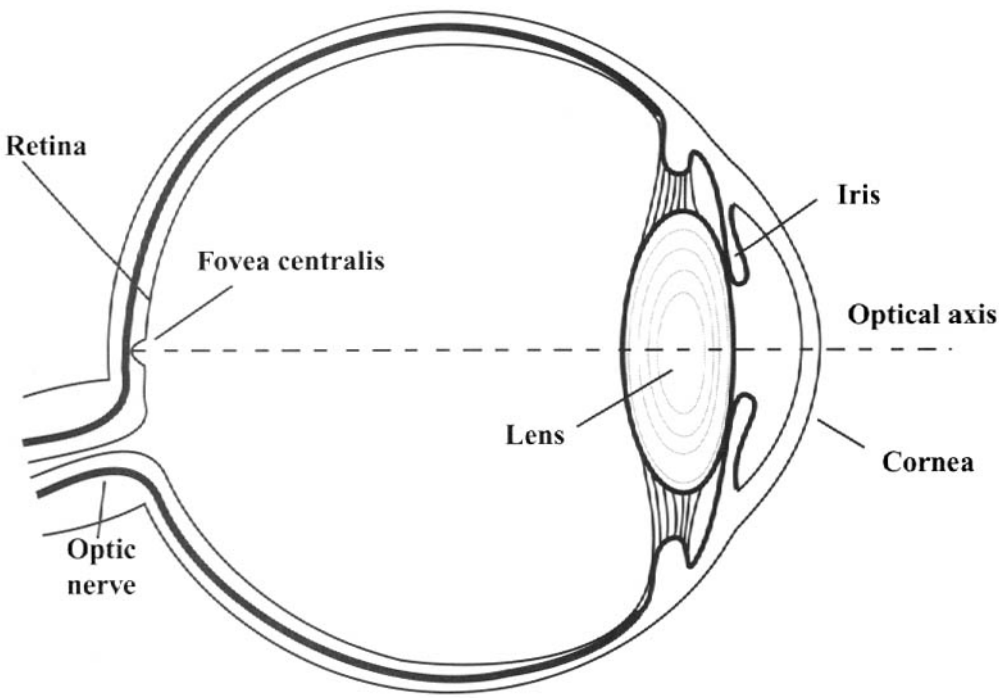


Figure 12.1 Elements of the human eye.

magnitude of illumination levels. However, when adapted to a complex reference level (e.g., as when viewing a scene), visual response is approximately logarithmic and covers a range of about 2.2 log units [cf. Gonzalez and Woods (2002)]. This has three important implications. First, the adaptive nature of visual response means that the eye–brain system is sensitive to differences in flux but not readily able to discern (measure) absolute levels. It is important, therefore, that we make sure that changes in digital count are properly translated into discernible changes in brightness since these are so important to visual analysis. Second, the logarithmic nature of the response means that it takes a larger change in flux in bright regions to generate the same perceived change in brightness that a smaller change would generate in a dark region. Finally, the dynamic range limit at any adaptation level means that the visual system can only perceive a limited radiometric range at a reference adaptation level. If a scene is displayed over a larger brightness range (3.3 log units), the eye would typically adapt to a midlevel, and variations in the brightest and darkest regions would be lost to the visual system.

Clearly, the size and spatial sampling of the visual receptors introduce visual resolution limits typically referred to as *acuity*. An average observer's angular resolution is limited to about 1.7 line pairs per milliradian. This angular resolution can easily be scaled to display resolution when the distance to the display is known, i.e.,

$$\frac{\#\text{ line pairs}}{\text{mm}} = \frac{\#\text{ line pairs}}{\text{milliradian}} \cdot \frac{1}{H} \tag{12.1}$$

where H is the distance from the observer to the display. For example, at 0.5 meters we would resolve approximately 3.4 line pairs per millimeter, which is better than a standard computer monitor (i.e., we would be limited by the monitor, not the observer). On the other hand, if we were observing hard copy with a resolution of 10 line pairs/mm at this same distance, we would be limited by the visual system, and some of the fine detail would not be observed by the user.

This brings us to another limitation of the visual system, namely, closeness of focus. The near point (closest point where the eye can comfortably focus) varies with individuals and often with age, but generally is about 0.2 meters or more. Thus, one can only slightly improve the spatial detail resolved by the eye by moving the image closer. After that, some form of magnification is required.

Unlike photographic and most electro-optical systems, the response of the visual receptors are a function of mean brightness levels, previous illumination conditions, and illumination on adjacent cells. The nerve cells that process the signals from the receptor cells (rods and cones) interact with surrounding cells in a number of ways. The lateral connection of nerve cells in the eye, coupled with additional processing in the visual cortex, results in a variety of unexpected perceptual effects [cf. Cornsweet, (1970)]. The subtle details of visual response effects are too numerous for this treatment, so we will restrict ourselves to the most dominant effect. One of the strongest interactions is called *lateral inhibition*. Each cell reduces the sensitivity of adjacent cells when it is excited (i.e., when the receptor is illuminated). The spatial response to a point source in a cell structure exhibiting

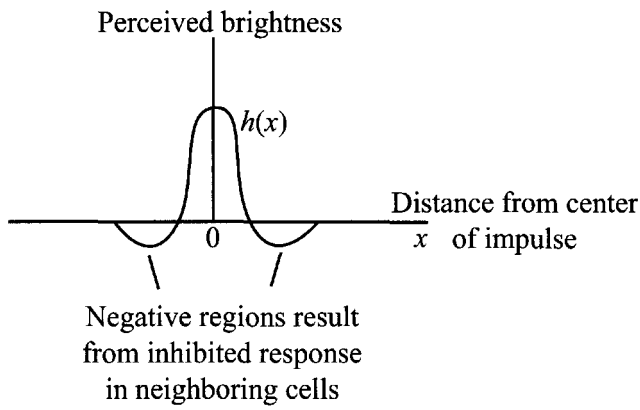


Figure 12.2 Simplified point spread function associated with lateral inhibition.

lateral inhibition is shown in cross section in Figure 12.2. This is called the point spread function of a sensor system (cf. Sec. 13.2). The perceived brightness can be approximated by the convolution of the point spread function with the illumination field onto the retina. Figure 12.3 shows the effect of convolving the point spread function of Figure 12.2 with a step change in brightness. The net effect of lateral inhibition is to exaggerate edges (the Mach band effect). This exaggeration of edges in the lowest level of visual processing is part of the overall cognitive process that keys heavily on edge elements. Thus, in displaying image data for visual analysis, we want to make sure we preserve, or possibly even exaggerate, edges to make them fully accessible to the image analyst's cognitive processes.

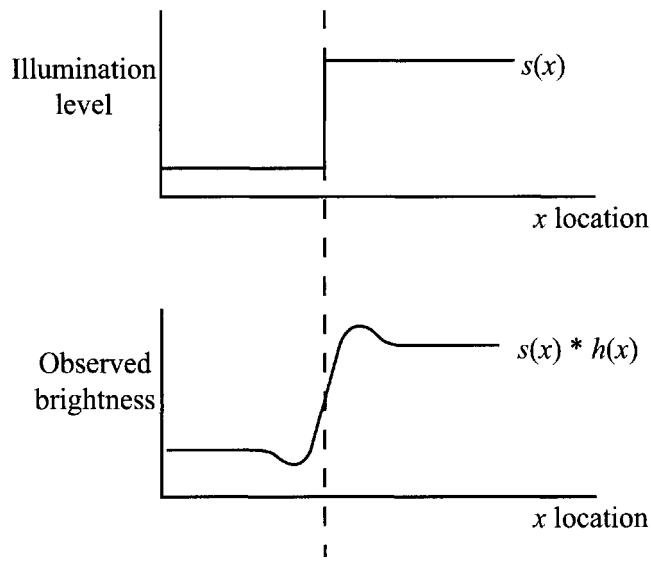


Figure 12.3 Convolution of the point spread function of Figure 12.2 with a step in illumination.

To this point we have ignored color vision, which is also a science in its own right. Our main concern will be in recognizing that the visual system responds to a range of spectral stimuli that can be described in a variety of color spaces [cf. Robertson (1992) and Fairchild (1998)]. The *International Commission on Illumination* (*CIE*) has defined one of the most commonly used color spaces based on the use of *tristimulus* values derived from color-matching functions designed to approximate the effective response of the human visual system [cf. Wyszecki and Stiles (1982)]. These color-matching functions (plotted in Fig. 12.4) are not the actual response functions of the red, green, and blue photo receptors but a set of weighting functions (based on the CIE 1931 Standard Observer) incorporating the average observer's perceived response to spectral radiance. By cascading the color-matching functions together with the incident spectral radiance, three tristimulus values are derived according to

$$\begin{aligned} X &= k\int L_\lambda x'(\lambda)d\lambda \\ Y &= k\int L_\lambda y'(\lambda)d\lambda \\ Z &= k\int L_\lambda z'(\lambda)d\lambda \end{aligned} \tag{12.2}$$

where X, Y, and Z are the tristimulus values; x', y', and z' are the color-matching functions shown in Figure 12.4; k is a constant used for units conversion; and the integral is over the visual response range. The Y tristimulus value is proportional to the overall luminance. For convenience of plotting color in a two-dimensional space, the x and y *chromaticity coordinates* are defined as

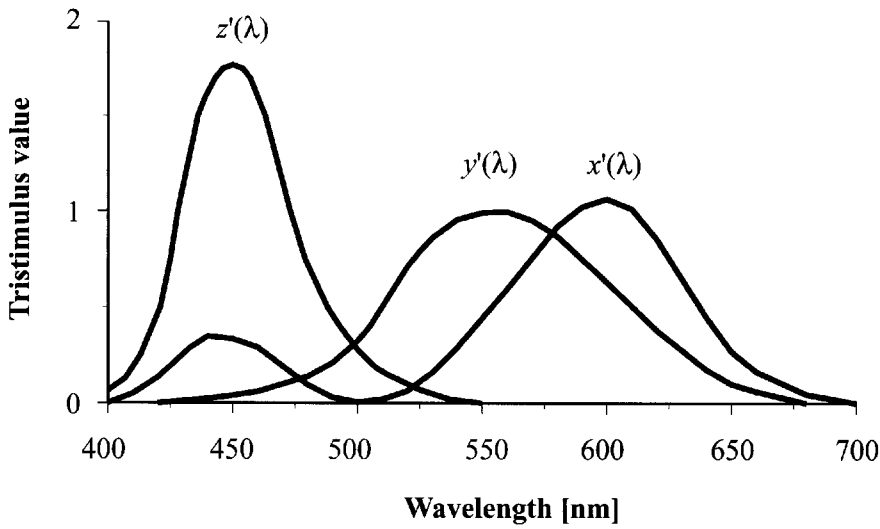


Figure 12.4 Color-matching functions.

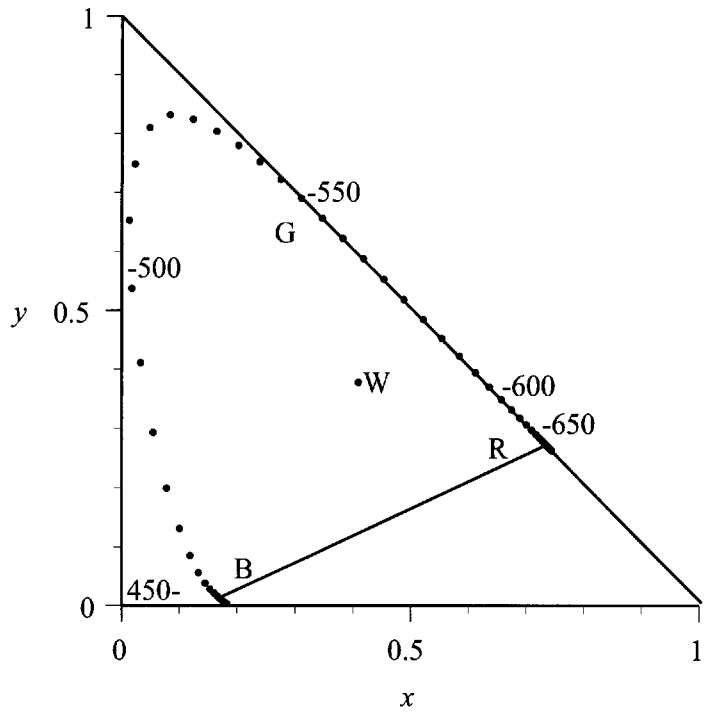


Figure 12.5 CIE 1931 chromaticity diagram; R, G, B, and W are the red, green, blue, and white regions, respectively.

$$x = \frac{X}{(X+Y+Z)}$$

and

$$y = \frac{Y}{(X+Y+Z)} \tag{12.3}$$

with any perceived color uniquely specified by (Y, x, y), where Y represents the luminance, and x and y provide what we think of as the color characteristics (hue and saturation). The color space is then characterized by plotting x versus y as shown in Figure 12.5. This is known as the CIE 1931 chromaticity diagram. The horseshoe-shaped region is constructed by computing the chromaticity coordinates for monochromatic flux at the indicated wavelengths and connecting the points to form the horseshoe shape referred to as the spectrum locus. The spectrum locus delimits the range of visually perceived color. The region inside the spectrum locus is referred to as the visual color gamut. Any given display system will be capable of displaying some subset of this gamut, as shown in Figure 12.6. Thus, the color gamut of the display system may limit the amount of information we can provide to the visual system and influence how we choose to display it. It is also important to recognize that all points in a color gamut are not perceived as unique colors. Surrounding any point in the chromaticity diagram, there is a region where changes in the chromaticity values are not noticeable as color changes. For display purposes, we want to make sure that if we want to show steps in color space, the step sizes

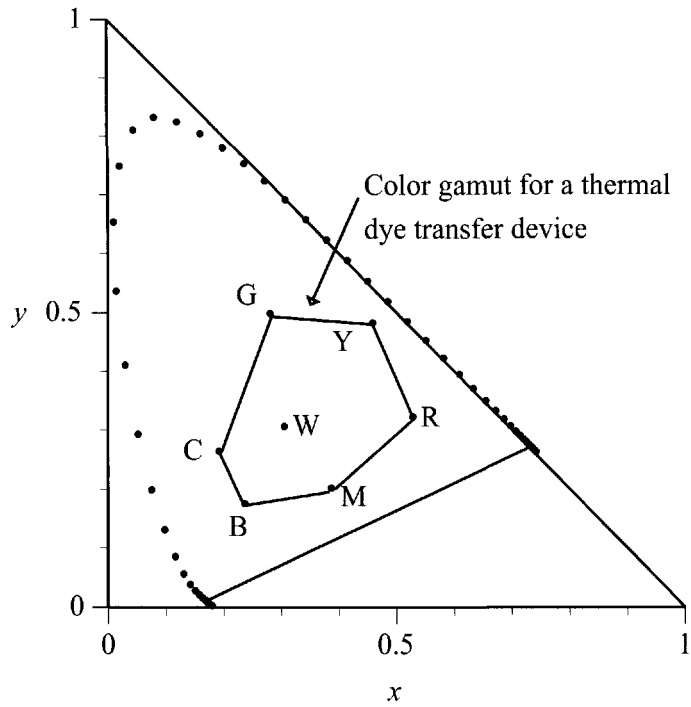


Figure 12.6 Chromaticity diagram showing the color gamut of an output device. Only colors inside the color gamut boundaries can be produced by this device. (Data courtesy of RIT's Munsell Color Science Lab.)

must be such that they exceed these *just noticeable difference* (*JND*) steps that were characterized in chromaticity units by MacAdam (1942).

When displaying images for hard or soft copy analysis, we need to factor in the visual response function. To accomplish this, the relationship between digital count and radiance from soft-copy displays needs to be calibrated. In the case of hard-copy displays, the relationship is usually between digital count and reflection or transmission density. For visual assessment, we would generally like to have this relationship approximately log linear with brightness (or linear with density since this is a log function). We also need to ensure that significant changes in digital count (i.e., changes greater than the overall system noise) translate into changes in brightness that are greater than the JNDs of the visual system. In cases where the display system is not readily adjusted, this is often best accomplished by adjusting the digital count driving the display using a lookup table approach (cf. Sec. 8.1). An example of this is shown in Figure 12.7, where a density versus digital count (DC) calibration curve is shown in the first quadrant and a lookup table used to linearize the relationship between image digital count and density is developed in the remaining quadrants. Similar display corrections need to be developed for color displays to ensure that the available color gamut is effectively used and to make sure that we don't try to display outside the color gamut. The background lighting conditions at the analyst's station can be very critical here, as they can have significant impact on the appearance and interpretation of color, as well as

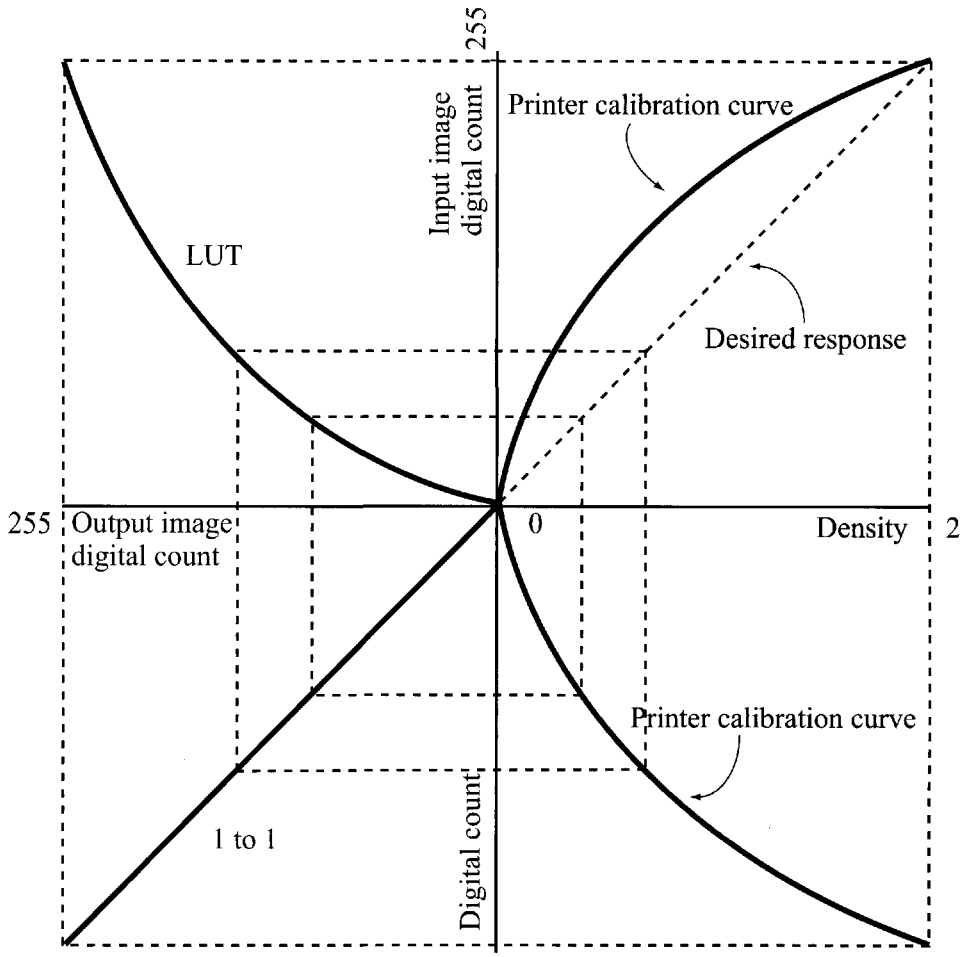


Figure 12.7 Calibration and correction of output device using a lookup table (LUT) to linearize the output.

affecting the ability to detect subtle gray-level variations. Where possible, these illumination factors should be included in the calibration of the output device [cf. Hunt, (1987)].

Resolution effects will be discussed in more detail in the next chapter. So for the moment, we will restrict our concern to ensuring that the observer is able to "see" all the resolution available in the data if required. For soft-copy displays, this means that the user should be able to zoom in to a point where the visual resolution is better than the resolution of the image displayed. Similarly, for hard-copy display, the observer should be able to magnify the display to a point where visual resolution is no longer a limiting factor. The reader should note, however, that most analysis will not be done at these extreme resolutions, and the synoptic perspective afforded by lower resolutions and large fields of view is also highly desirable. Furthermore, providing the user with a capability to observe the image at magnifications much beyond where the visual resolution matches the display

resolution can introduce unwanted and distracting artifacts (e.g., grain noise in the film can begin to be interpreted as image information). The ideal system would allow the user the full synoptic perspective at the maximum resolution. A practical compromise is to provide full synoptic coverage at reduced resolution and an ability to zoom in on localized regions to a point where the visual resolution is the limiting factor.

12.2 THEMATIC AND DERIVED INFORMATION

In many cases, the user needs information derived from the image rather than, or in addition to, the image itself. This is commonly thought of as *thematic* information where the image data have been processed to represent some form of spatial theme(s), such as land cover, vegetation stress, or water quality. Here again, it is critical that we ensure that any processed information is available to the analyst. For example, we will often color code themes (e.g., a land cover map), and we need to ensure that the colors selected are within the color gamut of the output device and clearly differentiable (i.e., several JNDs apart in color space). Furthermore, for both esthetics and ease of interpretation, it is often valuable to select color or gray shades in thematic maps that are intuitively meaningful to the user (e.g., blue water).

Another important and often overlooked aspect of conveying thematic information to the user is to recognize that decisions must be based on this information. It is often important to provide not only the results (e.g., the thematic map) but also information on the quality of the results so the user can decide how heavily to weigh them into any decision. This can be done, for example, by providing a confusion matrix of classification accuracies on independent data (cf. Sec. 9.2). In many cases where statistics are used in the processing algorithm (e.g., a GML classifier), it is even possible to provide confidence data on a pixel-by-pixel basis. For example, in a GML classifier, the a posteriori probabilities can be calculated and scaled so that they can be presented in image form with brightness varying with probability.

Recognize that the user needs this same type of information even when the output is in the form of numerical or tabulated data. For example, suppose an algorithm identifies 20% of a study region as forest and 17% of the forest area as dead or severely stressed due to an insect infestation. If possible, we should also provide the user with information on the confidence of these estimates, e.g., at the 95% confidence level the study region is $20.0 \pm 0.5\%$ forested, and of that forested region, $17 \pm 8\%$ is dead or severely stressed. As a result, it is important for the remote sensing scientist to always understand the algorithms being used and to perform sufficient analysis to determine what confidence should be placed on any results.

12.3 OTHER SOURCES OF INFORMATION

It is all too easy for experts to become myopic and neglect sources of information other than those with which they are most familiar. The remote sensing scientist is most often involved in extracting information from images. This is one very critical element of spatial information analysis. However, there are many other possible sources of information that may be helpful in addressing an information analysis problem. The whole field of geographic information systems (GIS) has evolved to address the problems of combining and analyzing spatial information. In this section, we will discuss a few of the fundamentals of GIS and point out other sources of information and analysis tools a remote sensoring scientist should look to for assistance. A more thorough treatment of GIS concepts can be found in Goodchild and Kemp (1990).

12.3.1 GIS Concepts

Geographic information systems (GIS) loosely refer to both databases and data processing tools used to analyze spatial data. Normally, the data are tied to some geographic reference system, so that each piece of information refers to a point or a region on the Earth. In a GIS, many types of information exist that describe spatially distributed variables. For any location covered by the GIS, all the attributes (the value of the spatial variables) can be accessed, or conversely, any attribute can be mapped to a spatial location or displayed as an image. Most important, algorithms can be implemented that allow processing of any or all of the attributes in order to search for certain combinations or relationships.

12.3.1.1 Spatial Data

Before we worry too much about GIS processing logic, we should briefly consider some of the forms of spatial data that may be of use in remote sensing and image analysis. To begin, each image is itself a set of spatial data that could be attributed on a pixel-by-pixel basis to the location on the ground where the pixel would be projected. Similarly, any image-derived data could be attributed to little parcels of real estate (ground sample spots) in the same fashion (e.g., land cover type). Another way to think about this process is to imagine the boundaries of thematic regions projected onto a map and the enclosed region assigned a theme (e.g., land cover type). This introduces the two primary GIS data formats: *raster* (pixelized data) and *vector* (points, line segments, and enclosed polygons), as illustrated in Figure 12.8. Different images or data derived from different images (e.g., thermal, multispectral, radar) can all be combined by having the data spatially registered to each other or to a common geographic coordinate system. Other types of spatial data can also be used to characterize the same piece of real estate represented by the image data. Any type of map data, for example, can be merged into a common GIS based on geolocation. This could include soils maps, road maps, and topographic data in the form of elevations, or derived data in the form of slope and aspect maps.

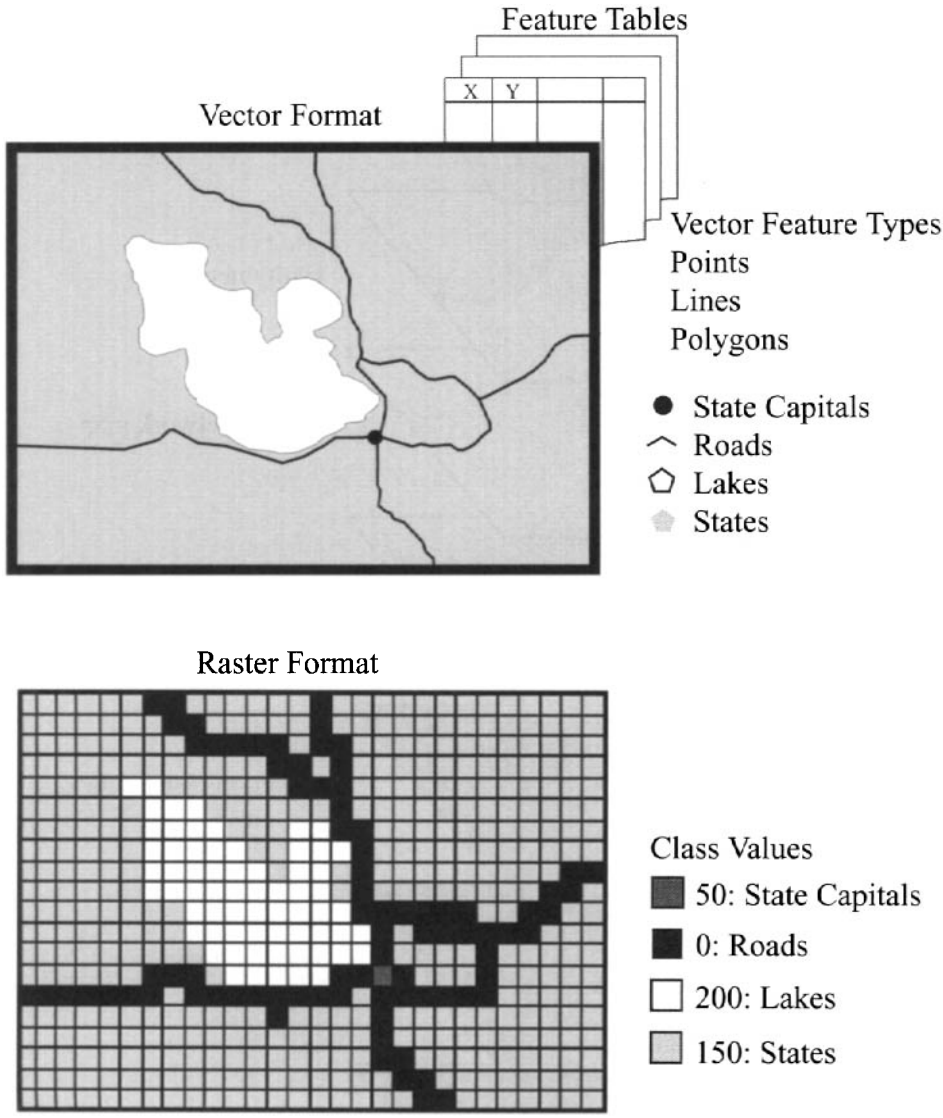


Figure 12.8 GIS data formats.

It is also possible to include geo-referenced point or line information such as the location of a pollution source or a telephone line with an information file describing the attributes of the point or line. For most of us used to working with images, the database is most easily thought of in raster form, with all of the features (or layers) in the database conceptualized as spectral bands or feature bands in a many-layered image (cf. Fig. 12.9). It is important to recognize that the resolution of the layers

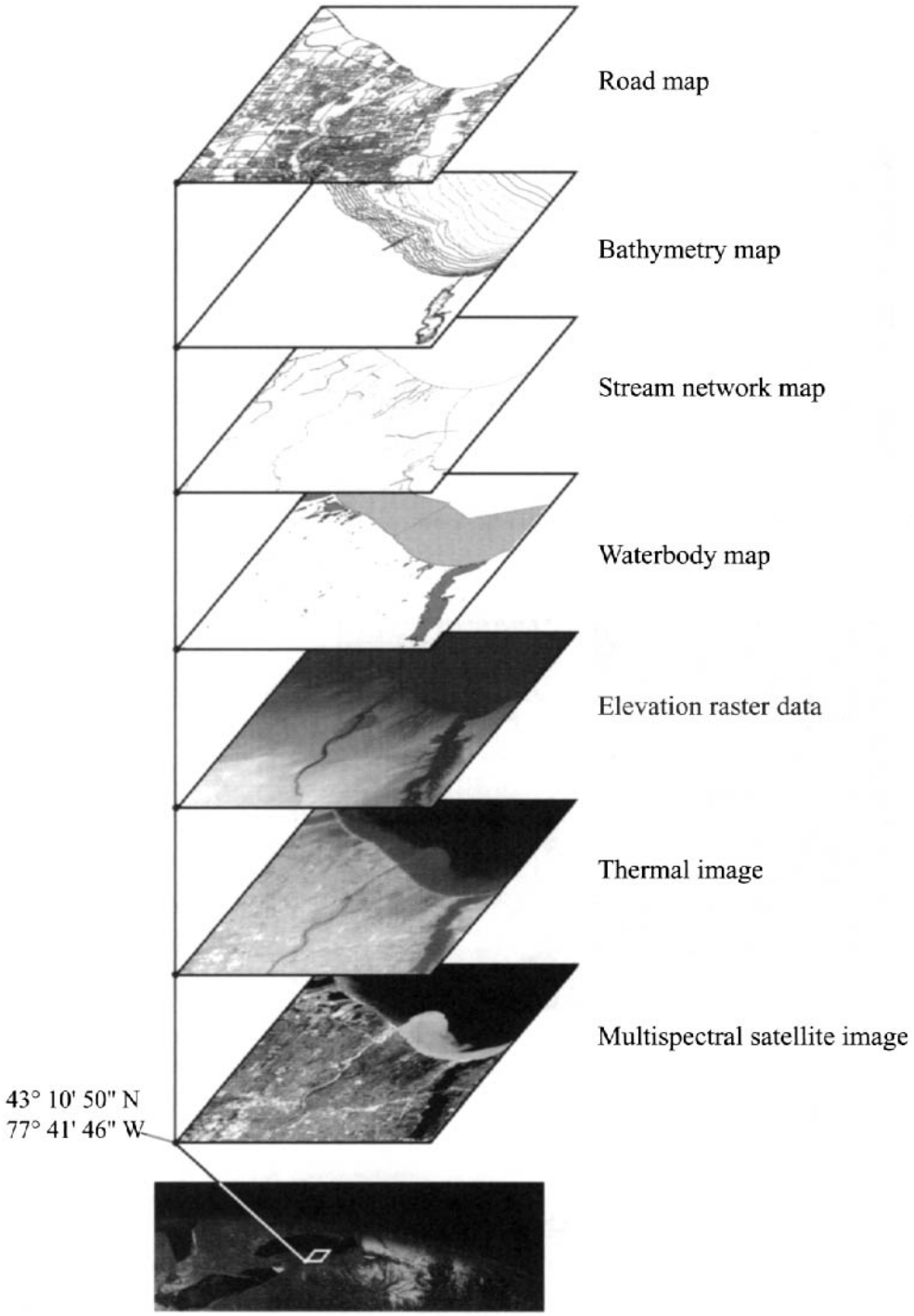


Figure 12.9 Layers in a GIS database.

need not all be the same. The lower resolution data can be resampled to a higher pixel density when necessary for comparison with the higher resolution data.

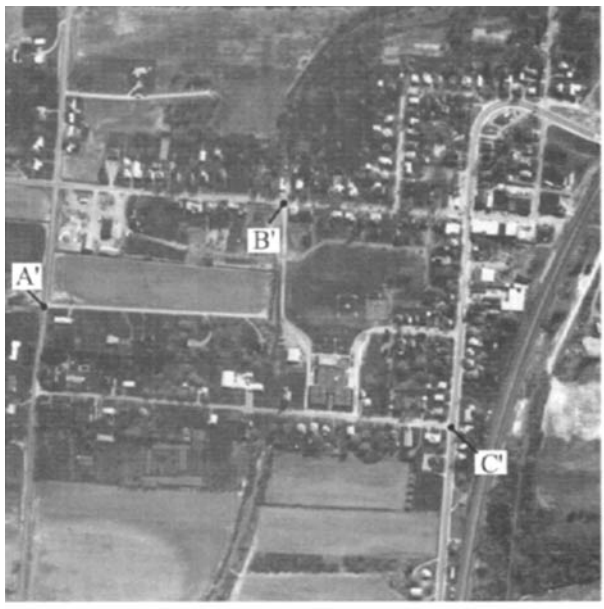
12.3.1.2 Registration and Resampling

The formatting and processing of GIS data rely very heavily on our ability to transfer spatial data into a common coordinate system (*registration*) and to re-sample the data so that we can easily access and process information from the same spatial location simultaneously. To begin to understand image registration, we will first take the case of two images of the same region. For convenience, we assume that the images are of relatively flat terrain so that image distortion due to terrain elevation is negligible. The images may, however, have different resolution, rotation, view angle, etc., so that pixel locations in one have no clear relationship to pixel locations in another. We desire a way to relate the geometric coordinate system in one image to that of the other image and eventually to transform one of the images so it has a common coordinate system with the other.

In the case of images, it is possible to use arbitrary coordinate systems, with pixel row and column values representing the location of pixel centers in a pair of Euclidean coordinate systems, as indicated in Figure 12.10. We use x', y' to designate points in the primary or target coordinate system of the reference image and x,y to designate points in the sample image coordinate system. We would like to warp the sample image so that it registers with the reference image. To accomplish this, we need to know the relationship between the two coordinate systems. We assume that the systems can be approximately related by a least-squares fit to a polynominal of the form

$$\begin{aligned} x &= a_0 + a_1 x' + a_2 y' + a_3 x' y' + a_4 x'^2 + a_5 y'^2 \cdots + \varepsilon_x \\ y &= b_0 + b_1 y' + b_2 x' + b_3 y' x' + b_4 y'^2 + b_5 x'^2 \cdots + \varepsilon_y \end{aligned} \tag{12.4}$$

where ε_x and ε_y are the residual errors after the transform. For simple distortions between the two images, a good transform can be achieved with a fairly low-order transform. For example, only the zeroth–order terms a_0 and b_0 are needed for a simple shift of the origin, and the first two terms (a_1, a_0, b_1, and b_0) are needed for a combined scale adjustment and shifting of the origin. Higher order terms are required to account for rotation, skew, and keystoning effects due to acquisition and perspective differences [cf. Schowengerdt, (1983)]. A coordinate transform of the form of Eq. (12.4) can only be assumed valid for images or portions of images that have good internal rectilinear geometry. For example, scanner images with residual roll distortion should be corrected for roll before use of Eq. (12.4). In scenes with considerable topographic variation, the simpler forms of Eq. (12.4) are only approximately valid for surfaces that approximate a plane. In rough terrain, piecewise approximations must be applied across the image using a different solution to Eq. (12.4) for each region (cf. Fig. 12.11). To avoid error in boundary regions, Eq. (12.4) must be truncated to four terms if quadrilaterals are used, or three terms if triangles are used. This produces a less precise but unique solution at all boundaries between solution regions. In cases where multiple transformations must be



(a) Target on reference image and coordinate system.

Control point coordinates

	x', y'			x, y	
A':	52	382	A:	9	91
B':	358	254	B:	332	157
C':	574	540	C:	344	518

(b) Sample image and coordinate system.

Figure 12.10 Coordinate systems for geometric registration.

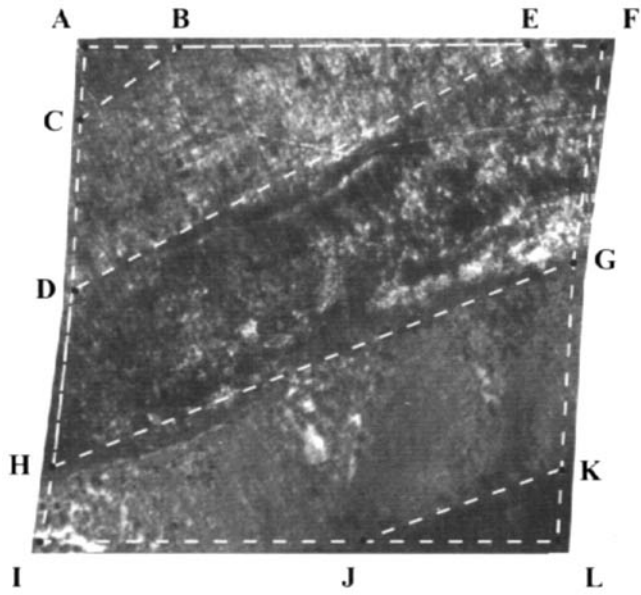


Figure 12.11 Use of quadrilateral regions in a piecewise geometric rectification for regions with high-terrain relief.

sequentially applied (e.g., if we have to roughly prewarp one image to another to facilitate a more rigorous automated registration), it is often convenient to restrict the transformation to affine (i.e., 2-D linear) transforms of the form

$$\begin{aligned} x &= a_0 + a_1 x' + a_2 y' + \varepsilon_x \\ y &= b_0 + b_2 x' + b_1 y' + \varepsilon_y \end{aligned} \tag{12.5}$$

which can be expressed in matrix form as

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix} + \begin{bmatrix} a_0 \\ b_0 \end{bmatrix} = \mathbf{T} \begin{bmatrix} x' \\ y' \end{bmatrix} + \mathbf{T_0} \tag{12.6}$$

or, more conveniently for cascading transformations,

$$[x, y, 1] = [x', y', 1] \begin{bmatrix} a_1 & b_2 & 0 \\ a_2 & b_1 & 0 \\ a_0 & b_0 & 1 \end{bmatrix} = [x', y', 1]\mathbf{A} \tag{12.7}$$

Linear transforms of this form can be cascaded together (via matrix multiplication) to obtain the final transform. Example transforms using this 3 × 3 matrix representation are shown in Figure 12.12 using the notation of Wolberg (1990). This means that the final images only need to be warped once, reducing the sampling errors that would be introduced by multiple resampling steps.

In order to apply any transform equation, common objects must be uniquely located in each image. The x,y and x',y' values of these control points generate the input data to a least-squares regression that is used to solve for the a_i and b_i coefficients. As with any regression solution, the input data should cover the entire solution space, and in general, the solution should not be extended beyond the sample space. This is particularly true when higher order terms are used in the solution for a whole-image transform. Severe distortion can occur beyond the sample space (for this reason, lower order solutions should be used whenever possible, and the user should make sure that sufficient control points are selected to ensure a robust solution).

The control points used for input to a coordinate transform can be manually located in each image. However, this can be a tedious, time-consuming task. A number of methods have been developed to automate the selection of control point pairs. One of the most straightforward involves the use of correlation; [cf. Eq. (8.8)]. A small window in one image is selected as a correlation kernel, and its normalized correlation with the other image is computed as

$$g(i,j) = \frac{1}{K_1 K_2} f(i,j) \bigstar h(i,j) \tag{12.8}$$

where K_1 is the sum of the values in the correlation kernel, K_2 is the sum of the values in the image under the correlation kernel, $f(i,j)$ is the image digital count, $h(i,j)$ is the correlation kernel and ★ is the correlation operator. Where the kernel passes

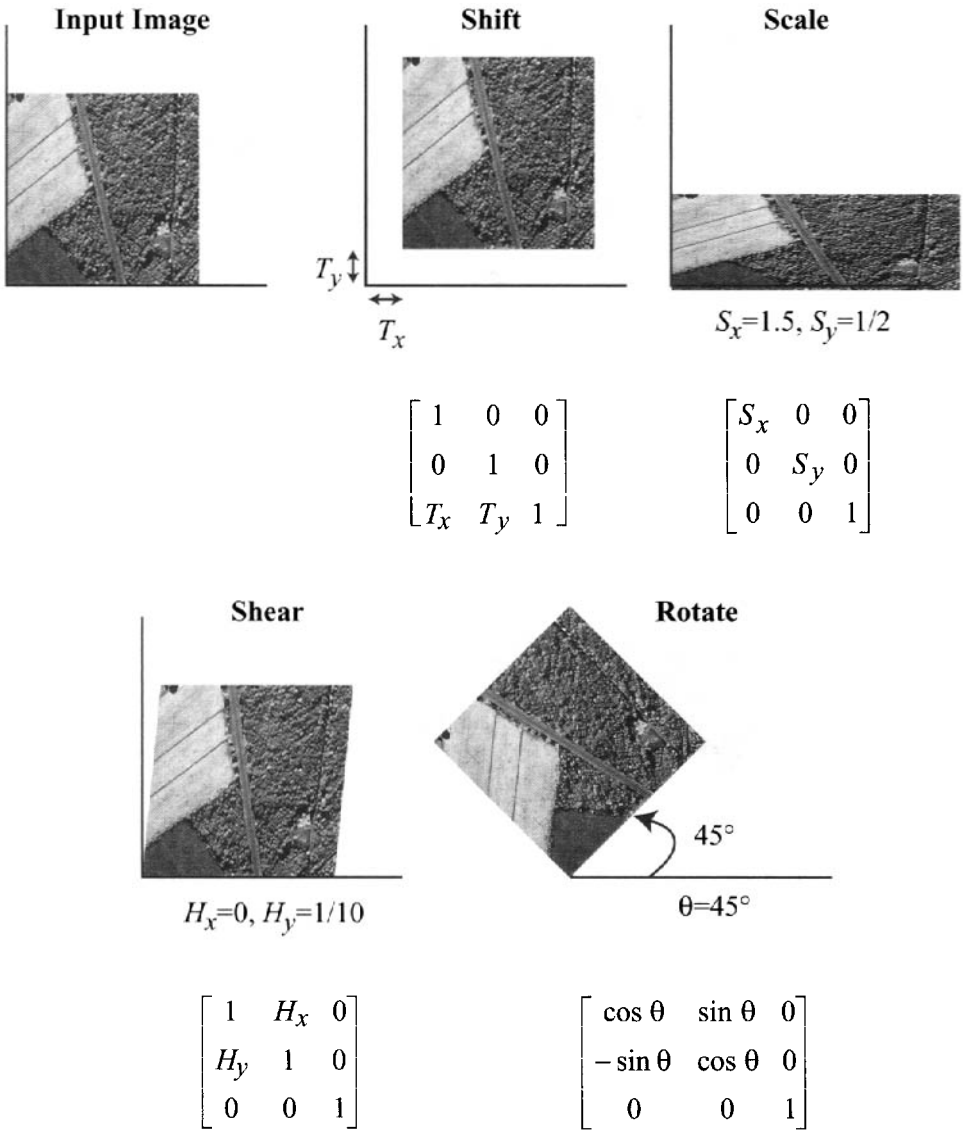


Figure 12.12 Examples of 3 × 3 matrix representation of affine transforms.

over the corresponding region in the search image, a maximum in the normalized correlation should occur. The original coordinates of the center of the correlation kernel and the location of the maximum in the normalized correlation image become a pair of control points.

Regrettably, the correlation value is sensitive to scale and rotation, so the images must be roughly scaled and rotated in order to use this approach. Most satellite images have had first-order geometric corrections applied, so this is often not a concern. To reduce the likelihood of finding a false maximum and to speed up the process, the search region for the correlation is often restricted to a region delimited by the expected error in nominal registration between the two images. For example, if the two-sigma (two standard deviations) pointing error in a satellite translates to 30 pixels in the image, we would expect a pair of control points

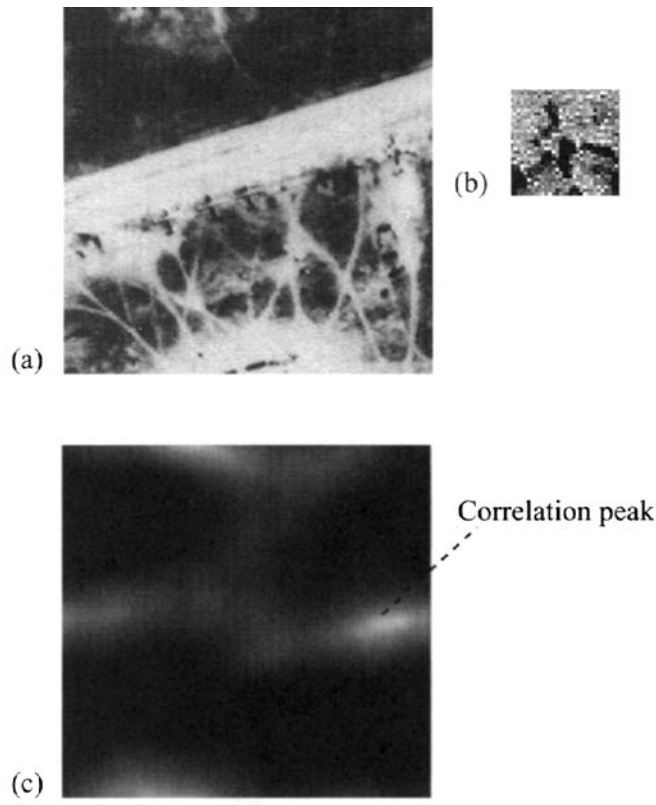


Figure 12.13 (a) search window, (b) image kernel, and (c) correlation surface. The peek in the correlation shows where the image kernal matches the brightness structure in the search window.

on two images from the satellite to be located within ±60 pixels of each other (i.e., the images should be registered to first order to within 60 pixels). Thus, the search window need only be 60 × 60 pixels rather than the entire scene. Furthermore, the correlation kernel (control point) can be selected to have high contrast and sharp edges to improve the sharpness of the correlation peak and, therefore, the precision of the conjugate control point location. An example of a control point kernel, search window, and correlation surface is shown in Figure 12.13. Many control points could be automatically located in this fashion and used as input to a regression solution to an equation of the form of Eq. (12.4). Regrettably, the control point selection process includes errors, some of which can be quite large (e.g., complete mismatch of points). Therefore, a screening process is often employed aimed at rejection of individual control points with excessive error.

Once the geometric relation between the two images is known, a method for mapping the image values from one coordinate system to the other must be defined. For each x',y' pixel center in the reference coordinate system, we must determine what value to assign from the input sample image. Using Eq. (12.4), we can locate the x,y location in the input image coordinate system corresponding to x',y'. Generally, this x,y value will not be an integer value falling exactly on a pixel center. A simple method of selecting the appropriate digital count value is to

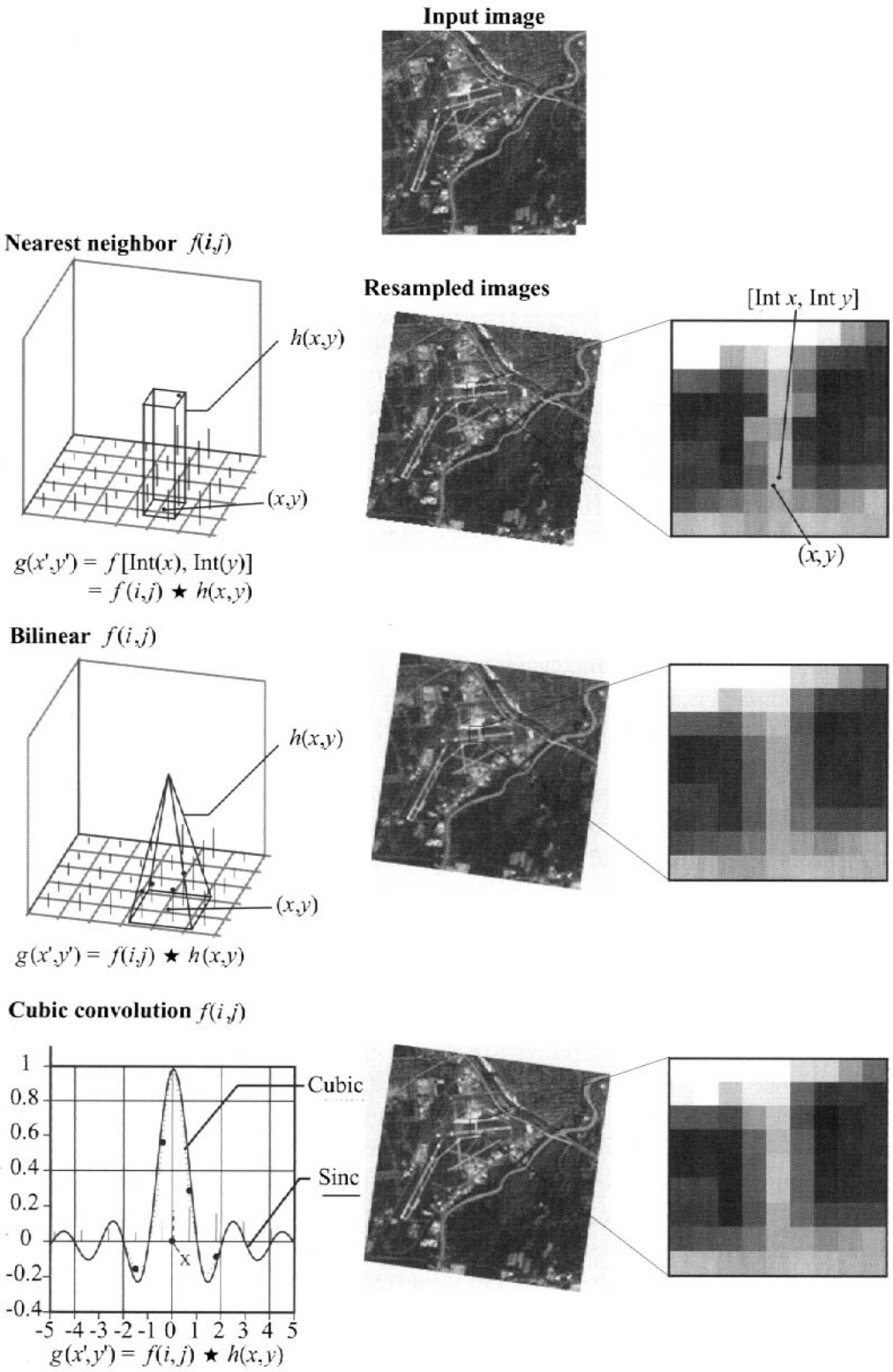


Figure 12.14 Resampling methods used in image registration.

select the value of the pixel center closest to the x,y location. This method of *nearest-neighbor resampling* can be easily implemented and is computationally very quick. The nearest-neighbor resampling method can introduce staircase artifacts at edges, as shown in Figure 12.14. This effect can be reduced using a simple bilinear resampling approach, as illustrated in Figure 12.14. *Bilinear resampling* reduces the edge artifacts but tends to blur the image slightly. Also, since it depends on four neighboring pixels, its computation time runs slightly longer than the nearest-neighbor method.

The resampling process can be thought of as convolving the image with a kernel that is centered at each consecutive x,y location in the input image corresponding to the x',y' values in the output image. The nearest-neighbor kernel would have a base of 1 pixel by 1 pixel and unit height. Thus, it will always just return the value of the 1 pixel located under the kernel. The bilinear interpolator is a four-sided pyramid kernel of base 2 pixels × 2 pixels and central height of unity. Linear systems theory indicates that the ideal sampling kernel would leave all frequencies unaffected (i.e., a rect function in frequency space), which means the convolution function should be a sync function centered at the (x,y) sampling location [cf. Eq. (13.9) and (13.10)]. Use of the sync function is impractical as it would require sampling the entire image, as shown in Figure 12.14. Instead, it is approximated with a cubic polynominal fit that uses the nearest 4 pixels in each direction as shown in one dimension in Figure 12.14. This *cubic convolution resampling* approach is a good approximation to the ideal sync function, producing less blur than the bilinear interpolator, but at the cost of increased run time [cf. Park and Schowengerdt, (1983)].

It is important to recognize that any resampling other than nearest neighbor introduces some blurring of pixel radiance values and some mixing of spectral signatures in a multispectral image. For this reason, where it is practical, it is often desirable to perform radiometric calculations before geometric transforms. For example, a multispectral land cover classifier could be run on the raw image, and then the class map image could be transformed.

To this point, we have concentrated on image-to-image registration. However, exactly the same principles apply to warping an image to a map. The control points in this case are map features that can be clearly identified on the image, and the reference coordinate system is the map coordinate system. To perform the resampling, a grid size is defined in the map coordinate system, and the image location is computed for each grid center, just as in image-to-image resampling. Any of the resampling methods can be used. Using this method, any image or digital map data can be transformed to a common coordinate system. Note that more sophisticated 3-D image warping can also be used to project image data onto digital elevation models, as shown in Figure 12.15 and subsequently onto any plane of interest to generate orthophotos (when projected onto the ground plane), as shown in the illustration, or the popular fly-through visualizations.

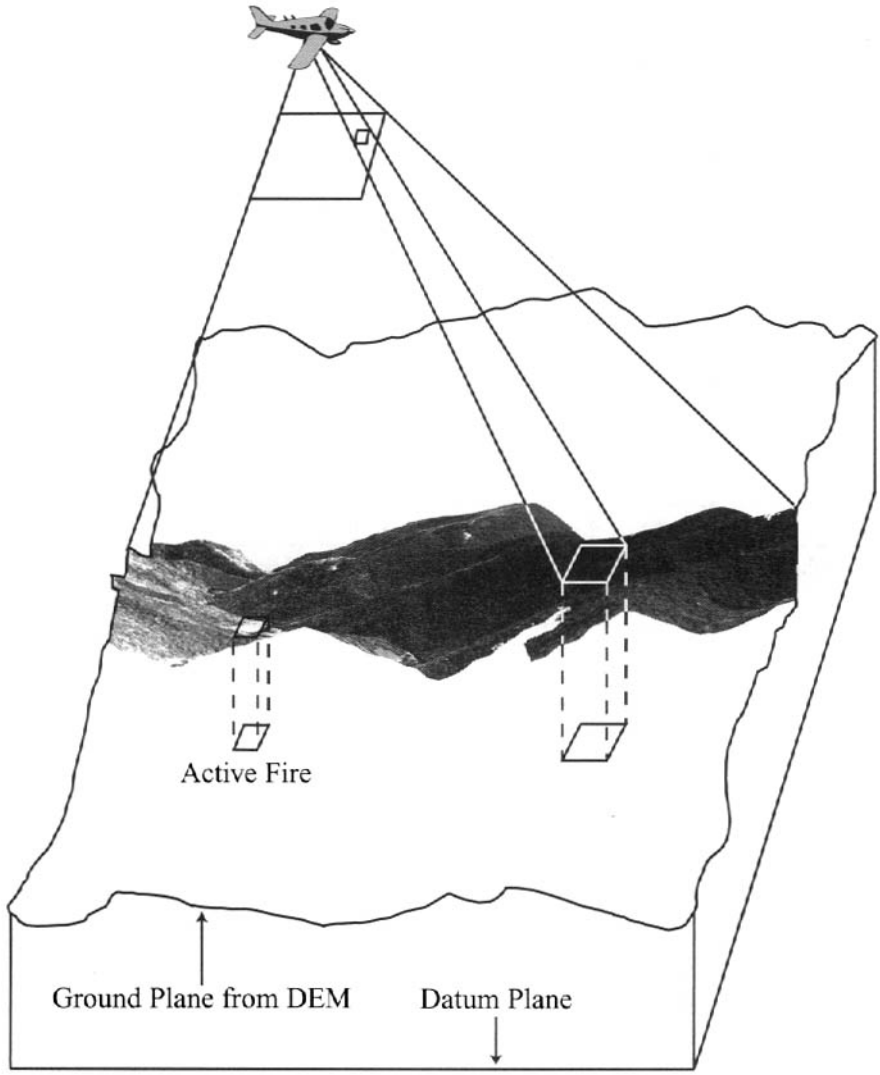


Figure 12.15 Illustration of how image warping can be combined with digital elevation model (DEM) data to generate perspective imagery or in this case an orthophoto. This example is of thermal infrared data of a fire being projected into map coordinates to support fire control activities (imagery from RIT's WASP camera system).

12.3.1.3 GIS Processing Logic

Once the spatial data are assembled and warped to a common geometric coordinate system, a host of GIS processing tools can be applied to the data. If the data are in raster form, all the image processing tools discussed in Chapters 8 and 9 can be applied, since the data can now be fashioned to look like multispectral image data. In addition, a number of logical operations can be performed to facilitate combining and extracting information from the various layers in a GIS. Two of the most useful of these operations that were not introduced in Chapter 8 are region growing and

Table 12.1 Layers in an Example GIS Database for Wildland Fire Risk Assessment

Source	Features	Code	
Landsat ETM+	Land use/Land cover (LULC)	Open Water	=11
		Low Intensity Residential	=21
		Commercial/Transportation	=23
		Bare Rock/Sand/Clay	=31
		Deciduous Forest	=41
		⋮	
	Fuel Type	Derived from LULC	
	Water Mask	Water (no risk)	= 0
		Other (potential risk)	= 1
DEM	Elevation	Meters	
	Slope	Percent	
Orthophotos	Roads	Roads	= 1
		Other	= 0
	Building Location	Structure	= 1
		Other	= 0

Boolean logic. Region growing is the simple process of locating all the grid cells (pixels) within some distance (number of pixels) of a feature. For example, from a rasterized road map (or a land cover map derived from multispectral imagery), we could find all pixels that were coded as roads. We could then add a new buffer layer to the GIS composed of all pixels within 100 meters of a road simply by having the computer locate all pixel centers within x pixels (x = 100 meters/grid cell size) of the road pixels. Similarly, we could make a map of distance ranges (e.g., 0 to 100 m, 100 to 200 m, etc.) from any feature of interest located in any of the GIS layers. The GIS layers can then be combined using combinations of simple algebra and Boolean logic (e.g., "and," "or," "not") to extract whatever data are required. Masks may also be created to exclude data from analysis by assigning values of zero and one to specific features in a layer and then using a multiplication operator to combine information from additional layers. For example, if we wanted to exclude all waterbodies from the analysis, we could first find all the pixels coded as water from a land cover map derived from Landsat imagery, and then assign values of zero to all water pixels and one to all other pixels. This water mask may be then multiplied with any other data layer to remove the water regions from analysis.

A full treatment of GIS processing logic is beyond the scope of this book, so we will restrict our discussion to one simple example to provide some insights into the possibilities of GIS processing. For simplicity, we will assume that all the data are preregistered and in raster format and that all image processing will use raster processing methods, as these will be similar to the image processing tools introduced in Chapters 8 and 9. Table 12.1 lists the initial layers that are available in the example GIS database and how they are coded. For this example, our

Table 12.2 Sample Criteria Required to Assess Wildland Fire Risk to Existing Buildings

Vegetation (Fuel Type)	Water	0 = No risk
	Grasses	5 = Low risk
	Light brush and small trees	10 = Medium risk
	Dense brush and large trees	20 = High risk
	Timber harvest residue	25 = Severe risk
Slope	0% - 10%	1 = Low risk
	11% - 20%	4 = Moderate risk
	21% - 40%	8 = High risk
	> 40%	10 = Severe risk
Roads	<100 m from a road	1 = Low risk
	>100 m from a road	10 = High risk
Buildings	<100 m buffer around buildings	1 = *Include*
	>100 m buffer around buildings	0 = *Exclude*
Total Weighted Fire Risk		0 = No risk
		1 = Low risk
		⋮
		45 = Severe risk

interest is in assessing the potential wildland fire risk within the study area based on a set of criteria listed in Table 12.2. Figure 12.16 shows the steps that would be used to generate some intermediate layers in the GIS and the logic that could be used to combine the layers to assign a fire risk to each pixel. Finally, a buffer algorithm is run to determine weighted fire risk for existing buildings. Note that to keep this example simple, other criteria such as building material, proximity to water sources, proximity to surrounding structures, and road and weather conditions have not been included.

One of the greatest advantages of a GIS system is that once the database has been populated, the data can be combined in many different ways in support of diverse interests. For example, commercial, transportation, environmental, agriculture, public health, and public utilities engineers can all access the same database in different ways to address application-specific questions. The information derived from a GIS may include images, maps, data tables, location indices, or simply decisions (e.g., no suitable site). In all cases, the display philosophy of making sure the data are in a format meaningful to the user and easily understood, and whose limitations are clearly delineated, should be a major design consideration

12.3.2 Databases and Models

In addition to spatial databases, analysis of remotely sensed data increasingly relies on other sources of data and modeling tools to increase our ability to extract information. Some of the data are simply tabulated databases. For example, there

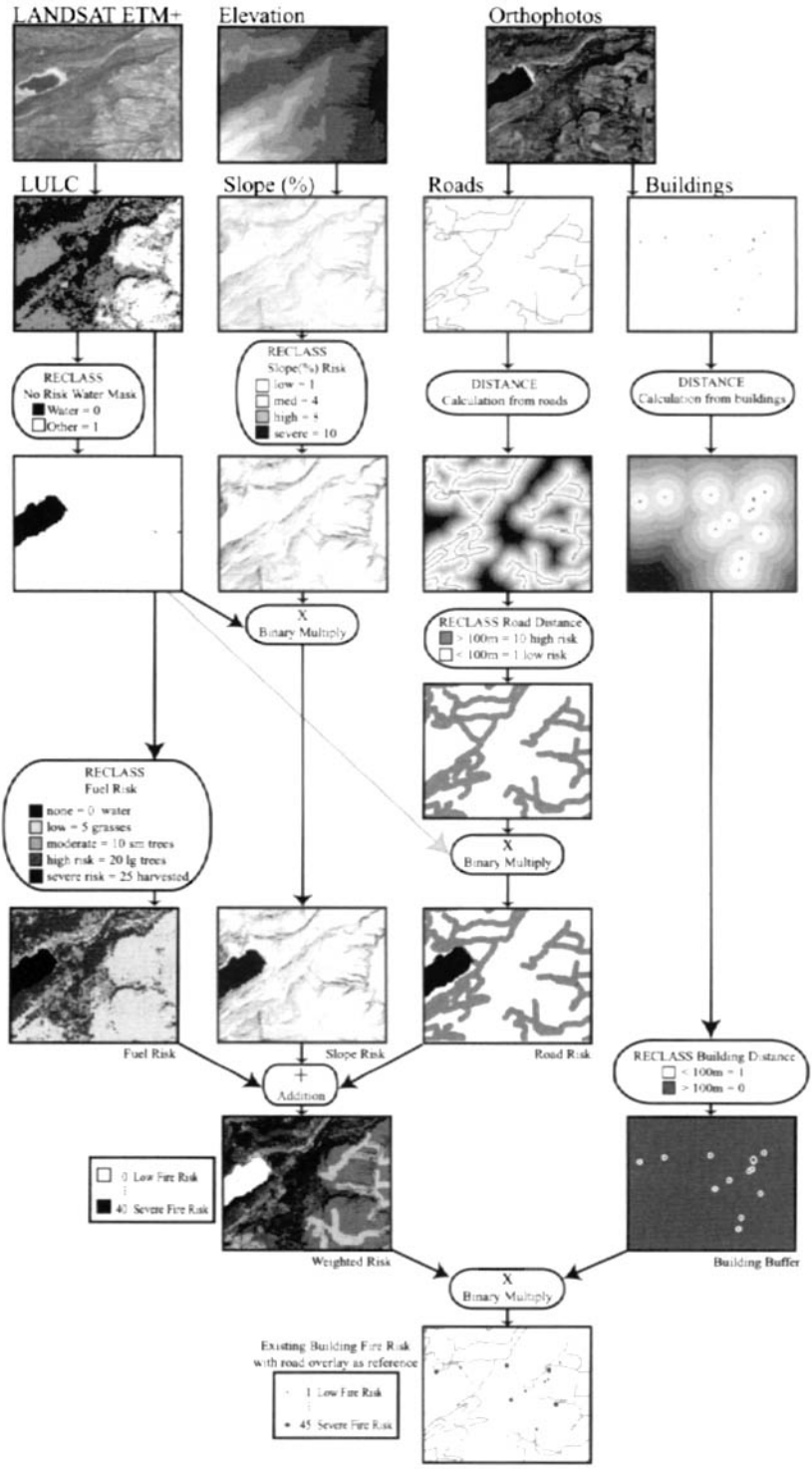


Figure 12.16 Logical steps used in assessing the wildland fire risk defined in Table 12.2 for GIS data set listed in Table 12.1.

are a variety of databases containing laboratory or field reflectance spectra for various material types. Some of these databases also include thermal and dielectric properties that can be useful in thermal infrared and radar image analysis. In many cases, the databases require analytical or stochastic models to make the support data useful to the image analyst. For example, the thermal properties of materials would typically only be of use if a thermal model were available to translate materials information along with meteorological variables into temperature estimates (cf. Chap. 14). Radiation propagation models such as LOWTRAN and MODTRAN discussed in Chapter 7 are another example where atmospheric databases and models are combined to provide the image analyst with tools to extract better information about a target's temperature or reflectivity. There are also a whole host of application-specific models that relate reflectance or temperature data to phenomena of interest. For example, Suits (1972) describes a model relating canopy reflectance to the structure, layering, and individual leaf reflectance in a vegetation canopy, and Gordon et al. (1983) describe a model relating water reflectance to the concentration of various parameters in the water.

While we will not delve into these specific models here, they reflect a growing trend toward image analysis that focuses on information extraction by coupling conventional image processing with physical models of the scene and the imaging process. This coupling more closely relates the final image parameters to the phenomena that caused them. This provides the user more ability to understand not only what was imaged but also, in many cases, why it appears as it does (i.e., is the change in appearance of the vegetation due to a species change, a change in canopy density, a change in slope and azimuth, or a change in vegetation vigor?). This model-based image analysis, coupled with the physics-based algorithms introduced in Chapter 11, essentially continues the image chain all the way to the application-specific phenomenology of interest. We will return to this topic again in Chapter 14, where we discuss a particular set of models aimed at synthetic image generation.

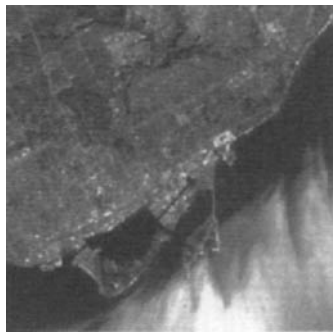
12.4 IMAGE FUSION

Loosely interpreted, *data fusion* is the process of merging data from different sources to improve our knowledge of a scene. Thus, most GIS- and model-based analyses could be construed to be data fusion. In this section, we want to concentrate on *image fusion* where two types of image data are somehow combined to form a new type of image data that ideally contain more interpretable information than could be readily accessed from the two image sets separately. Often, fusion involves merging data from different sensors with different resolutions to attempt to improve the interpretability of the lower resolution system.

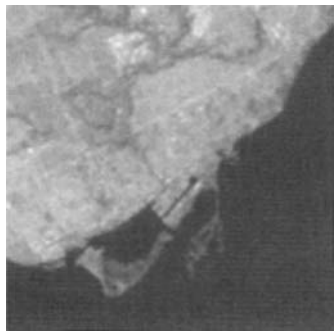
Haydn (1982) used this approach to improve the interpretability of LWIR images from the Heat Capacity Mapping Mission (HCMM) sensor. The LWIR images had a 500m GIFOV. The HCMM image was geometrically registered and resampled to the GIFOV of a Landsat MSS image of the same region (80m GIFOV). The data were "fused" by having one band of the MSS data represent intensity and

the thermal band from HCMM represent hue in an *intensity hue and saturation* (*IHS*) color space. This color space was chosen because the visual-cognitive system tends to treat intensity (brightness), hue (color), and saturation (purity of color or how pastel the shade is) as roughly orthogonal perceptual axes. While it may be very difficult to decide the relative change in redness, greenness, or blueness between two colors, it is much easier to say what the relative intensity, hue, and saturation are. Haydn took advantage of this by displaying the high-spatial-resolution data from MSS as intensity, since the visual system's spatial cues are most sensitive to intensity variations. The thermal data were displayed as hue, with saturation left at a fixed level. This is accomplished by a simple transformation of the data from IHS space to the red, green, and blue (RGB) space used in most display devices [cf. Conrac Corporation (1985)]. The resulting image "appears" to be a high-resolution thermal image, since the visual system's spatial cues are triggered by intensity variations, and thermal variations are interpreted by the variation in hue (color). There is actually no more thermal information, but by fusing it in this fashion, it is more readily interpreted. Figure 12.17 shows an example of this approach applied to a Landsat TM image, where the 30-m GIFOV data from the red spectral band is used to control intensity, and the 120 m LWIR data controls the hue.

Chavez et al. (1991) compared three methods for fusing a high-resolution panchromatic (pan) band image (e.g., SPOT) with lower resolution multispectral band images (e.g., SPOT Multispectral or TM). In all cases, the multispectral data are registered and resampled to the pixel size of the high-resolution pan band. The first method uses the IHS transform described above by taking three TM bands as RGB inputs and transforming to IHS color space. The resulting intensity image is found to be roughly proportional to the pan image brightness. The pan image is then histogram equalized to the intensity image (cf. Chap. 8) and used to replace the intensity in the IHS image. The "high-resolution" IHS image can then be back-transformed to RGB space for display. The resulting image appears to be a high-resolution three-band multispectral image. This method works reasonably well if the three input bands are highly correlated with each other and the pan band. When this is not the case, the substitution of the pan band for intensity can produce substantial shifts in the radiometry of the hybrid scene. This method is also limited in that only three bands can be processed at a time. A second method using the principal components transform has many similarities to the IHS transform method, but all the bands can be operated on simultaneously. In this case, the multispectral data are transformed into principal component (PC) images (cf. Sec. 9.3). The first PC image is often roughly proportional to the pan brightness. Again, a histogram equalization and substitution approach is used. The PC images with the high-resolution histogram-equalized pan band substituted for PC band 1 are then back-transformed to produce the hybrid high-resolution multispectral bands. This method yields results similar to the IHS transform. The images appear to have a higher resolution and are reasonably correct in terms of brightness when the first PC image and the pan band are highly correlated. However, when this is not the case, the radiometry of the hybrid image can be significantly different than in the original image. The third method was suggested by Schowengerdt (1980) and uses quite a different approach. In this case, the high-resolution image is high-



(a)



(b)

(c)

Figure 12.17 Image fusion of the red channel of TM (a), with the LWIR channel (b) using the IHS transform to produce (c). See color plate 12.17.

pass filtered, leaving an image with only the high-frequency edges. This image is summed with each band of the low-resolution multispectral image to yield a hybrid high-resolution image. This method is attractive in that the resultant image appears to have all the resolution of the pan image, and since low-frequency regions are left unchanged, the mean level radiometry is preserved. The limitation of this approach comes into play if the edge information in the pan band is not highly correlated with the edge information in one or more of the multispectral bands. In this case, while the image may "look" nice, false edges may be introduced or true edges neglected. A variety of similar approaches to multiresolution image fusion has been extensively reviewed by Zhang and Blum (1999). These approaches use

wavelet decomposition pyramids to extract different spatial frequencies from registered images and then recombine them for analysis.

Most of the methods presented thus far were primarily focused on enhancing the appearance of the hybrid high-resolution image to facilitate visual image exploitation. Another class of algorithms are primarily focused on fusion techniques designed to preserve the radiometric integrity of the high-resolution hybrid image for quantitative or machine exploitation. The visual improvement associated with these methods is comparable to the methods described above but comes as a by-product of the algorithm's attempt to generate a radiometrically correct estimate of what a high-resolution scene would look like. All of these methods must, of course, use some method of estimating how the high-resolution radiance is distributed in the hybrid scene. This cannot be known exactly, since even with the high-resolution pan image, there is too little information for a perfect reconstruction. As a result, this is an active area for ongoing research since many estimation methods exist and their utility may be both scene and application dependent.

One of the most straightforward methods of fusing multispectral data with higher resolution panchromatic data relies on the assumption that there is some degree of correlation between the multispectral band and the pan band brightness values. The method operates on superpixels that have GIFOVs the size of the original multispectral (MS) pixels. Each superpixel is composed of subpixels that are resampled to the same GIFOV as the pan band. The hybrid high-resolution pixel brightness can then be expressed as

$$DC_i' = \frac{DC_i}{DC_{pavg}} DC_p \tag{12.9}$$

where DC'_i is the digital count of the hybrid output subpixel value in the ith spectral band, DC_i is the digital count of the superpixel in the ith spectral band, DC_{pavg} is the mean digital count in the pan band for the subpixels contained in the low-resolution superpixel, and DC_p is the digital count in the pan band for the high-resolution pixel of interest. The algorithm implicitly assumes that DC_i/DC_{pavg} represents a fixed proportional relationship that is approximately constant over all subpixels in the superpixel.

This method is very attractive in that, on average over a hybrid superpixel, it must provide exactly the same radiance as in the original superpixel, i.e.,

$$DC'_{iavg} = \frac{\sum_{j=1}^{N} DC'_{ij}}{N} = \frac{\sum_{j=1}^{N} \left(\frac{DC_i}{DC_{pavg}} DC_{pj} \right)}{N} = \frac{DC_i}{DC_{pavg}} \frac{\sum_{j=1}^{N} DC_{pj}}{N} = DC_i \tag{12.10}$$

where DC'_{iavg} is the mean digital count over the hybrid superpixel in band i, and the sum is over the N subpixels in the superpixel. This means that, on average, at the resolution of the original MS imagery, the radiometry is preserved exactly. Pradines (1986), Price (1987), and Munechika et al. (1993) all used methods similar to this to enhance TM or SPOT MS data using the 10-m SPOT panchromatic band. This approach yields both radiometrically and visually improved images. It is, however,

subject to error when the assumptions are violated. This occurs when the MS band and the pan band are only weakly correlated or negatively correlated. It also occurs even with well-correlated bands if the correlation ratio (DC_i/DC_{pavg}) is not a constant over the superpixel, as might be the case for a mixed pixel (i.e., a superpixel composed of subpixels representing two or more material classes). Munechika et al. (1993) describe improvements to the baseline methods that incorporate adaptations for mixed pixels, and Price (1987) suggests corrections for correlation effects. Braun (1992) combined some of these approaches by developing algorithms that adapt to both poor correlation and mixed pixel effects. Both Munechika et al. (1993) and Braun (1992) used MS land cover classification accuracy to evaluate the quality of the hybrid images and showed that the hybrid images provided scene classification at least as good as, and often significantly better than, the original MS scenes. The "apparent" visual improvements in the MS imagery can be seen in Figure 12.18. Robinson et al. (2000) suggested an extension to this approach that is applicable to imaging spectrometer data. The algorithmic implementation of the Robinson et al. (2000) approach presented below was developed by Rhody (2002). The assumption is that a low-spatial-resolution image spectrometer data set of h spectral bands is resampled to the nominal pixel size of a corresponding high-spatial-resolution panchromatic or multispectral image having j bands. The high-spatial-resolution data are then registered and resampled to the spectroscopic data set to form a concatenated registered data set of $(h + j)$ bands and pixels with the same size as the high-resolution data. These data are then segmented into spectral classes using an unsupervised classifier (e.g., the k means classifier described in Chap. 9). The relationship for the kth spectral class between the brightness in the ith spectral band of the spectrometer data and the j spectral bands of the sharpening image can be solved by least-squares analysis of an expression of the form

$$DC_{ki} = a_{io} + a_{i1}DC_1 + a_{i2}DC_2 \cdots + a_{ij}DC_j + \varepsilon_{ki} \tag{12.11}$$

where DC_{ki} is the resampled low-resolution brightness in band i of the spectrometer for a pixel assigned to class k, DC_1 through DC_j are the corresponding brightness values for that pixel in the j sharpening bands, a_{io} through a_{ij} are the coefficients found by least-squares analysis of all the data in the kth class for the ith band of the spectrometer data, and ε_{ki} is the residual error. The hybrid or sharpened brightness is then obtained by applying the least-squared coefficients on a pixel-by-pixel basis according to

$$DC'_{ki} = a_{io} + a_{i1}DC_1 + a_{i2}DC_2 \cdots + a_{ij}DC_j \tag{12.12}$$

This process is repeated for all h bands in spectral class k and then for all of the spectral classes. Robinson et al. (2000) suggest ways to simplify or expand on this approach based on the level of correlation between the ith band and the sharpening bands.

Note that this sharpening of spectral data by fusion has all the limitations of the multispectral approaches in that it implicitly assumes brightness correlation exists between the bands to be sharpened and the high-resolution bands. As a result, the sharpened images are always more attractive to the visual system because they

Figure 12.18 Fusion of multispectral TM data (a) with geometrically registered and resampled SPOT data (b) to yield the fused product (c). See color plate 12.18.

have sharper edges even if the edges are falsely introduced. For example, Robinson et al. (2000) point out that fraction maps generated by unmixing the sharpened image cubes, while visually more pleasing, yield poorer quantitative results than an unmixing process proposed by Gross and Schott (1998) that first unmixes the low-resolution cube and then sharpens the fraction maps using higher spatial resolution data.

The subject of data fusion will continue to grow in importance as we generate higher spectral resolution sensors of modest resolution and attempt to apply the data to spatial phenomena at or below the GIFOV of the spectrometer images.

Since the limited radiometric flux density will make high-spatial-resolution spectrometers very difficult and expensive to build, fusion will be an interim solution. Regrettably, most fusion techniques make the image "appear" better, since the eye keys on the increased spatial structure whether or not it is radiometrically correct. Braun (1992) points out the need for quantitative, machine-based, image fidelity metrics to aid in evaluating the relative merits of image restoration algorithms (fusion can be considered a subset of the overall image restoration problem). He points out that most of the conventional metrics (e.g., classification accuracy and RMS error) are inadequate.

12.5 KNOW YOUR CUSTOMER

This chapter has emphasized ways to combine data and prepare them for dissemination to a later stage of analysis or to final users. These last stages of the image chain are often the most neglected. The remote sensing community must be careful to recognize that if the information passed to the next stage of analysis is not in a form and in terms (e.g., units) that are readily understood, then the information effectively does not exist. Furthermore, for users to obtain full value from the information, they must be able to assess its integrity. This means that the data must be analyzed in such a fashion that the confidence or error limits on the final data must be known and the limitations (e.g., robustness) of each approach clearly delineated. In order to prepare these final output products effectively, the remote sensing specialist must become familiar with the end user's requirements and final applications.

12.6 REFERENCES

Braun, G.J. (1992). "Quantitative evaluation of six multi-spectral, multi-resolution image merger routines." M.S. thesis, Center for Imaging Science, Rochester Institute of Technology, NY.

Chavez, P.S., Sides, S.C., & Anderson, J.A. (1991). Comparison of three different methods to merge multiresolution and multispectral data: Landsat TM and SPOT panchromatic. *Photogrammetric Engineering and Remote Sensing*, Vol. 57, No. 3, pp. 295-303.

Conrac Corporation. (1985). *Raster Graphics Handbook*. 2nd ed., Van Nostrand Reinhold, NY.

Cornsweet, T.N. (1970). *Visual Perception*. Academic, NY.

Fairchild, M.D. (1998). Color Appearance Models. Addison-Wesley, Reading, MA.

Gonzalez, R.C., & Woods, R.E. (2002). *Digital Image Processing*. 2nd ed., Prentice Hall, Upper Sadler River, NJ.

Goodchild, M.F., & Kemp, K.K., eds. (1990). "NCGIA core curriculum." National Center for Geographic Information and Analysis, University of California.

Gordon, H.R., Brown, D.K., Brown, J.W., Evans, O.B., Broenkow, R.H.B., & Broenkow, W.W. (1983). Phytoplankton pigment concentrations in the Middle Atlantic Bight: comparison of ship determinations and CZCS estimates. *Applied Optics*., Vol. 22, pp. 20-36.

Gross, H.N., & Schott, J.R. (1998). Application of spectral mixture analysis and image fusion techniques for image sharpening. *Remote Sensing of Environment*, Vol. 63, No. 2, pp. 85-94.

Haydn, R. (1982). "Multidisciplinary investigations on HCMM data over Middle Europe and Morocco...Southern Germany and Marrakech Morocco." NTIS No. 82N24588.

Hunt, R.W.G. (1987). *The Reproduction of Colour in Photography, Printing & Television*. 4th. ed., Fountain Press, Tolworth, UK.

Levine, M.S. (1985). *Vision in Man and Machine*. McGraw-Hill, NY.

MacAdam, D.L. (1942). Visual sensitivities to color differences in daylight. JOSA, Vol. 32, No. 5, pp. 247-274.

Munechika, C., Warnick, J.S., Salvaggio, C., & Schott, J.R. (1993). Resolution enhancement of multispectral image data to improve classification accuracy. *Photogrammetric Engineering & Remote Sensing*, Vol. 59, No. 1, pp. 67-72.

Park, S., & Schowengerdt, R.A. (1983). Image reconstruction by parametric cubic convolution. *Computer Vision, Graphics and Image Processing*, Vol. 20, No. 3.

Pradines, D. (1986). Improved SPOT image size and multispectral resolution. *Proceedings SPIE*, Vol. 660, pp. 98-102.

Price, J.C. (1987). Combining panchromatic and multispectral imagery from dual resolution satellite instruments. *Remote Sensing of Environment*, Vol. 21, pp. 119-128.

Rhody, H.E. (2002). "Enhancing spatial resolution for exploitation in hyperspectral imagery." IEEE Proceedings Applied Imagery Pattern Recognition Workshop, 31st, pp. 19-25.

Robertson, A.R. (1992). Color perception. *Physics Today*, pp. 24-29.

Robinson, G.D., Gross, H.N., & Schott, J.R. (2000). Evaluation of two applications of spectral mixing models to image fusion. *Remote Sensing of Environment*, Vol.71, pp. 272-281.

Schott, J.R. (1993). Contribution to *The Focal Encyclopedia of Photography*, 3rd. ed., Edited by L. Stroebel & R. Zakia, Focal Press, Boston, MA.

Schowengerdt, R.A. (1983). *Techniques for Image Processing and Classification in Remote Sensing*. Academic, Orlando, FL.

Schowengerdt, R.A. (1980). Reconstruction of multispatial, multispectral image data using spatial frequency content. *Photogrammetric Engineering and Remote Sensing*, Vol. 46, No. 10, pp. 1325-34.

Society for Information Display, Palisades Institute for Research Services, Inc., 201 Varick Street, Suite 1006, NY.

Suits, G.H. (1972). The calculation of the directional reflectance of a vegetative canopy. *Remote Sensing of Environment*, Vol. 2, No. 117.

Wolberg, G. (1990). *Digital Image Warping*. IEEE Computer Society Press, CA.

Wyszecki, G., & Stiles, W.S. (1982). *Color Science*. Wiley, NY.

Zhang, Z. & Blum, R.S. (1999). A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application. *Proceedings IEEE*, Vol. 87, Is. 8, pp. 1315-1326.

CHAPTER 13
WEAK LINKS IN THE CHAIN

In much of our discussion to this point, we have simplified many of the equations and concepts in order to introduce and deal with the fundamental or governing equations. In doing so, we recognize that there are times when the simplifications will be less valid than desired and that, in general, noise and/or error sources should become part of the equation. In analyzing an individual image or image product, we have to realize that it is the net result of the entire image chain: acquisition conditions, sensor system, processing procedures, and display or analysis. The end product will be limited by errors or degradations introduced all along the image chain and, like any chain will be no stronger than its weakest link. Understanding the subtleties of the image chain can provide insight into sources of limitations in the images, how much confidence to place where, and what processing techniques are viable, and it may improve the potential for information extraction from the image. The same image chain analysis can show the system analyst where the weak links exist so that improvements are focused where they can do the most good. Making a strong link even stronger has little impact on the quality of a chain, whereas even the slightest improvement in the weakest link impacts the entire chain. In this chapter, we will focus on spatial and radiometric fidelity issues that have been simplified to this point for clarity of presentation but that are often sources of weakness in the image chain.

We will continue to assume that most of the image chain can be approximately modeled as a linear shift invariant (LSI) system as described in Section 13.1.1. This is generally a good assumption for EO sensors and can typically be used, though more cautiously, for the image output links of the chain. However,

as discussed in Section 13.1.1, the LSI approximations are less applicable if there are visual links in the image chain and may have no meaning for nonimaging links (e.g., land cover tables, decision trees, etc.). The reader should recognize that the LSI approximations are widely used and quite robust, but that their validity should be evaluated for any system under study.

13.1 RESOLUTION EFFECTS (SPATIAL FIDELITY)

In this section, we want to look in more detail at the factors that impact the spatial resolution of the final image and to introduce a formalism for characterizing spatial image fidelity and cascading it through the system.

13.1.1 Spatial Image Fidelity Metrics

In earlier chapters, we encountered some simple methods for characterizing the *spatial resolution* of imaging systems (e.g., IFOV, GIFOV, and just-resolvable line pairs in a tribar target). The detector spot-size metrics such as IFOV and GIFOV provide intuitive estimates of resolution but imply a spatially detector-limited system. The discernible contrast of a tribar measure is more satisfying in that it incorporates end-to-end system effects but only provides performance information about the highest resolution, and then in subjective form. To overcome this limitation, the concept of the *square-wave* or *contrast transfer function* (*CTF*) is used as a measure of image fidelity. The CTF is an objective measure and characterizes image quality over all the spatial frequencies in the image. It is based on the concept of image contrast (C) or square-wave modulation defined as (cf. Fig. 13.1)

$$C = \frac{A - B}{A + B} \tag{13.1}$$

where A and B are the observed brightness for white and dark bars in a square-wave target. For example, given a printed image of a tribar target on the ground acquired with an imaging system, we would like the image to reproduce exactly the modulation on the ground. Figure 13.1 shows an example of the radiance measured from cross sections of a ground target and the corresponding radiance measured from an image (in arbitrary units). The contrast transfer (CT) is defined as the ratio of the contrast of the input and output images. Since the absolute magnitude of the image contrast will be a result of the tone reproduction process, the contrast transfer is often normalized to unity at zero spatial frequency, either by extrapolation or by using the lowest spatial frequency as a reference. As a result, we will often be interested in relative contrast transfer (i.e., relative to zero spatial frequency) under the implicit assumption that most imaging systems reproduce low frequencies very well (which they do), and we are interested in how well higher frequencies are reproduced relative to the low-frequency "ideal." This concept of contrast reproduction as a function of the spatial frequency of line pairs in a tribar target is captured in the CTF as illustrated in Figure 13.1. The normalized contrast is measured for

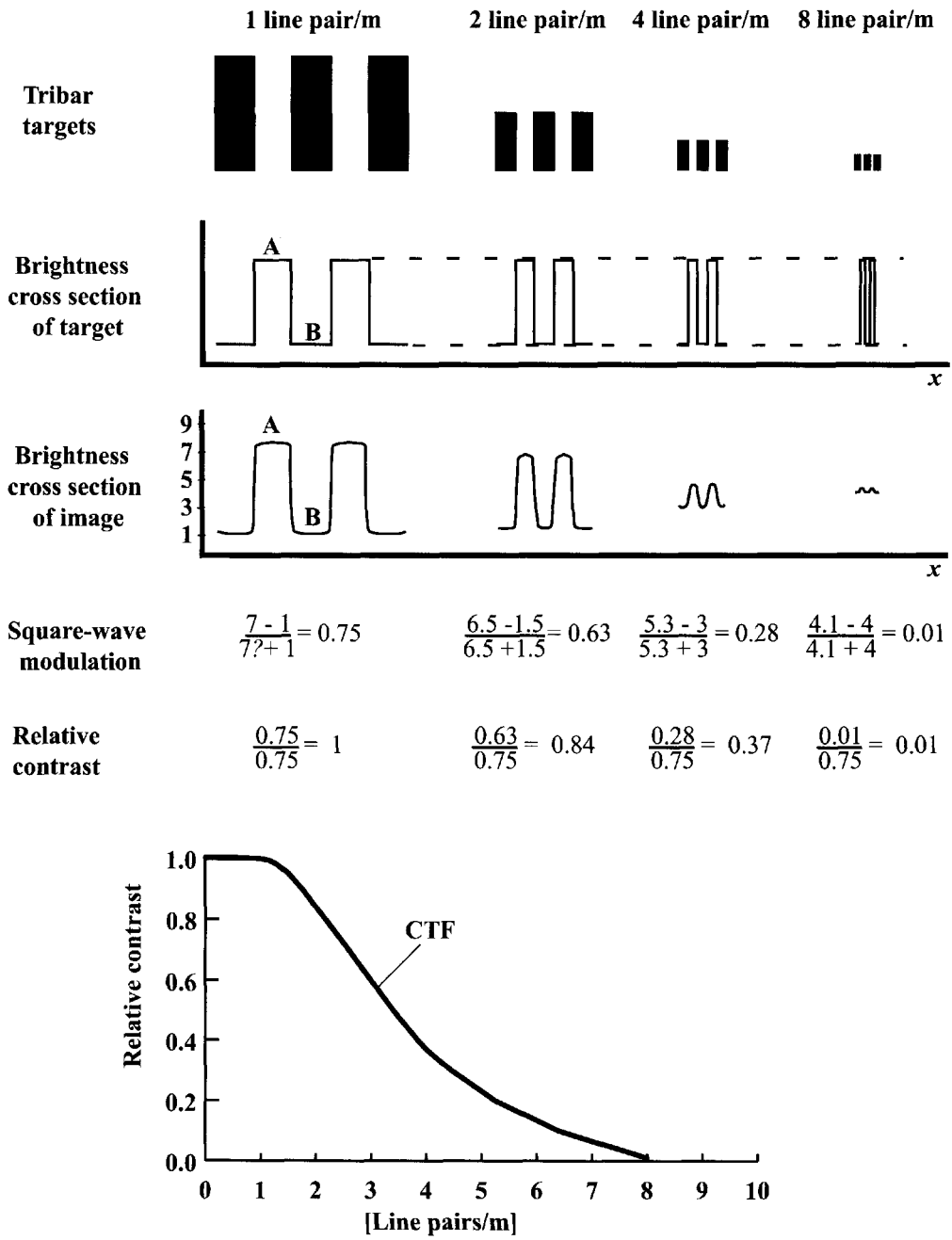


Figure 13.1 Square-wave modulation transfer function (MTF), contrast transfer function (CTF).

several line pair patterns in an image of a tribar target (note that these could be measurements from a print, transparency, digital image, or soft-copy display). The normalized contrast is then plotted as a function of the pattern frequency, and a curve fit through the data provides the CTF. The value of using the CTF rather than the frequency of the just-resolvable line pairs as a resolution metric is shown in Figure 13.2. If we assume that a relative contrast of 0.1 is where an observer could just

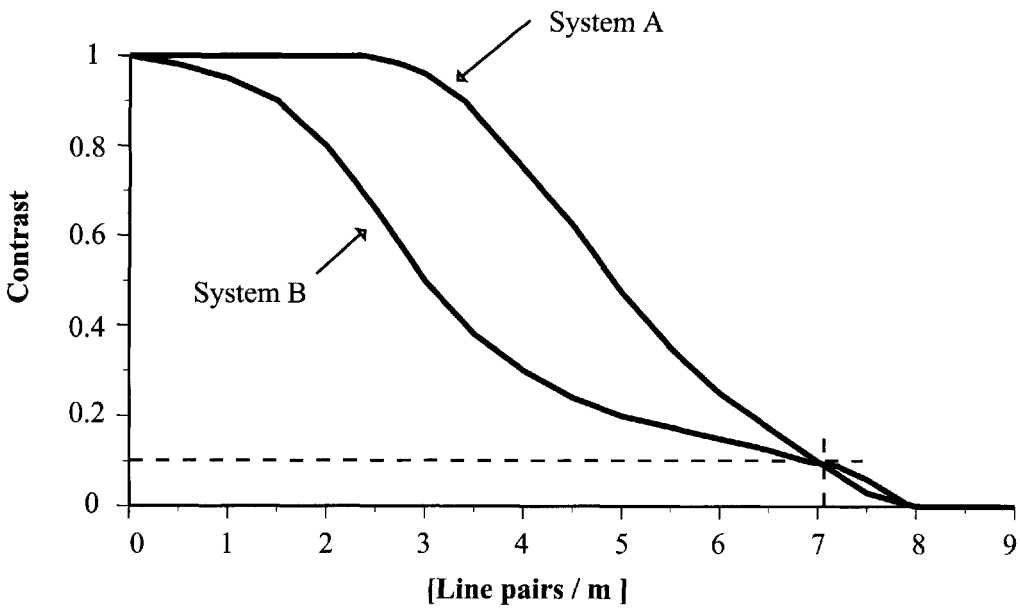
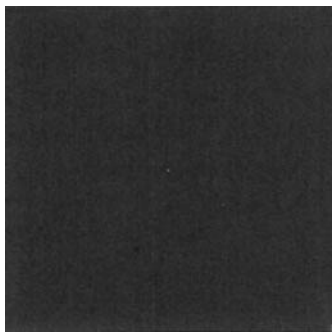


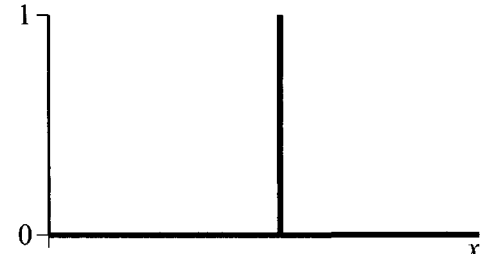
Figure 13.2 Comparison of the CTFs for two hypothetical systems.

distinguish the lines, then systems A and B illustrated in Figure 13.2 would have the same performance against the simple cutoff metric. However, using the CTF metric, it is clear that system A performs much better in the midrange frequencies and very similarly at higher frequencies. If all other considerations are equal, system A would be a much better system.

While having obvious intuitive appeal, the CTF metric is not commonly used because of the greater mathematical flexibility available through the use of the sinusoidal *modulation transfer function* (*MTF*). The MTF is defined in essentially the same fashion as the CTF, except that sinusoidally varying brightness functions are used as input (sine waves in cross section). The MTF tells how well a sinusoidally varying brightness of a given frequency will be reproduced by the imaging system. Again, the modulation (peak minus trough divided by peak plus trough) is normalized usually by the modulation at zero frequency. In cases where low-frequency response differs significantly from unity, the modulation recorded by the system can be carefully measured and compared with the modulation in radiance of the scene to compute the actual low-frequency MTF. The MTF carries essentially the same information as the CTF in terms of its intuitive appeal, while having a number of properties that make it easy to manipulate and solve for using linear systems theory. For example, the two-dimensional inverse Fourier transform of the MTF is the system impulse response or *point spread function* (*PSF*). Recall that the PSF is the response of the system to a point source of radiance (mathematically the response to a delta function). Conceptually, this is the shape of the blurred image of a point source (i.e., the blur spot). The PSF is also often used as a measure of resolution by measuring its *full width at half the maximum* (*FWHM*), as illustrated in Figure 13.3. When the systems PSF is projected onto the ground, its FWHM is often referred to as the ground spot or *ground sample size* (*GSS*).

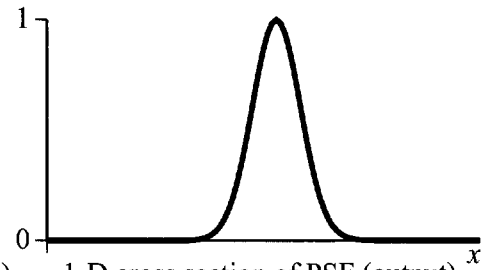


(a) 2-D delta function (input)

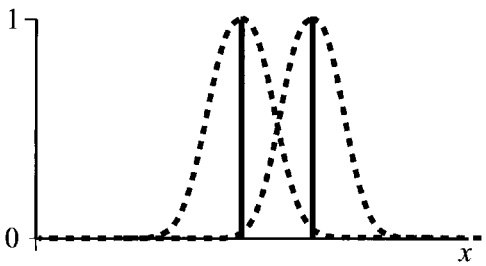


(b) 1-D delta function (input)

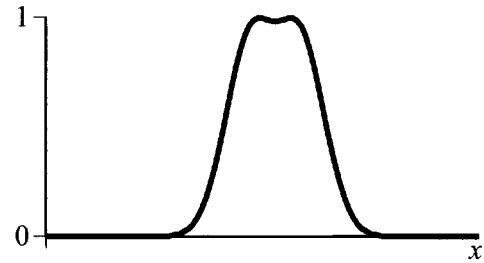
(c) 2-D PSF (output)



(d) 1-D cross section of PSF (output)



(e) 2 delta functions (input)



(f) Blurred representation of delta functions (output)

Figure 13.3 Point spread functions (PSFs) as a measure of resolution.

For simplicity in this treatment, we will often assume that the system can be treated as separable in x and y. Gaskill (1978) shows that in this restricted case the one-dimensional impulse response along the x axis corresponds to the one-dimensional inverse Fourier transform of the MTF along the u axis, i.e.,

$$h(x,0) = F^{-1}[H(u,0)] \tag{13.2}$$

where $h(x,0)$ is the profile of the PSF along the x axis, $H(u,0)$ is the profile of the MTF along the u axis (i.e., the profiles along corresponding axes form one-dimensional transform pairs), and the variables and nomenclature are those introduced in Chapter 8. In most cases, the separability assumption is a good approximation for many EO systems. However, a full two-dimensional treatment should be undertaken when detailed quantitative results are required. Gaskill (1978) points out that if the PSF is circularly symmetric, the Hankel transform can be used to compute the PSF from a radial slice through the system MTF.

Coltman (1954) showed that the MTF and CTF can be calculated from each other according to

$$\begin{aligned} \mathrm{CTF}(u) &= \frac{4}{\pi}\left[\mathrm{MTF}(u) - \frac{\mathrm{MTF}(3u)}{3} + \frac{\mathrm{MTF}(5u)}{5} - \frac{\mathrm{MTF}(7u)}{7} + \cdots\right] \\ \mathrm{MTF}(u) &= \frac{\pi}{4}\left[\mathrm{CTF}(u) + \frac{\mathrm{CTF}(3u)}{3} - \frac{\mathrm{CTF}(5u)}{5} + \frac{\mathrm{CTF}(7u)}{7} - \cdots\right] \end{aligned} \tag{13.3}$$

where u is spatial frequency [cycles/mm]. This yields a CTF approximately 10% higher than the MTF in the mid frequencies, with comparable values at the highest and lowest frequencies.

Rather than solving for the MTF by studying many sine wave inputs, it is often more convenient to take advantage of some concepts from linear systems [cf. Gaskill (1978) or Goodman (1968) for a more thorough treatment of these concepts]. Recall that a step function or edge can be thought of as being constructed by summing many sine waves of varying frequency and amplitude. A perfect edge contains all frequencies. If we introduce a knife edge (i.e., a step function) as the input to an optical system, then the image along the perpendicular to the edge is the *edge spread function* (*ESF*). The derivative of the edge spread function with respect to position is the *line spread function* (*LSF*) in that direction. Furthermore, one line through the 2-D MTF is the Fourier transform of the LSF, as illustrated in Figure 13.4. [Note that for separable systems LSF(x) = PSF(x)] The interrelationship of the image quality metrics plus the ease of combining the effects of components of the imaging system on the overall system MTF makes this the most common metric for describing the spatial resolution of imaging systems. The cascading principles of linear systems tell us that the MTFs of the components multiply to yield the system MTF (strictly speaking, this is only valid for incoherent systems where phase effects can be ignored). Recall from Section 8.4 that the convolution theorem states that multiplication in the frequency domain is equivalent to convolution in the spatial domain. Thus, if we know the impulse response (PSF) of each component of a system, they could be convolved together to produce the system impulse response function (PSF). We will utilize this cascading process extensively in the next section to study the impact of the image chain on the overall system MTF.

It is important to recognize that in many cases a human image analyst is the final consumer of the image. From one perspective, we are preparing the image for consumption by the analyst. However, I think it is often useful to consider the

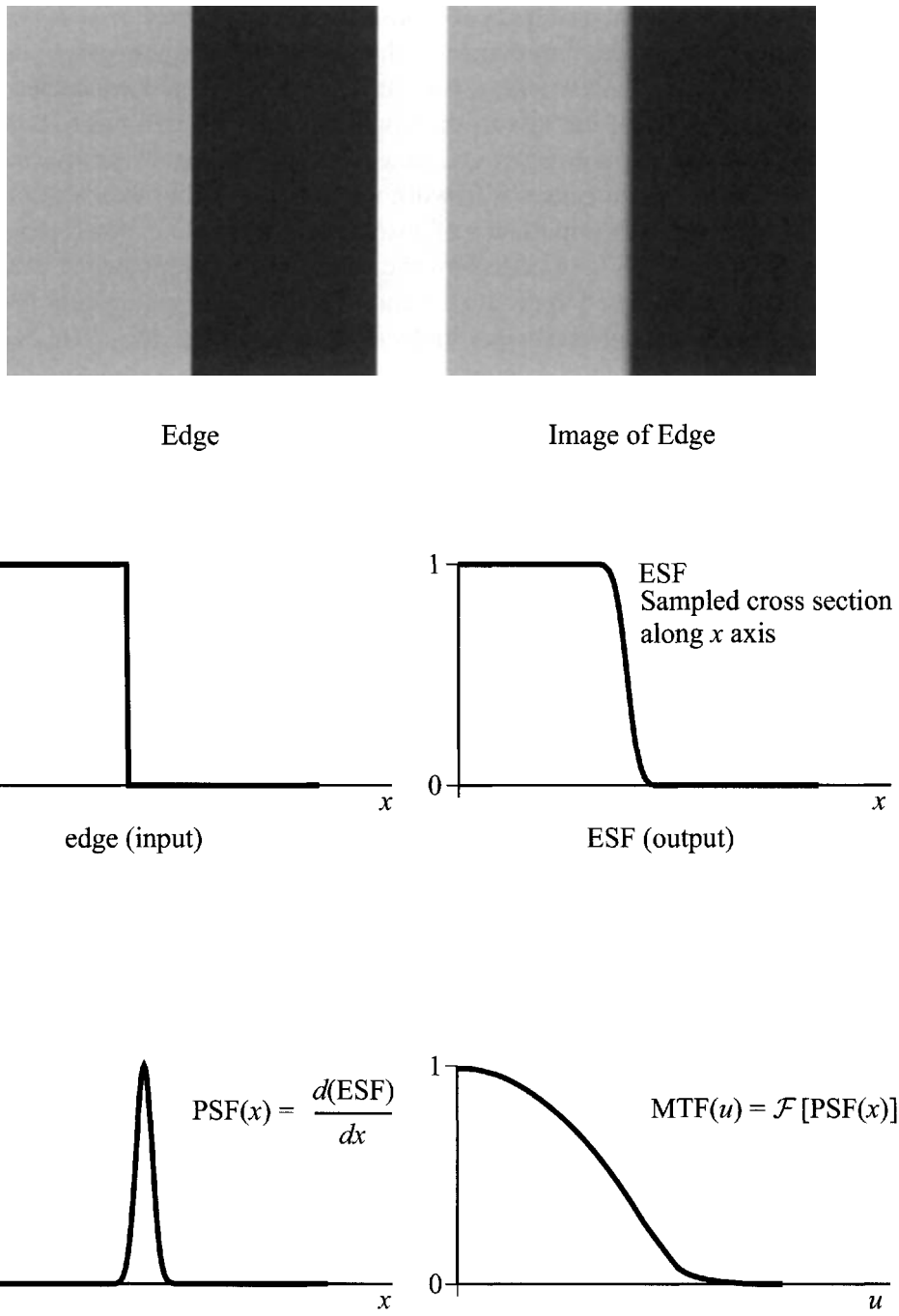


Figure 13.4 Relationship between the point spread function (PSF), edge spread function (ESF), and the modulation transfer function (MTF) of a system.

analyst as another link in the image chain. From either perspective, it is of the utmost importance to recognize the spatial frequency response of the visual system. Figure 13.5 shows an estimate of the contrast sensitivity of the visual system, with the maximum response occurring characteristically upward from the lowest frequencies. The contrast sensitivity can be thought of as being analogous to the

MTF. The visual spatial frequency response can be considered to be a weighting factor that determines the importance of the spatial frequency to visual analysis. For example, in the simplest case, it is clear that the midrange frequencies are the most important and that the loss of the lowest frequencies will have less impact than degradation of the midrange frequencies. This is due to the adaptive nature of the visual system, which causes it to work with relative rather than absolute gray levels. This reduces the importance of absolute gray levels for visual perception. Granger and Cupery (1972) discuss how the visual system MTF and the image system MTF can be combined to generate a combined image quality metric for visual assessment called the *subjective quality factor* (*SQF*).

13.1.2 System MTF

In this section, we want to assess how various links in the image chain impact the spatial image fidelity of the final image. In particular, we want to evaluate how the MTF concept can be used to characterize and assess the spatial characteristics of the image chain. From this standpoint, each link of the system is treated as a potential source of image degradation due to blurring. We generally assume that the world is composed of objects of interest that exhibit all spatial frequencies, and an ideal system would perfectly reproduce all those frequencies in the final image (i.e., have an MTF of unity at all observable frequencies). Each link can then be thought of as a filter that attenuates some frequencies of interest as characterized by the MTF of the link. The overall system performance is then the product of the individual MTFs. From the image chain standpoint, we are interested in the performance of the entire chain so that we understand the characteristics of the final image. We also are interested in determining the weak links so that we know where to concentrate our efforts at improving the system.

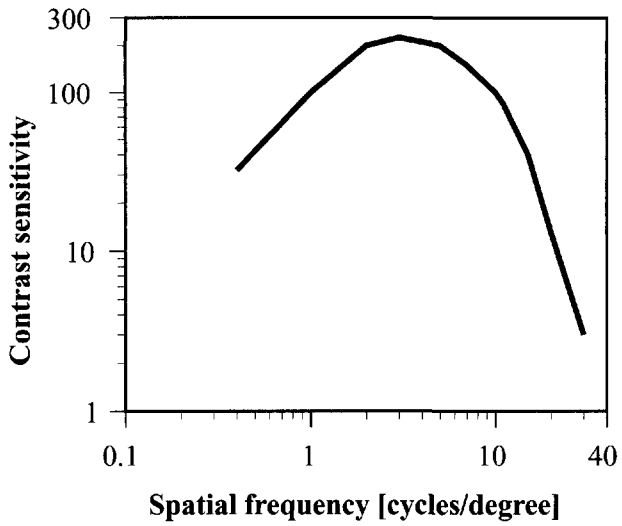


Figure 13.5 Contrast sensitivity of the human visual system. [Adapted from Hall and Hall (1977)].

13.1.2.1 Example System

The next several sections (13.1.2.2 to 13.1.2.10) examine the impact of various links in the image chain on the MTF of the entire chain. In each subsection, some example calculations will be included for a sample system. For this treatment, we want to take a hypothetical system of modest complexity and examine how the components might impact the final system performance. The components of an actual system would vary in complexity, but the approach would remain fundamentally the same. The image chain of interest in this case is illustrated in Figure 13.6. We are analyzing a single-channel airborne push-broom imager that utilizes an onboard digital recording system. The system specifications of interest are listed in Table 13.1. The image is geometrically rectified on the ground and printed as a transparent photographic image using a thermal dye diffusion film writer. From the point of view of the image analyst, we might tend to emphasize the image fidelity on the final transparency (i.e., at the end of the chain). However, the entire chain must be understood if we begin to look at tradeoffs to improve the image fidelity. For example, can we use an alternative film writer to improve the image fidelity or by using the same device write out a smaller portion of the image and achieve better fidelity over a smaller area? If we ship the imager overseas to be used with a different image processing system, what kind of performance can be expected? In this case, at a minimum we need to know the performance of the airborne imager as a unit, the imager with the digital ground processing taken together, and the film writer together with the prior links (i.e., the entire chain).

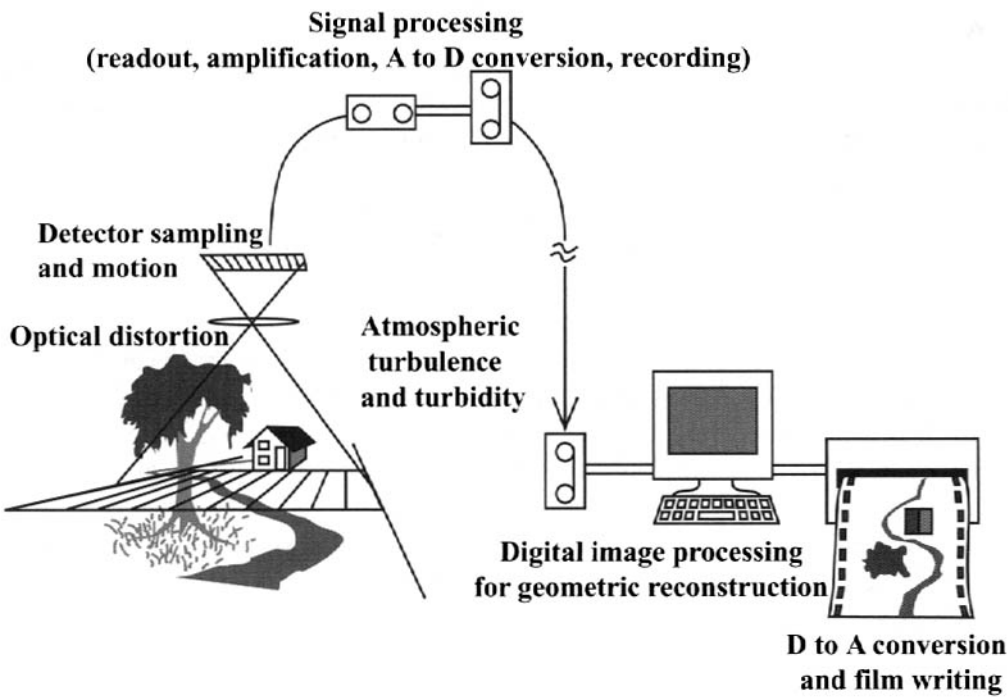


Figure 13.6 Components of a hypothetical imaging system for use in a spatial image quality assessment.

Table 13.1 Specification of Example System		
Description	Parameter	Value
F number	$F\#$	5.6
Focal length	f	0.56 m
Lens diameter (pupil)	d	0.1 m
Bandpass	$\Delta\lambda$	0.4 - 0.7 μm
Nominal wavelength	λ_{test}	0.55 μm
Side dimension of detector element	ℓ	15 μm
Nominal flying height	H	2000 m
Number of pixels in array	N	2048
Scale	s	$2.8 \cdot 10^{-4}$
Ground instantaneous field of view	GIFOV	0.05 m
Exposure time	t	$1 \cdot 10^{-3}$ sec
Field of view	FOV	±1.6°
Ground speed	W	90 mph
General: The system is an airborne push-broom scanner employing a silicon-based linear array detector and conventional refractive optics.		

13.1.2.2 Atmospheric Effects on Spatial Image Fidelity

The first link that must be considered is distortion to the wave front reaching the optical system due to atmospheric turbulence and scattering of flux from adjacent pixels into the pixel of interest (i.e., effects from type I photons in Fig. 3.1). The distortion due to atmospheric *turbulence* results from density fluctuations in the lower atmosphere causing variations in the index of refraction of the air that in turn induce random deviations in the wave front. For our purposes, we can normally think of the aggregate of these statistical fluctuations as a blur. Rees (1990) suggests that for vertical viewing, the angular dispersion will be of the order of $3 \cdot 10^{-6}$ radians. Or, assuming the effective height of the atmosphere (scale height) for turbulence to be 6666 m, the blur spot due solely to turbulence will be the product of the angular dispersion and the scale height. Thus, even when viewing down from space, the PSF due to atmospheric turbulence will be quite small (of the order of 0.02 m in diameter). This should not imply that turbulence can always be neglected. It can become very important for long paths in the lower atmosphere such as encountered by a *forward-looking infrared* (*FLIR*) system [cf. Beland (1993)] for a more thorough treatment).

The reduction in contrast due to the scattering of flux from adjacent pixels into the line of sight is a function of the *turbidity* of the atmosphere. Kaufman (1982) shows that this *adjacency effect* can be used to generate an MTF of the atmosphere using atmospheric radiative transport models. His results are in agreement with Pearce (1977), who used Monte Carlo simulations of the paths of the individual photons through radiative transport models and found that for low-to-moderate turbidity levels, the MTF is decreased but remains fairly constant over

a wide range of spatial frequencies representing target scales from a few to a few hundred meters (cf. Fig. 13.7). Thus, over a wide range of conditions, the dominant effect would appear as an overall reduction in scene contrast and would often be accounted for to the first order by mean-level corrections for atmospheric transmission and path radiance. However, some residual effects persist and will be more pronounced under hazy conditions and near-large high-contrast boundaries, i.e., large land–water interfaces. This effect does not fall off rapidly at higher frequencies and often goes unnoticed since it does not appreciably impact the appearance of the image. However, it can impact precise radiometric calculations (cf. Sec. 11.2.1) and should not be neglected under the conditions indicated above (i.e., large high-contrast boundaries).

The units of measure of spatial resolution or spatial frequency response will often be different for the different components of a system. To cascade the MTFs together, we need to convert all information to common units such as cycles per unit distance on the ground, which are used here. Pearce's data are already in this form. To convert the MTF due to atmospheric turbulence into similar units, we assume that the point spread function due to averaging of the randomly refracted beam can be represented as a Gaussian blur spot with an FWHM of 0.02 m. The MTF is estimated by computing the Fourier transform of the PSF. Figure 13.8 shows how this would roll off the atmospheric MTF but only at very high frequencies. The overall MTF of the atmosphere would have the general shape obtained

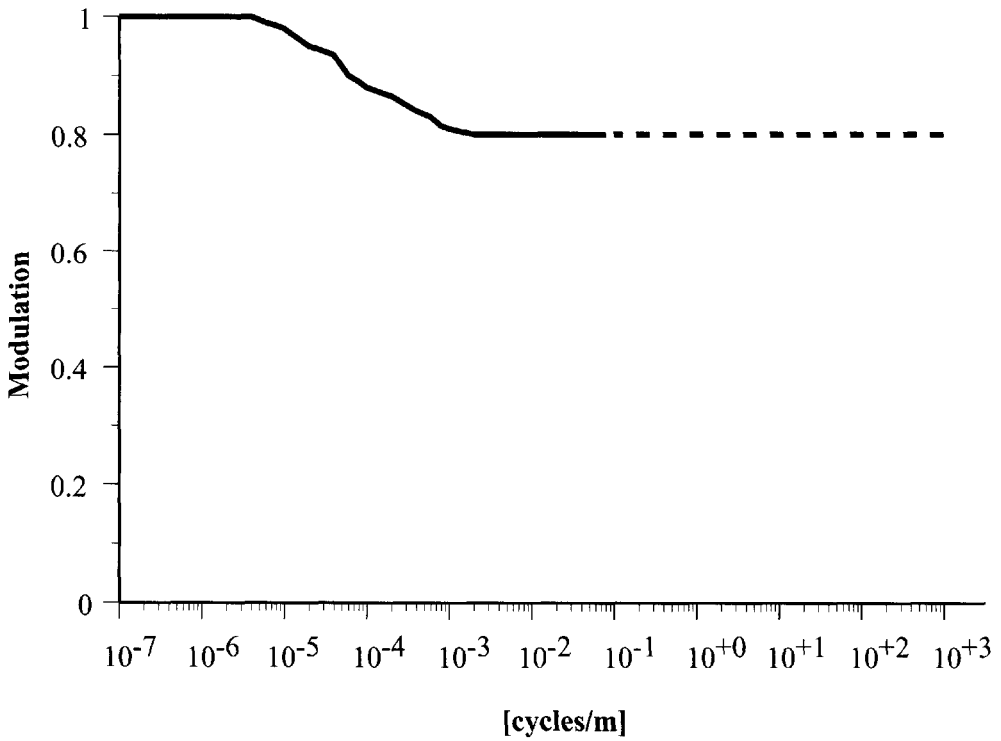


Figure 13.7 Modulation transfer function of the atmosphere due to turbidity effects [after Pearce (1977)].

by multiplying the MTF due to turbulence by the MTF due to turbidity. This is the general shape of this function, but it will depend on the viewing conditions and the composition and temperature of the atmosphere.

13.1.2.3 Optical Effects on Spatial Image Fidelity

Assessment of Figure 13.8 tells us that, from a spatial image fidelity standpoint, the image should reach an airborne or space-based system largely undistorted by the atmosphere except at extremely high frequencies (spot sizes of the order of 2 cm). The next link to be considered is the optical system, which for convenience we will treat as a unit including all windows, mirrors, and lens elements. Most on-axis remote sensing optical systems can achieve near diffraction-limited performance, making it relatively easy to calculate their expected performance to first order. For wider fields of view, aberrations (coma in particular) can become important, and a more detailed assessment of the lens design is required to determine expected performance. For a diffraction-limited incoherent system, the MTF is proportional to the correlation of the pupil function of the optical system with itself (i.e., the *autocorrelation* of the pupil function). For a simple optical system with an unobstructed circular pupil, the autocorrelation of the pupil function looks approximately like a circularly symmetrical triangular function. The high-frequency cutoff for incoherent irradiance is given by

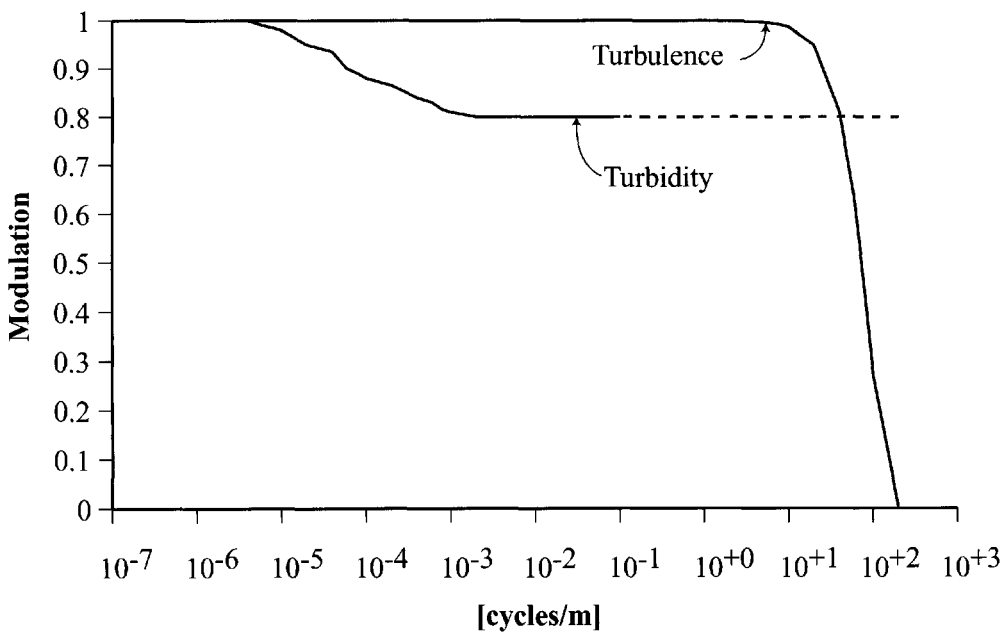


Figure 13.8 MTF of the atmosphere due to the combined effects of turbidity and turbulence.

$$u_{\max} = \frac{d}{\lambda f} = \frac{1}{\lambda F\#} \tag{13.4}$$

where $u_{\max}$ is the maximum spatial frequency [cycles/mm], λ is the wavelength, d is the aperture diameter, f is the focal length, and $F\#$ is the f number of the optical system. The PSF of the optics, obtained by taking the Fourier transform, is the well-known *Airy disc* pattern [cf. Fig. 13.9(a)]. The diameter of the Airy disc (to the first zero) is given by

$$D = 2.44\frac{f\lambda}{d} = 2.44\lambda F\# \tag{13.5}$$

where D is the diameter of the Airy disc. The central peak in the Airy disc (area enclosed by the first zero) contains 84% of the energy from a point source. The radius of the Airy disc is commonly used as a measure of the resolving power of optical elements according to the *Rayleigh criterion*. This is the minimum distance on the focal plane that must separate the images of two point sources (e.g., stars) so that they appear as visually distinct (i.e., the peak of one Airy disc will fall on the first zero of the second; [cf. Hecht (1987) and Fig. 13.3]. The MTF of the optical system is typically computed and expressed in terms of cycles per millimeter at the focal plane. This can be converted to cycles per unit distance on the ground simply by multiplying the spatial frequency by the scale factor, i.e.,

$$(u') = (u \cdot s) = \left(u \cdot \frac{f}{H}\right) \tag{13.6}$$

where u' is the spatial frequency expressed in cycles per unit length on the ground, u is spatial frequency expressed in cycles per unit length (typically mm) on the focal plane, s is the scale, f is the focal length, and H is the range from the sensor to the target (i.e., height above ground level for a vertical viewing system). The MTF due to diffraction limited optics for a clear-aperture F5.6 system flown at 2000 m is shown in Figure 13.10 ($d = 0.1$ m, $\lambda = 0.55$ μm).

As discussed in Chapter 5, many remote sensing systems use Cassegrain-style optics that have a central obscuration. The pupil function for a Cassegrain system has the general appearance shown in Figure 13.9. The MTF, obtained from the autocorrelation of the pupil function, and the PSF (Fourier transform of the MTF) for a typical Cassegrain system are also shown in Figure 13.9(b). Note that for comparison, the diameter of the primary and the effective focal length in the Cassegrain are the same as in the clear aperture example. The MTF of the Cassegrain has a characteristic hipped shape showing more filtering of the low- to mid-range frequencies and slightly better performance than the unobscured system at higher frequencies (note, however, that the cutoff frequency would be the same for both optics). As a result, the central peak in the PSF of the Cassegrain is slightly narrower, with more of the power shifted to the first lobe.

For the hypothetical clear aperture system, we see from Figure 13.10 that the optics have approximately the same spatial image fidelity as the atmosphere for this very high-resolution system (projection of the Airy disc onto the ground

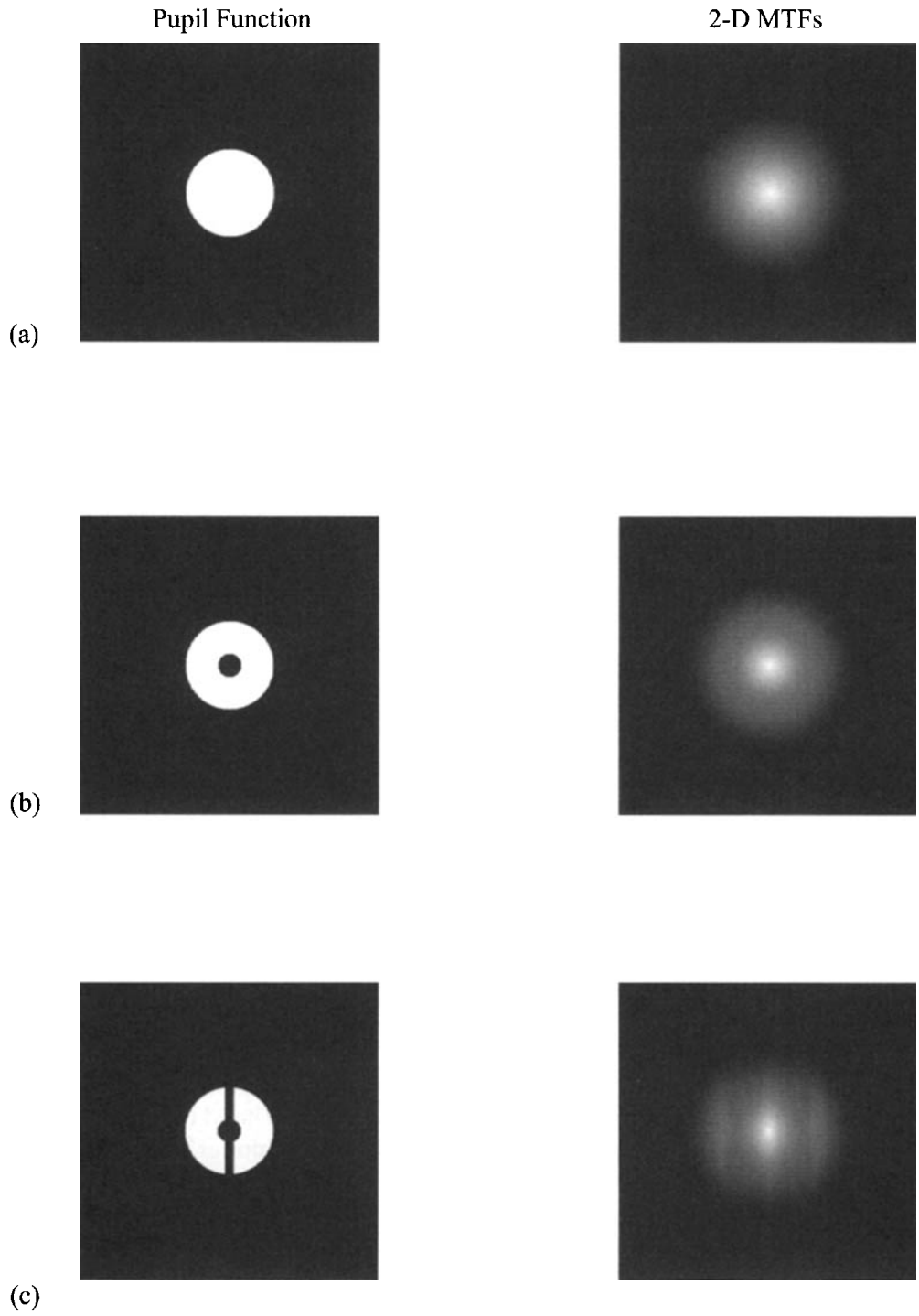


Figure 13.9 Pupil functions, 2-D MTFs, 1-D MTFs, and PSFs for some typical optical systems.

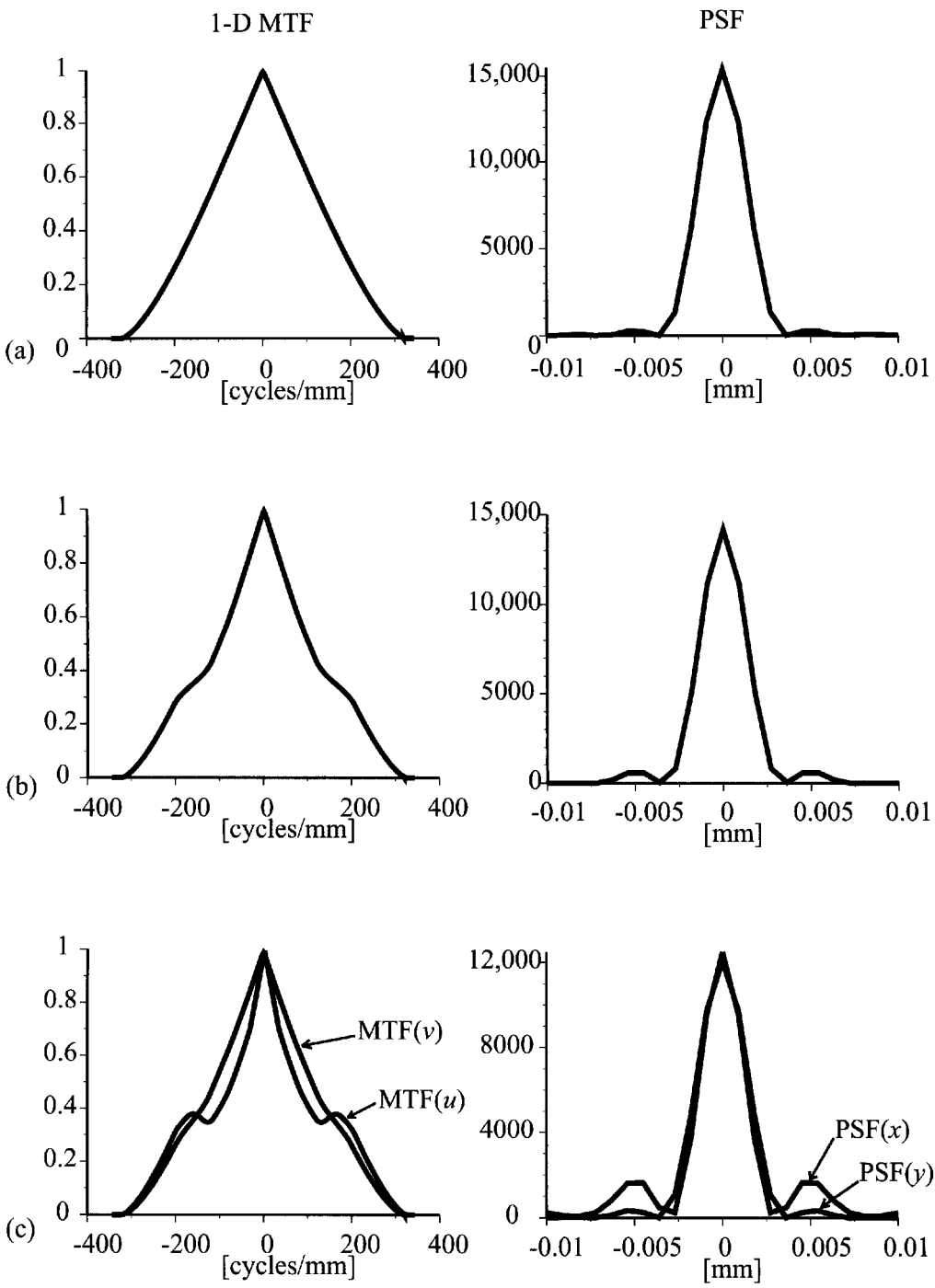
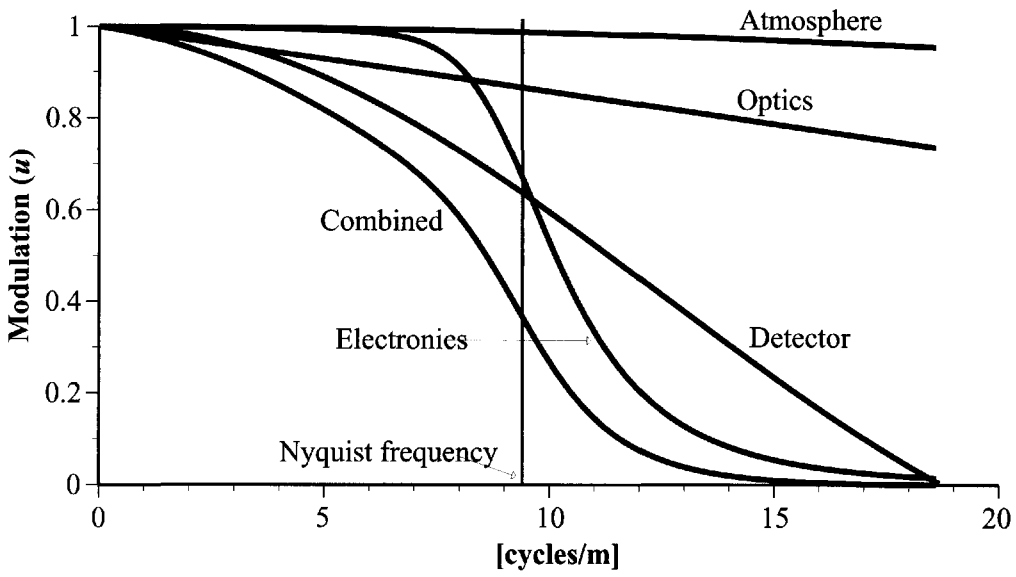
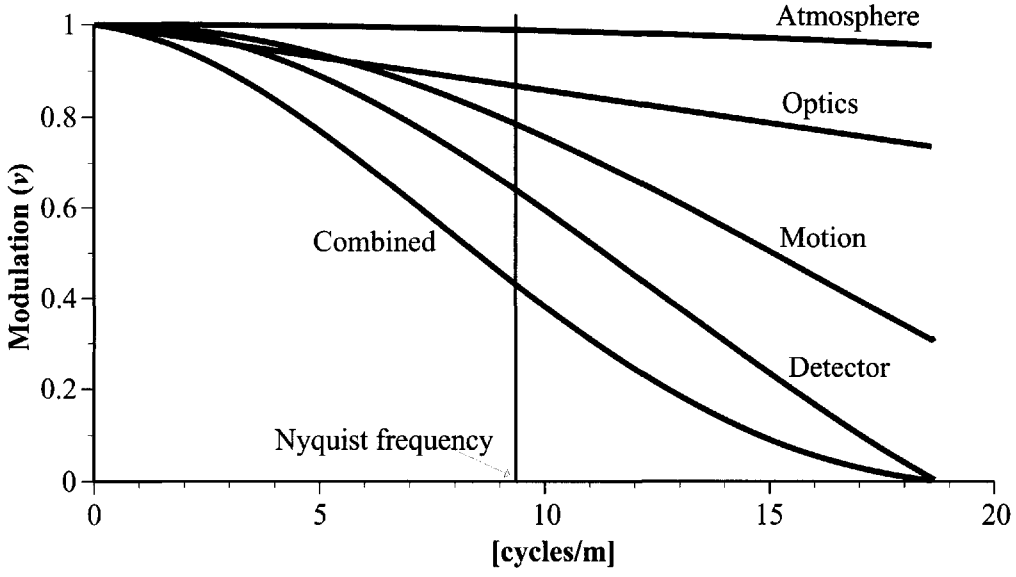


Figure 13.9 continued.



Across track



Along track

Figure 13.10 Cascaded MTFs of the atmosphere, optics, detector, detector motion, and electronics for a hypothetical system.

yields a spot size of 0.03 m). For any system of more modest resolution, the high-frequency atmospheric effect can usually be considered negligible.

13.1.2.4 Effect of Sampling by the Detector on Spatial Image Fidelity

In many system designs, the physical size of the detector is the limiting spatial factor in the optical radiation collection, because its dimensions are often larger than the blur spot (PSF) of the optics. To characterize the impact of the detector on spatial image quality, we need to briefly review sampling theory since the EO detector samples the scene. The sampling theorem states that the highest spatial frequency that can be reproduced by an ideal sampling system [i.e., where the sampling is performed by a series of unit area delta functions with spacing Δx referred to as $1/\Delta x$ comb $(x/\Delta x)$] is

$$u_N = \frac{1}{2\Delta x} \tag{13.7}$$

where u_N is the *Nyquist frequency* and Δx is the distance between sample centers. The MTF in an idealized system would be unity for frequencies less than the Nyquist frequency and zero for frequencies higher than Nyquist. Note that this assumes that the image does not contain frequencies higher than the Nyquist frequency. Higher frequencies must be filtered prior to sampling to avoid aliasing. Figure 13.11 shows a simplified one-dimensional representation of the sampling process. For linear array EO systems, Δx is the distance between detector centers. For scanning systems, it is the effective distance the detector has moved between samples (i.e., the distance the image has moved across the focal plane between sample acquisitions). In a push-broom system, for example, if we define the along-track direction to be y, then the sample interval Δy will be defined by the period of the sample acquisition. In the across-track direction x, the sampling interval Δx will be defined by the distance between detector centers in the array, Δx, which is referred to as the detector *pitch*. In practice, the detectors do not sample at a point, but occupy some finite dimension. In this case, the sampling process can be represented as a convolution of the detector impulse response with the image, which is then sampled by the *comb* function, i.e.,

$$g(x) = [f(x) * h(x)] \cdot \frac{1}{\Delta x} \text{comb}\left(\frac{x}{\Delta x}\right) \tag{13.8}$$

where $g(x)$ is the sampled image, $f(x)$ is the input image brightness, $h(x)$ is the impulse response of the detector (which is normalized to unit magnitude), and $1/\Delta x$ comb $(x/\Delta x)$ is the sampling function, which is unity for integer values of $x/\Delta x$ and 0 elsewhere. A rectangular detector can be treated as a rectangular averaging process in each dimension having a one-dimensional impulse response defined by

$$h(x) = \frac{1}{|\ell|} \text{rect}\left(\frac{x - x_0}{\ell}\right) \tag{13.9}$$

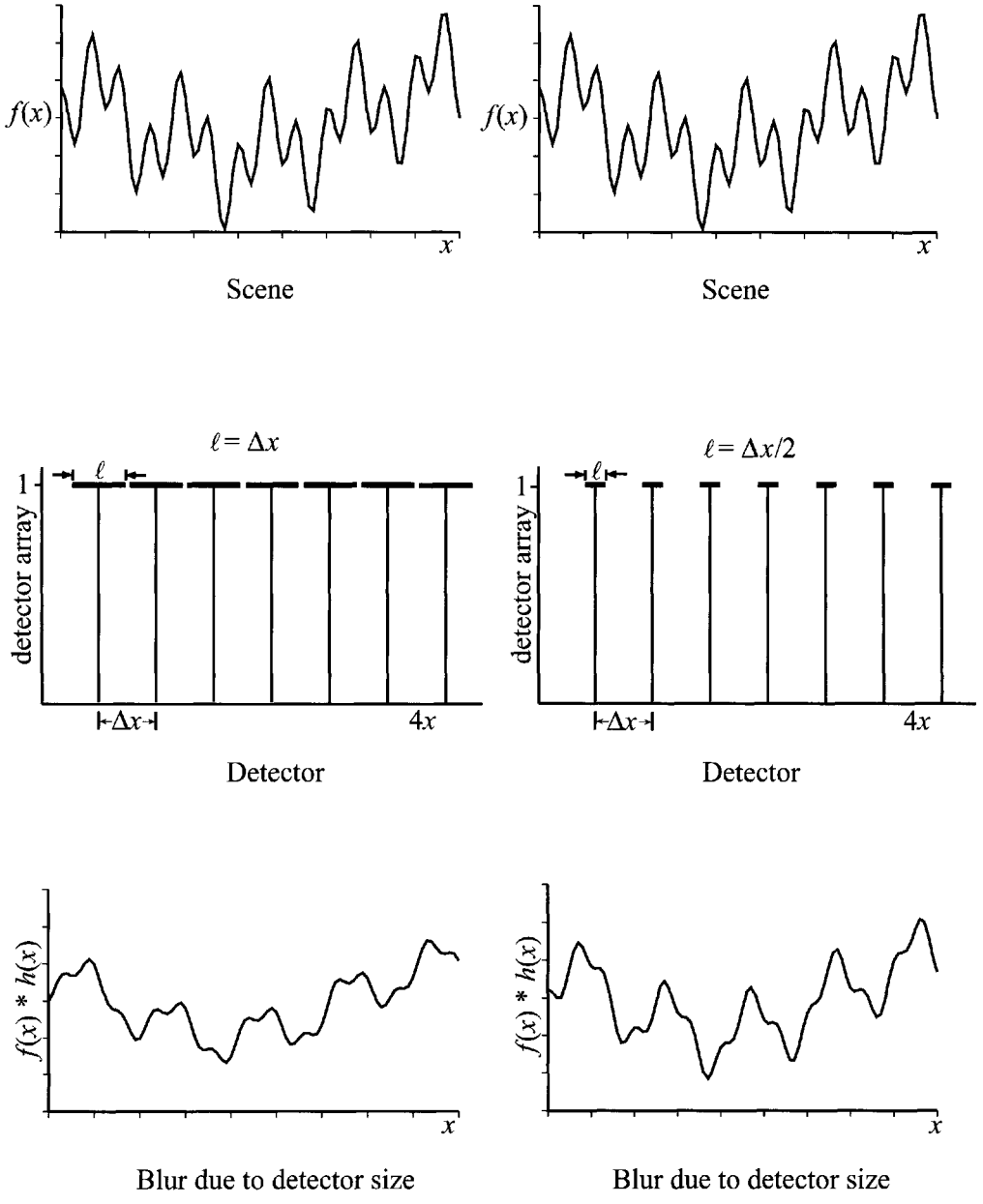


Figure 13.11 Detector sampling concepts and MTFs.

where rect[$(x-x_0)/\ell$] is a function equal to unity over the range x_0-$\ell/2$ to x_0+$\ell/2$, equal to 1/2 at $x_0 \pm \ell/2$, and zero elsewhere. The MTF due to the detector blurring is then given by the Fourier transform of the impulse response (PSF) of the detector; in one dimension,

$$MTF(u) = \mathcal{F}\left[\frac{1}{|\ell|}\text{rect}\left(\frac{x}{\ell}\right)\right] = \frac{\sin(\pi \ell u)}{\pi \ell u} = \text{sinc}\, \ell u \tag{13.10}$$

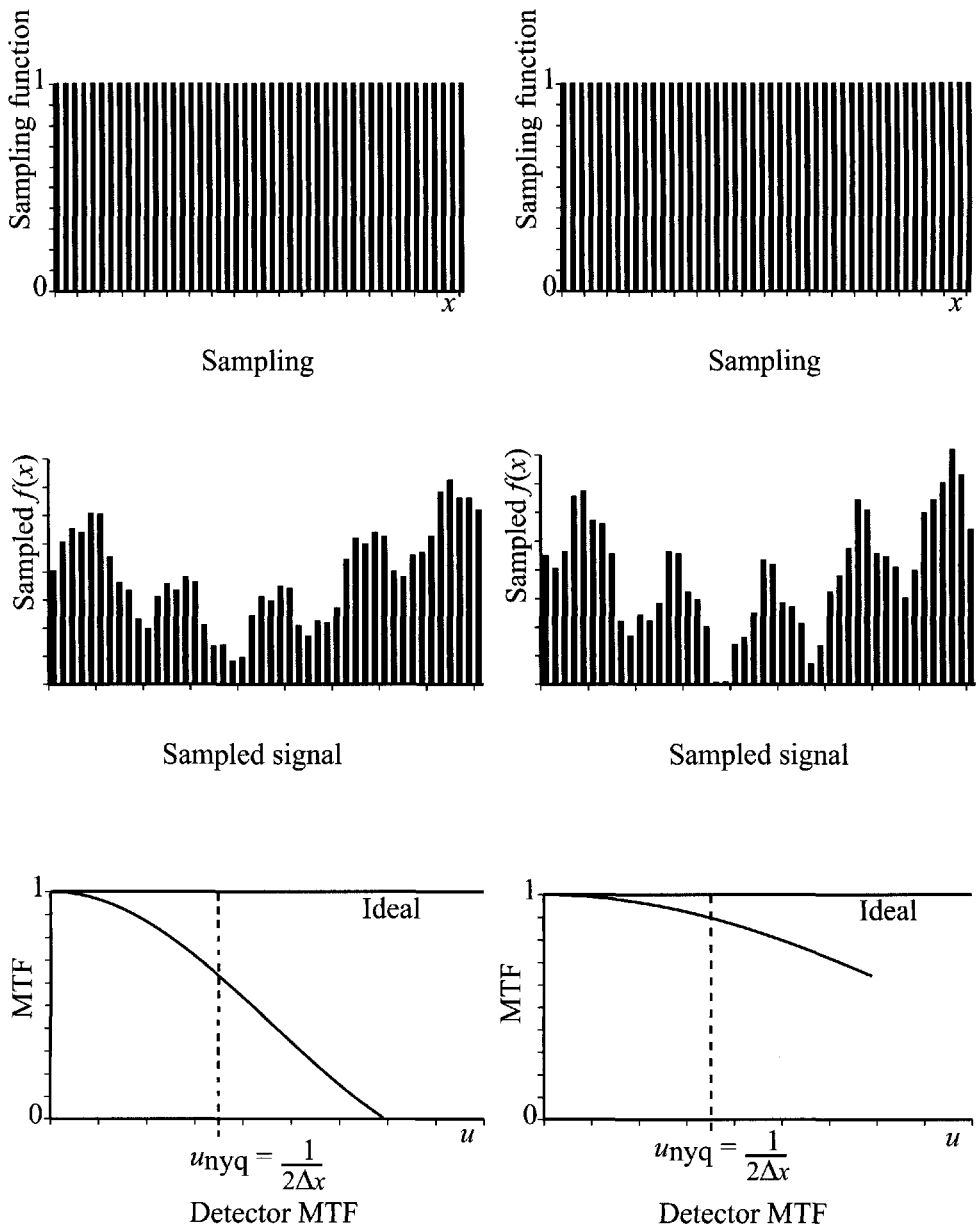


Figure 13.11 continued.

In many cases, ℓ is approximately equal to the detector spacing (Δx) in the across-track and the along-track directions. This results from using adjacent square detectors. The shape of the MTF for this case is shown in Figure 13.11. In some cases, particularly in the across-track direction for push-broom imagers or in either dimension for staring arrays, the effective detector size does not completely fill the space between detector centers. In this case, ℓ is less than Δx (more closely approximating the ideal delta function), and the MTF is improved as shown in Figure 13.11 for a the case where $\ell = 0.5\Delta x$. The downside of this is that fewer photons are detected by low-fill factor detectors (i.e., detectors whose sensitive area is less than their physical size), thus reducing the signal to noise.

It is important to recognize that to this point we have been able to assume that the PSFs of the links and, therefore, the MTFs are symmetric in x and y. In many cases, due to detector shape or differential sampling schemes in the x and y directions, the PSFs and MTFs are asymmetric. While ideally we would always treat the full two-dimensional functions in all our analysis, the mathematics and graphic representation of the results can become cumbersome. For this reason, it is common practice to just present the x and y or u and v components of the PSFs and MTFs for simplicity. The user must evaluate the characteristics of the particular system under study to determine when more detailed assessment is warranted (e.g., when the system response is not separable in x and y).

For our example case, we are going to assume that we have fully filled square detectors on 15-μm centers and that they are sampled in the along-track direction at time intervals that correspond to 15 μm of image motion, so that we should have symmetry in x and y. Using a linear array of 15 μm square detectors in our hypothetical push-broom system, we would expect to be resolution-limited by the detector. The detector dimension ($\ell = 15$ μm) is approximately twice the Airy disc diameter ($D = 2.44\lambda F\# = 2.44 \cdot 0.55$ μm $\cdot$ 5.6 = 7.5 μm). The impact on the MTF scaled to ground units is shown in Figure 13.12. We see that for this system the detector limits the high-frequency response. However, the MTFs of the optics and detector are both significant contributors to the reduced fidelity at higher frequency.

Fiete (1999) points out that if we Nyquist sample at the diffraction-limited cutoff frequency (u_c) to attempt to sample the highest possible spatial frequencies passed by the optics, then

$$u_c = \frac{1}{2\Delta x} = \frac{1}{F\#\lambda} \tag{13.11}$$

or the pitch would be given by

$$\Delta x = \frac{F\#\lambda}{2} \tag{13.12}$$

This is referred to as Q equal to 2 sampling, where

$$Q = \frac{F\#\lambda}{\Delta x} \tag{13.13}$$

However, from an image quality/image interpretability perspective, Fiete (1999) points out that a good compromise between spatial fidelity and signal to noise (which is improved with larger Δx values) is often achieved with Q values near 1. He points out that for Q equal to 1 sampling, there is some power from the optics at the Nyquist sample limit (i.e., the MTF is not zero), and that the aliased frequencies (frequencies between Nyquist and u_c that would be sampled in a $Q = 2$ system) have almost no power, and boosting them to where they contribute to the image's appearance will typically adversely boost the noise.

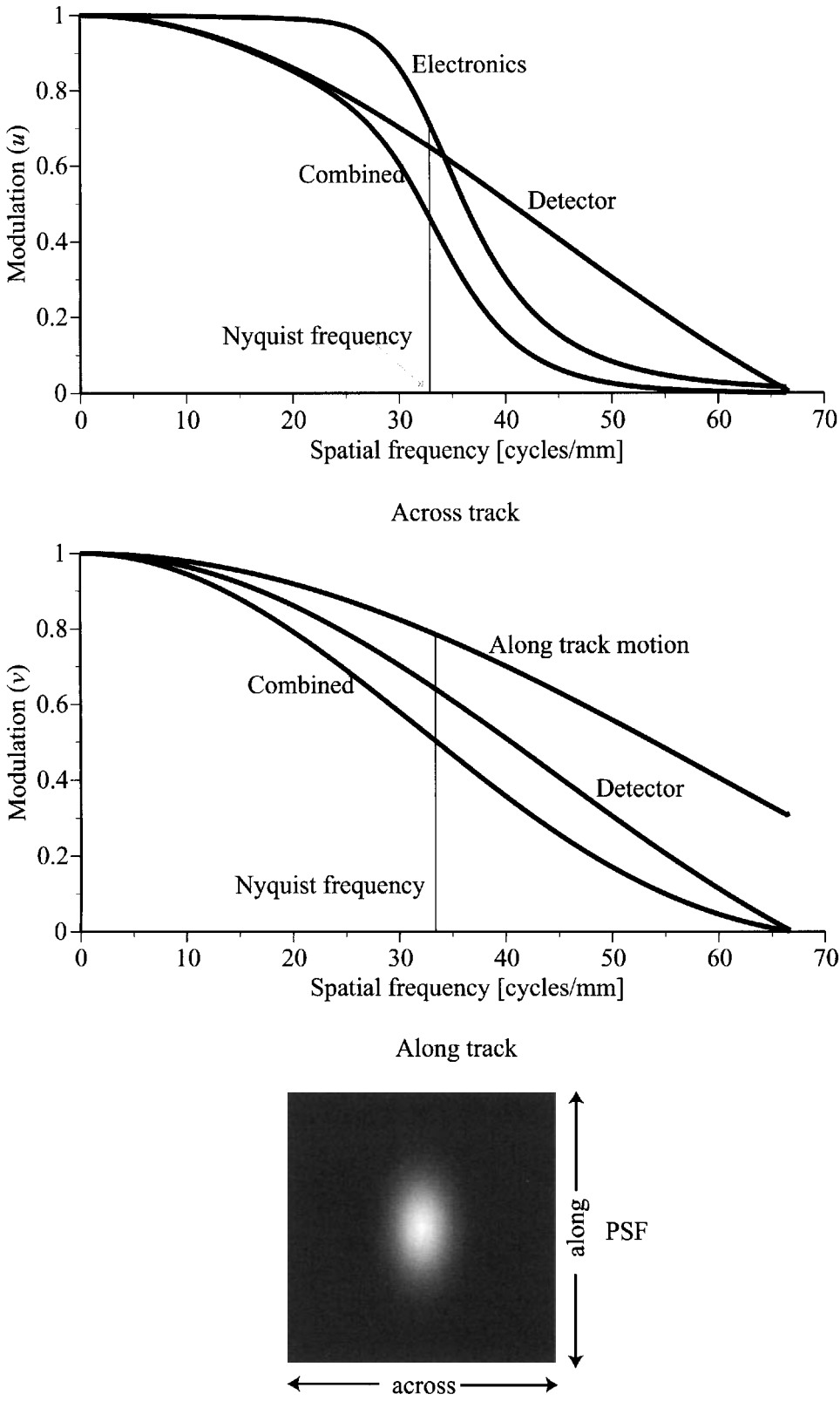


Figure 13.12 Effects of motion and detector readout on image MTF and 2-D PSF.

13.1.2.5 Motion Effects on Spatial Image Fidelity

Another factor that must be considered in the detector sampling process is the time interval over which the sensor is integrating the flux. Because the image is moving across the detector during this process, there is additional blurring in the motion direction (i.e., along-track for a push broom or across track for a scanner). This process can be represented by an impulse response function in the form of a rect. The rect would have a base as long as the distance the image moves during the integration time. For example, in the system we are discussing, if the aircraft ground speed (W) is 90 mph (40 m/sec) and the integration time (t) is $1.0 \cdot 10^{-3}$ sec, then the PSF in the y direction due to motion is

$$PSF(Y) = \frac{1}{|Wt|}\text{rect}\left(\frac{Y}{Wt}\right) \tag{13.14}$$

and the PSF is expressed in ground coordinates. The MTF in terms of spatial frequencies at the image plane (v) is

$$MTF(v) = \mathcal{F}\left[\frac{1}{sWt}\text{rect}\left(\frac{y}{sWt}\right)\right] = \text{sinc}(vsWt) \tag{13.15}$$

where s is the scale factor, sWt = 11 μm, and Wt = 0.04 m. This has been scaled to ground coordinates and multiplied by the detector MTF to yield a combined detector- and motion-induced MTF relative to the Y direction for plotting in Figure 13.12. Note that in this case the blur due to motion is quite substantial: nearly as large as the blur due to the size of the detector.

13.1.2.6 Signal Processing Effects on Spatial Image Fidelity

In the across-track direction, a different set of phenomena introduce incremental signal degradation. These include blurring of the signal due to less-than-perfect charge transfer out of the linear array and less-than-perfect amplification and digitizing of the frequencies in the electronic processing. These effects can be combined and represented by the MTF of the electronic readout and processing (MTF_e). Since this is an electronic signal, it is normally expressed initially in terms of temporal frequencies. The MTF_e can be converted to spatial frequencies on the focal plane based on the detector readout rate and the image scale if required, i.e.,

$$u = u_e\left[\frac{\text{cycles}}{\text{sec}}\right] \cdot t_c\left[\frac{\text{sec}}{\text{pixel}}\right] \cdot \frac{1}{\Delta x}\left[\frac{\text{pixels}}{\text{mm}}\right] = \frac{u_e t_c}{\Delta x}\left[\frac{\text{cycles}}{\text{mm}}\right] \tag{13.16}$$

where u_e [cycles/sec] is the temporal frequency of the electronics associated with detector readout and electronic processing including the A-to-D converter, t_c [sec/pixel] is the clocking time of the readout electronics, and Δx [mm/pixel] is the spacing between pixel centers on the array. In the hypothetical system we are considering, each line of data is transferred simultaneously to a data storage buffer employing *first in, first out* (*FIFO*) technology and then read out at a slower rate to

minimize degradation of the MTF. The electronic amplifier is designed to cut off at the Nyquist limit to reduce aliasing. The final MTF due to the readout and processing electronics is shown in Figure 13.12 in terms of the spatial frequency at the focal plane, and in Figure 13.10 in terms of spatial frequency on the ground. Note that the spatial frequency effects due to the electronics adversely impact the MTF, but in this case they are not as significant as the motion effects in the y dimension. This is shown in Figure 13.12 where the two-dimensional point-spread function due to the size of the detector, motion in y, and electronic readout, electronic processing, and sampling in x are shown. This asymmetric PSF is typical of most EO systems.

13.1.2.7 Sensor System MTF

The final MTF of the imaging system, including the atmosphere, is shown in Figure 13.10. This is obtained by cascading (multiplying) the other curves in Figure 13.10 to obtain the effective sensor MTF with respect to Y [MTF(v')] and with respect to X [MTF(u')]. An analysis of these curves shows that the spatial frequency response is dominated by the size of the detector, but in this case the motion, electronics, and optical effects are all significant. This is an example of a system where the GIFOV is not a particularly good indicator of resolution. If we wanted a simpler way of characterizing the resolution than the PSF or MTF, we could use the FWHM of the PSF scaled to the ground (the GSS) as shown in Figure 13.13. Alternatively, Slater (1980) suggests use of the *effective instantaneous field of view* (*EIFOV*), which is derived from the frequency where the MTF is 50% of maximum ($u_{0.5}$):

$$\text{EIFOV} = \frac{1}{2u_{0.5}s} \tag{13.17}$$

where s is the scale factor. Note that both the GSS and the EIFOV can and should be defined in both the along-track and across-track directions because of the asymmetry of the EO imaging process.

These full-system resolution metrics are much better estimates of system performance than simple measures such as GIFOV. However, for many systems the detector remains by far the dominant factor in determining resolution, such that the GIFOV is a good estimate of performance, and for this reason remains widely used because of its simplicity. It should be recognized that while all the resolution metrics introduced are attempting to characterize common performance characteristics and will result in similar values, they should only be directly compared to performance metrics derived in a similar fashion (i.e., EIFOV to EIFOV, not GIFOV to GSS).

13.1.2.8 Effect of Image Processing on Spatial Image Fidelity

For assessment and comparison of sensor performance, analysis of the spatial frequency performance of most systems stops here. However, from an end user's standpoint, the image the sensor "sees" typically is not the final product, and other

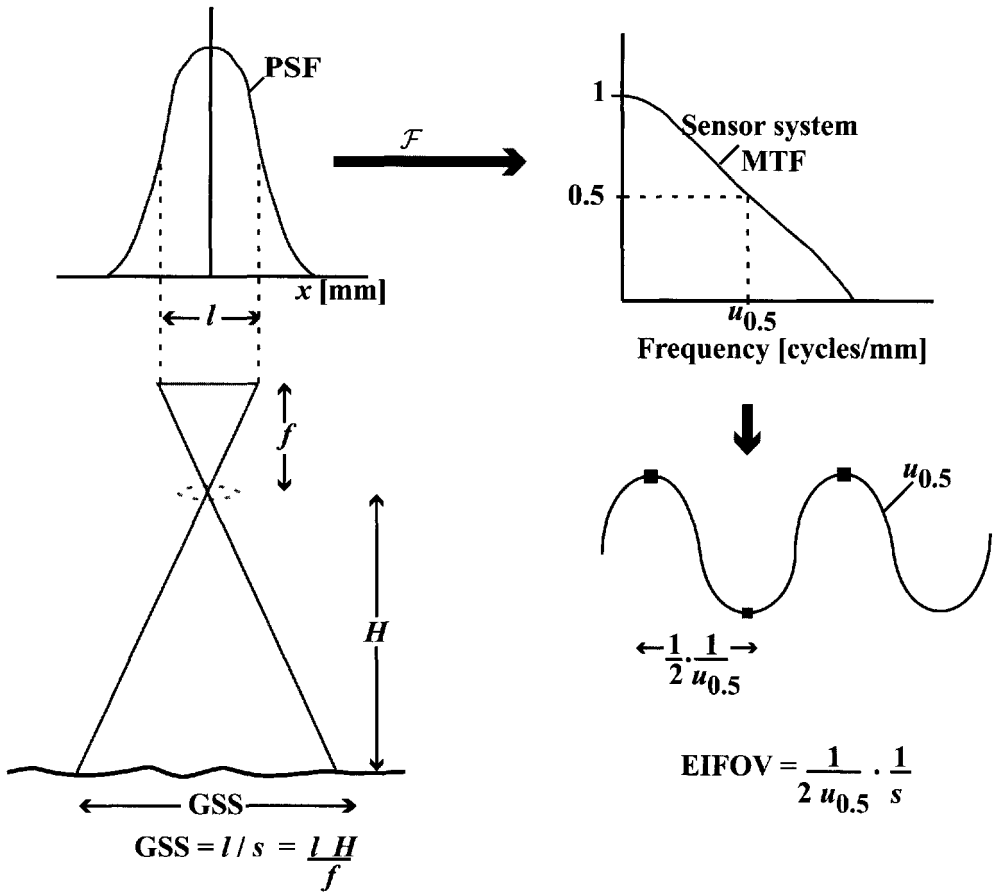


Figure 13.13 GSS and EIFOV resolution metrics derived from a full characterization of the spatial response of an imager.

factors must also be considered. In a well-designed image chain, one would expect that little or no incremental degradation would take place. However, this is not always the case, and only a full image chain assessment will adequately characterize the final image fidelity.

The next major step in the chain is digital image processing. All possible processing operations to which an image may be subjected are far beyond the scope of this effort, so we will treat them in the aggregate for simplicity and mention only two of the most common processes explicitly. Many images, particularly satellite images, are transmitted to ground receiving stations. This process can degrade the spatial frequency content of the image. If the data are still in an analog form, this degradation will be a function of the bandwidth and performance characteristics of the transmission channel. However, most data are in digital form such that degradation due to transmission is much less likely. With today's error correction schemes, digital data can be transmitted with essentially no degradation to the signal.

The most common problem associated with image transmission is that the growing volume of data due to increased spatial resolution, coverage, and number of spectral bands may exceed the bandwidth of the downlink systems. To compensate for this, various forms of *data compression* are employed. When lossless methods

do not provide enough compression, lossy approaches are often employed. The most primitive form of lossy compression is to transmit only some of the data. For example, in a multispectral system, if we know the data are to be used for water-quality studies, and three NIR bands are collected by the sensor, we might choose to only downlink one of the NIR channels. This saves channel capacity so that all the visible channels could be transmitted in a lossless fashion. Since nearly all the information content for water quality is in the visible region and the single IR channel would provide input for atmospheric correction, this might be considered an acceptable "lossy" approach. This approach may be among the most acceptable methods of lossy data compression as we evolve to sensor systems with tens or hundreds of channels. On the other hand, in many cases we may not know the eventual use of the data, and so we want to preserve as much of the full data set as possible. In this case, more conventional forms of lossy data compression are employed that attempt to minimize the distortion in the imagery. Rabbani and Jones (1991) describe a number of these approaches. In general, these lossy compression methods tend to degrade high frequencies in an image and may introduce artifacts. In general, the dominant effect of compression schemes can often be characterized to first order as a filter impacting the spatial frequencies in the image (i.e., the effect of the compression can be estimated by an MTF). There are cases where the impact of the compression on the image is much more difficult to characterize, and changes can be introduced of which the user is unaware and may inadvertently interpret as real phenomena. For this reason, the user is cautioned to evaluate fully the characteristics of a compression algorithm before analyzing compressed data. For our purposes in this discussion, we will limit our considerations to the impact of the compression as a filter on the spatial frequencies by assuming that the compressor has a characterizable MTF.

The other image processing step applied to essentially all images is some form of *image reconstruction*. This process may occur at a central ground station or distribution center, or it may be performed by the individual analyst. In general, it is the step where the geometric sampling of the image is taken into account. It may involve corrections for roll, pitch, and yaw in an airborne system, Earth rotation effects in a satellite system, or registration to a ground coordinate system. For some systems (e.g., radar) this can be a very involved process; for others it may involve only simple line shifting and/or image rotation. In all cases, there is some impact on the image fidelity that can be characterized by the MTF of the reconstruction (or resampling) process. In addition, a number of other image processing steps may be performed prior to display or printing of the image (e.g., radiometric corrections, requantization of gray levels, or noise suppressions algorithms). For convenience, we will treat all of these effects simultaneously by referring to the cascaded product of their MTF as the processing MTF and recognize that it may be asymmetric with respect to the along- and across-track directions. We should also point out that in many cases it is difficult to develop an analytical form for the MTF of these processes, and so empirical methods are employed.

For the hypothetical example system we are considering, there is no data compression, and the primary image reconstruction step would be shifting of lines to correct for aircraft roll. The imaging time is assumed to be short compared to

the pitch-and-yaw periods of the aircraft, so only roll distortion is significant. Furthermore, the frequency of the aircraft roll is low enough that only line-to-line motion (not pixel-to-pixel) need be considered. The line shifting and resampling will result in a slight degradation that can be minimized by the use of cubic convolution resampling, as described in Section 12.3.1.2.

13.1.2.9 Effect of Image Output Devices on MTF

The fidelity of the image presented to the analyst may be further impacted by the display process itself. If the image is interpreted from a monitor, the MTF of the monitor must be evaluated under the viewing conditions to be used for analysis. If the image is to be interpreted from hard copy, then the MTF of the film writing system must be evaluated. Let's take this second case for our example. If the image were to be viewed unaided, then we would need to factor the angular resolution of the visual system into our final calculations. In this case, we will assume that the analyst will be using optical magnification to assist in viewing the images and that this will introduce no significant degradation to the image fidelity. Thus the MTF of the film writer will be the only factor under consideration.

First consider the case where 2048 pixels of each row of the digital scene are written with a continuous tone film writer onto an 8 × 10 inch transparency. We will assume that 7 × 9 inches are actually available for the image. The 2048 pixels in each line would normally be written across the width of the image using the 7 inches addressable, and 2633 (2048•9/7) lines could be written using the 9 inches addressable along the length of the film. However, the electronics that drive the film writer can introduce spatial frequency degradation. More important, the individual spots (pixels) will not be perfectly reproduced due to spreading of the illuminating source in the film writer and diffusion in the film. Overall, the film writer might have an MTF such as illustrated in Figure 13.14. When scaled to the ground, this yields the MTF shown in Figure 13.15. This would indicate that for this overall system the film writer was, by far, the limiting factor. On the other hand, if we were willing to look at only half of the image width at a time written over the whole output image area of the film writer, then the MTF of the film writer would not change, but the image written to it would not contain the high frequencies that are so heavily filtered by the film writer. This comes about because when we scale the film writer MTF to the ground, the spatial frequencies are shifted upward by a factor of 2 when the image scale is doubled. This significantly reduces the adverse effects of the film writer on image fidelity, as shown in Figure 13.15. Obviously other tradeoffs are involved in determining whether writing half a line of the image at a time is an acceptable procedure.

13.1.2.10 MTF of the Entire Image Chain

Figures 13.10, 13.12, and 13.15 show how the various links in the imaging chain can be analyzed, not only to see the net effect on spatial frequencies but also to identify the weak links in the image chain. This list will indicate where efforts to

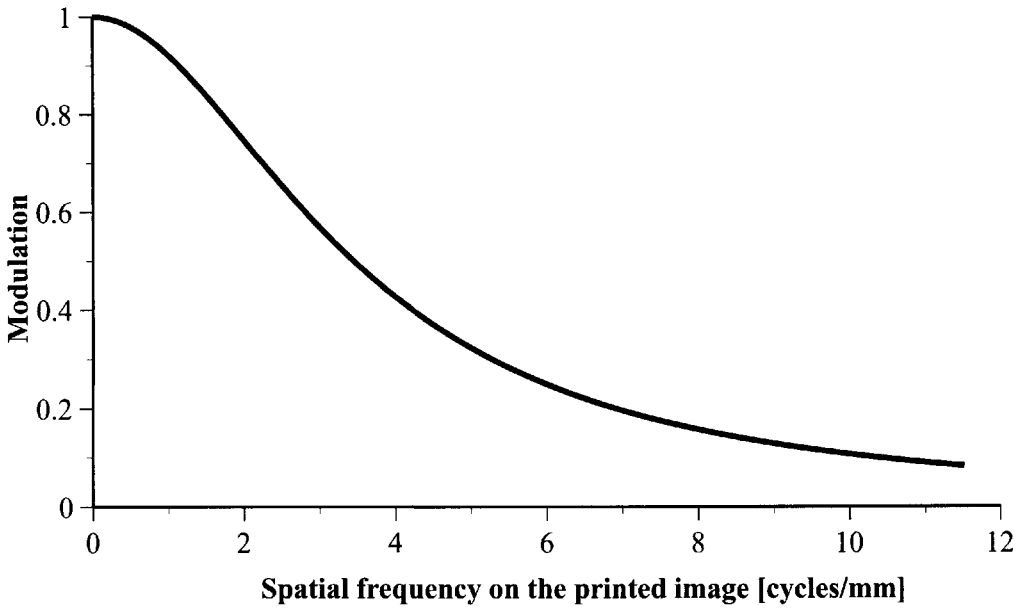


Figure 13.14 MTF of a hypothetical film writer.

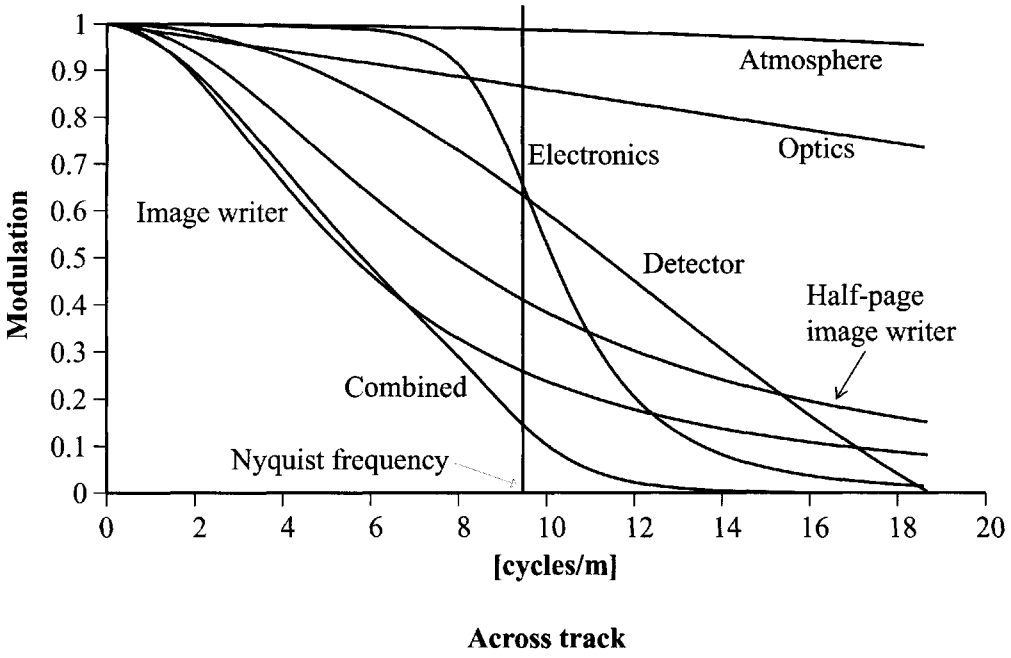
improve system performance would be most productive. In the case treated here, the film writer is by far the weakest link, and improvements elsewhere in the chain would have had almost no impact on final image fidelity. By simply changing the way the film was written, a significant improvement in spatial image fidelity could be achieved. However, the productivity of the unit is reduced by 4, and materials costs quadruple. This solution would be only an interim step until a higher performance film writer could be brought on line.

This section has discussed how image chain analysis of the spatial frequency response of a system can be used in image fidelity assessment and system performance analysis. It is also a powerful tool in end-to-end system design, not only in evaluation of expected system performance but also to avoid over engineering components that are not limiting factors in system performance.

13.1.3 Measurement of, and Correction for, MTF Effects

In the previous section, we explored how the MTF can be used to characterize the spatial image fidelity of components and entire image chains. In this section, we will briefly discuss how the MTF of a system can be measured. It is always best to assess the overall spatial fidelity of a system after final assembly or at least as far down the assembly stage as is physically possible to ensure that the integrated system matches design expectations.

A variety of approaches may be used to perform this type of evaluation. For our purposes, we will consider only two simple methods to illustrate the type of analysis that can be performed. The first method uses step functions (edges) where the brightness drops at a straight edge from a uniform high brightness level to a uniform low brightness. The second method will involve the use of a line source.



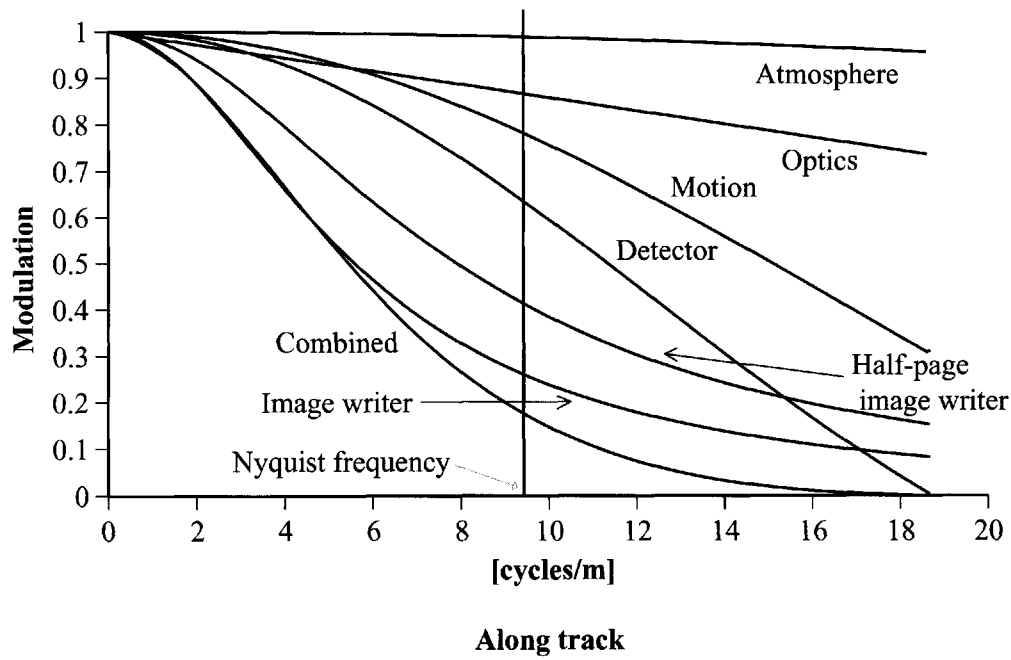


Figure 13.15 MTF of the overall image system.

To begin, let's consider the case of evaluating the MTF of a staring sensor system using a two-dimensional array. We assume that, like most remote sensing systems, the sensor has a fixed focus at infinity, so we would like to evaluate the MTF for the image of an object that appears to be at infinity (i.e., the rays incident on the first components of the sensors are approximately parallel). This is achieved in the laboratory with a collimator, as shown in Figure 13.16. An object at the focal plane of the collimator will appear to be located at infinity to a sensor looking into the collimator. If we introduce a knife edge as the object, then an image of the knife edge should be a step function. In fact, it is a blurred step and can be analyzed as described in Section 13.1.1 to yield the MTF of the entire system along one direction (cf. Fig. 13.4). In this case, the measured MTF would be the product of the sensor MTF and the system (i.e., the collimator). The MTF of the collimator can be treated as known from previous tests (e.g., of distant point sources), so the sensor MTF is then

$$\mathrm{MTF}_s = \frac{\mathrm{MTF}_m}{\mathrm{MTF}_c} \tag{13.18}$$

where MTF_s is the sensor MTF along the axis perpendicular to the edge, MTF_m is the measured MTF, and MTF_c is the MTF of the collimator. This process must be repeated several times at different edge locations to avoid any artifacts due to the location of the projection of the edge onto the individual detectors in the array (phase effects) and the results averaged. This is most often done with a drive stage that steps the edge across the focal plane in step sizes a small fraction of the detector size. This method is sensitive to noise effects because of the derivative operation employed in generating the LSF. Wherever possible, averaging is employed to minimize the adverse impact of noise [e.g., long, narrow sampling slits oriented perpendicular to the edge are scanned over the edge to generate an estimate of the edge spread function (ESF)]. This entire process must, of course, be repeated for the perpendicular direction so that any differences between the along-track and across-track direction are determined. To avoid singularities, the MTF of the collimator should be significantly larger than zero for all spatial frequencies relevant

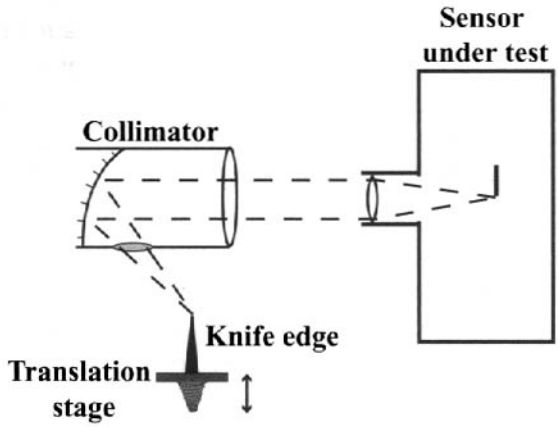


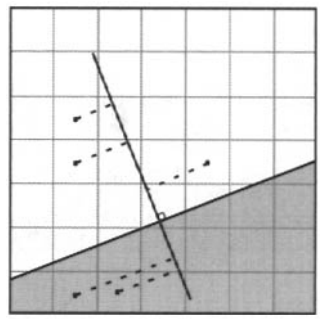
Figure 13.16 Test setup for measuring the MTF of an imaging sensor.

to the sensor. The entrance aperture of the collimator must be at least as large as the sensor entrance aperture (larger for scanning systems). Ideally, an end-to-end sensor test such as this should include all sensor components, including signal conditioning, A-to-D conversion, and recording where applicable. Clearly, the signal should also be sampled along the processing chain so that the component MTFs can also be determined.

Warnick (1990) used a more elegant procedure suggested by Foos and Fintel (1990) to determine the MTF of a digital imaging sensor. This method used a test setup similar to that in Figure 13.16. However, the edge was introduced at an angle to the sensor array, as shown in Figure 13.17. The pixel locations in the vicinity of the edge in the digital image were projected onto a perpendicular to the edge and their brightnesses plotted as a function of the distance from the edge as determined by the projected value. The resultant sampled edge spread function was analyzed using the method of Tatian (1965) to generate the MTF of the measurement setup along the direction perpendicular to the edge. This method has the distinct advantage of reducing the number of experimental measurements required, since phase effects are included in the single measurement. This approach also has the advantage of generating an ESF with reduced noise because many points are included in the estimate, and aliasing can be reduced by sampling at numerous points along the ESF. The final sensor MTF calculation must still use Eq. (13.18) to remove any effects due to the collimator.

This same general approach can be used to compute the MTF of push-broom and scanning sensors. The MTF in the along-track direction can be difficult to measure unless a precise translation stage is used to scan the edge across the sensor. For many sensors, it may be possible to perform much of this end-to-end evaluation using a test range. For example, in calibrating the RIT airborne line scanner, we often locate the scanner on a test pad on the roof of one building such that the scan mirror sweeps perpendicular to the roof of a second building. With the scanner located on a rotating stage, panoramic images can be acquired, as shown in Figure 13.18. Building edges can then be used as step functions to generate the sensor MTF. It is also possible to install test targets (e.g., tribars) on the roof or face of the adjacent structure for use in resolution studies. For this approach to be useful, the range distance must be comparable to the operational imaging distance so that changes in the PSF due to differences in focal depth are negligible. Also, the edges of the buildings must have uniform backgrounds and be good approximations to straight edges at the relevant resolutions. For example, the RIT line scanners tested in this manner were 1- or 2-milliradian systems usually flown at 0.4 km. The buildings used as targets were approximately 0.44 km away, which would result in GIFOVs of 0.44 or 0.88 m. The variation in the edge of the buildings was of the order of 0.015 m and should introduce no error in the computations.

The other source of error in this measurement is turbulence due to the long horizontal path. Because the IFOV of this device is large compared to turbulence effects under low-turbulence conditions, this effect is often negligible. For operational testing, measurements were used only when a visual assessment of turbulence was negligible for angular resolutions significantly higher than those of the sensor under test. Watkins et al. (1991) describes a range target and procedures



Pixel location projected onto perpendicular

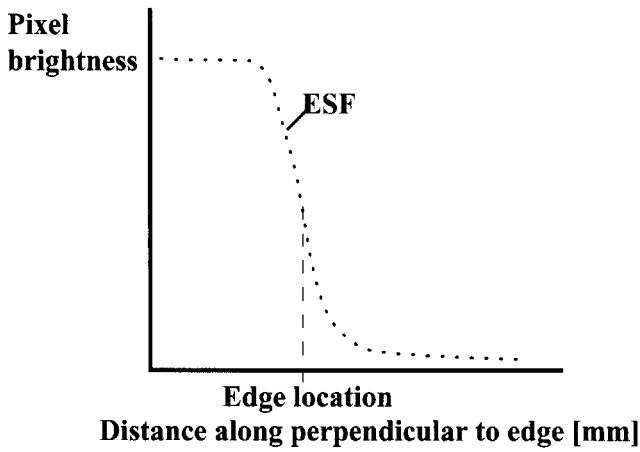


Figure 13.17 MTF measurements using an edge rotated slightly from the detector axes.

used to characterize sensor performance and performance degradations under conditions when turbulence is important. The use of a test range as an alternative or supplement to laboratory bench tests is particularly attractive for sensors with large entrance apertures or large fields of view where the collimators can become very expensive and the physical size of the test arrangement becomes quite cumbersome.

The one feature of the sensor MTF that is usually not tested by either the bench or range tests is blur due to the forward motion of the platform. This is normally treated as simple linear motion and cascaded with the along-track MTF to generate the final MTF of the sensor package. In addition, blur due to variation in the optical axis associated with platform jitter during signal integration is not easily characterized in the lab, and platform dynamics data or full up system tests must be employed.

For some systems, it can be difficult to generate a step function that is both very straight and has uniform high and low brightness. This can be a particular problem in the thermal infrared, where the contrast is usually generated by thermal differences or by emissivity differences with a hot or cold surround. To avoid the

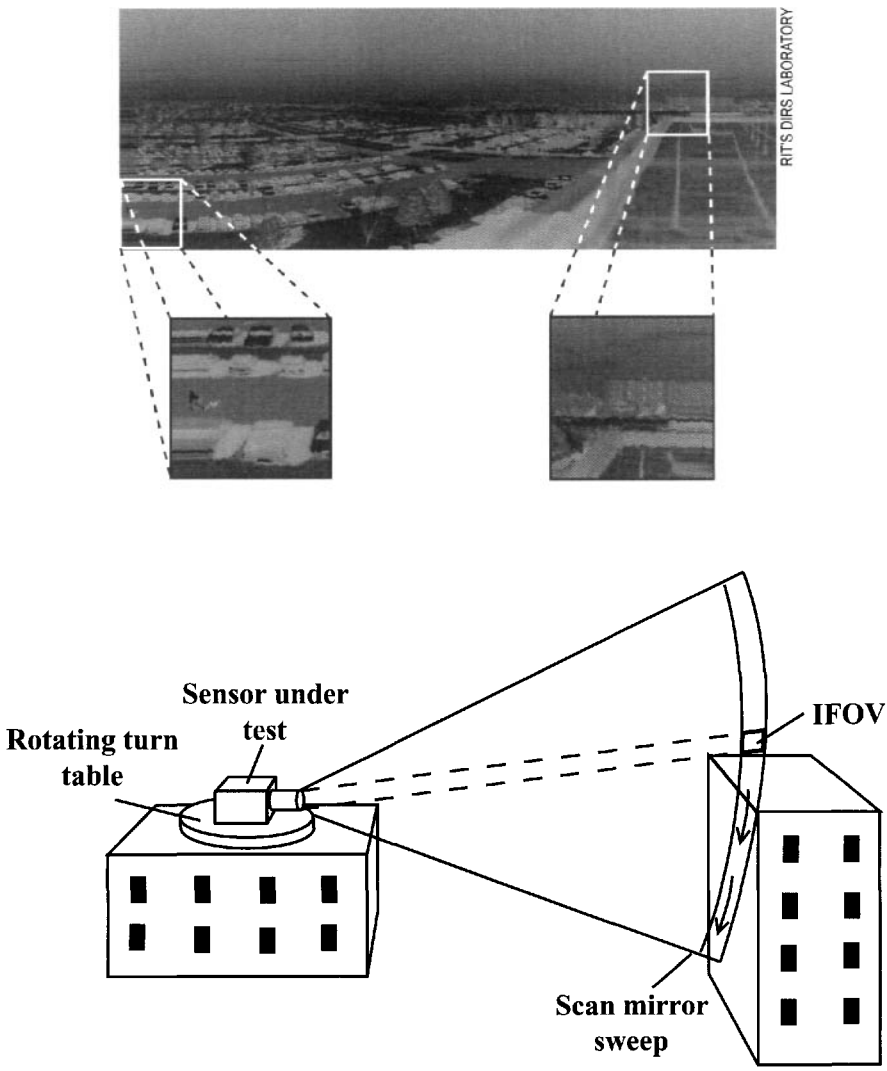


Figure 13.18 Use of a test range for assessment of sensor MTF in the LWIR. Note that image at the far left is before turntable rotation, and horizontal image is due to the motion of the pedestrian.

effort of trying to ensure uniformity, an alternative target is often employed. This is simply a very thin resistive wire located at the focal plane of the collimator. The wire is heated by running a current through it. If a perfect image of the wire would be small compared to the PSF of the system, then the actual image can be treated as a line spread function LSF (an image of a mathematical line source having no width and infinite extent). The LSF image can be analyzed as illustrated in Figure 13.4 to yield the MTF of the sensor.

We have emphasized approaches for characterizing the spatial frequency response of the entire sensor. In general, each component is analyzed in the same fashion. Known signals (sine wave, step, etc.) or image representations of a well-

defined function are input to a link in the chain (e.g., film writer, image reconstruction algorithm), and the output signals (images) are analyzed by treating the output signal (image) as a convolution of the input signal with the impulse response (PSF) of the link under test. The impulse response of the link is then determined by deconvolution, which is most easily done in the frequency domain using Eq. (13.18) or by using simple input functions, so that the approach illustrated in Figure 13.4 can be used. The overall system performance can be obtained by cascading the component MTFs, or the original input images can be passed through the entire image chain and the final images deconvolved to yield the image chain MTF.

For most airborne or space-based systems, even range tests must be treated merely as a baseline assessment, with final performance determined under flight conditions. In many cases, it is very difficult and expensive to perform detailed in-flight assessment of spatial image fidelity, so the normal procedure is to generate detailed performance data before flying the instrument. Flight programs are then used to verify that no changes have occurred due to flight conditions or launch stresses. For airborne systems with reasonably high resolution, the flight tests can be simple replications of the laboratory or range tests employing similar targets on the ground at scales appropriate to the imaging conditions. For lower resolution systems or in the thermal infrared where target costs can become substantial (although the line target is still often affordable), the use of ground targets can become quite cumbersome. Under these conditions, it is often possible to use serendipitous acquisition of targets that approximate the type of well-behaved patterns (lines, steps, points, line pairs) that would be used as test targets. Schowengerdt et al. (1985) describe two variations on this approach for calibration of the postlaunch spatial frequency response of the Landsat TM. In one case, a long, narrow bridge over a uniform water background is used as the input target (cf. Fig. 13.19). The brightness distribution of the bridge (the input function) can be simply defined along an axis perpendicular to the bridge. The deconvolution process can be accomplished in the frequency domain by division, i.e.,

$$\mathrm{MTF}_s(v) = \frac{\mathrm{MTF}_m(v)}{\mathcal{F}[o(l)]} \tag{13.19}$$

where $\mathrm{MTF}_s(v)$ is the MTF of the sensor with respect to spatial frequency (v) along the axis perpendicular to the bridge, $\mathrm{MTF}_m(v)$ is the modulus of the Fourier transform of a subset of the TM image that only contains the bridge and the water background (the modulus is sampled along the axis corresponding to spatial frequencies v), and $o(l)$ is the function that describes the brightness variation of the bridge as a function of distance l along the axis perpendicular to the long axis of the bridge. This approach suffers from two limitations. The first is that the target is a rather unusual feature not readily available in most images, and the second is that the method does not allow for easy separation of the along- and across-track MTF since the result will normally be a composite value. To overcome these limitations, Schowengerdt et al. (1985) demonstrated an approach using high-resolution aircraft data to assess the MTF of lower resolution satellite data. In effect, one assumes that the high-resolution image contains all relevant spatial frequencies and

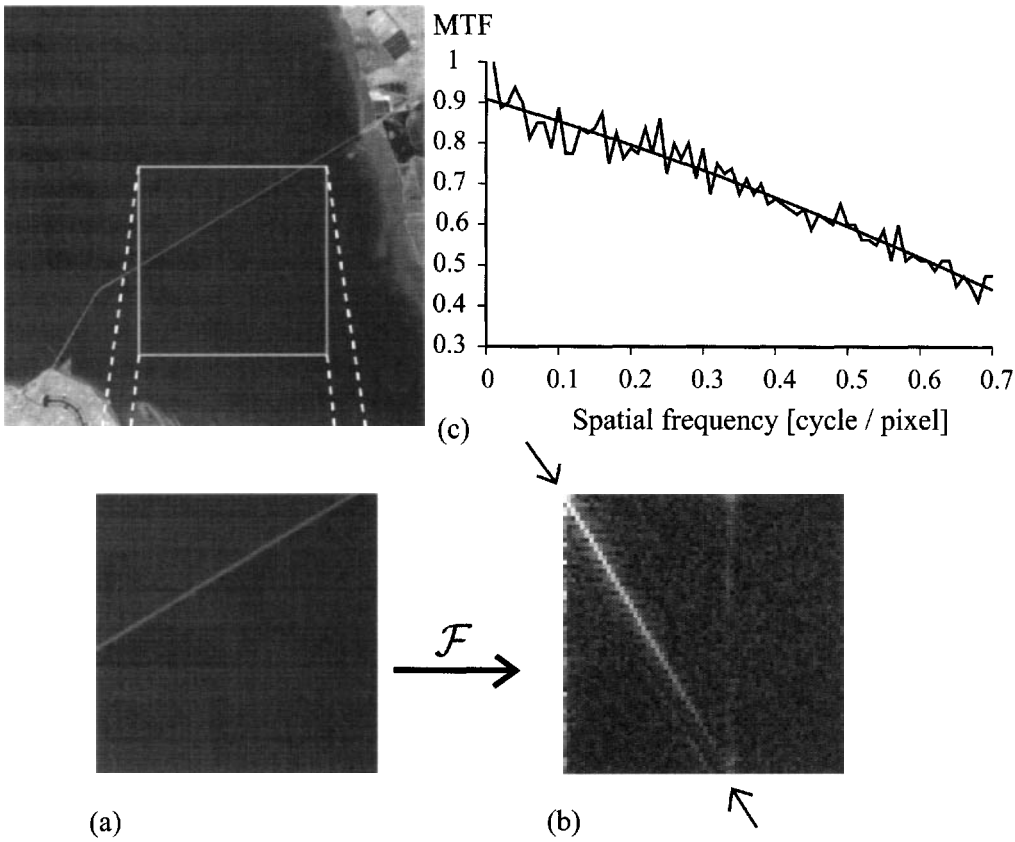


Figure 13.19 Image derived MTF: (a) 128 × 128 pixel subimage of a TM image of the San Mateo Bridge; (b) one quadrant of the Fourier transform of (a) showing the modulus and frequency dimension sampled; (c) resulting sensor system MTF showing noise levels and a smooth polynomial approximation [after Schowengerdt et al. (1985)].

can be treated as the object. The resultant image, after registration and gray-level matching, can be expressed in the frequency domain as

$$\mathrm{MTF}_m(u,v) = O(u,v) \cdot R(u,v) \cdot \mathrm{MTF}_s(u,v) \tag{13.20}$$

where $\mathrm{MTF}_m(u,v)$ is the modulus of the Fourier transform of the satellite image, $O(u,v)$ is the modulus of the Fourier transform of the registered aircraft image, $R(u,v)$ is the MTF associated with the resampling process required for registration, and $\mathrm{MTF}_s(u,v)$ is the sensor system MTF. The resultant sensor MTF_s can then be found from

$$\mathrm{MTF}_s(u,v) = \frac{\mathrm{MTF}_m(u,v)}{O(u,v) \cdot R(u,v)} \tag{13.21}$$

This process is usually quite noisy, and several noise reduction steps are necessary (as discussed in the reference) to yield final MTF values. To be effective, the aircraft imagery must be taken in essentially the same spectral band, at the same time, and from the same perspective as the satellite to ensure the validity of the assumption that it contains the same spatial frequencies as the scene. As a result,

while this method yields a good assessment of the postlaunch performance, it can be expensive and cumbersome to acquire the necessary data. Wherever possible, the use of naturally occurring targets is a more attractive alternative, particularly for operational verification of the ongoing performance of a sensor.

Once the sensor or system MTF is known, it is often desirable to attempt to compensate for the degradation in high-frequency response that is common in most systems. Conceptually, this is quite simple. An *inverse filter* is applied to the observed image by dividing its Fourier transform by the transfer function of the sensor, resulting in a boost of the high frequencies. The inverse transformed image should therefore appear sharper and more closely resemble an undegraded image. In operation, noise in the process often results in the presence of frequencies with zero amplitude in the computed transfer function, thus causing the process to break down. Even where methods are used to avoid zero values in the transfer function, division by small-amplitude values tends to exaggerate periodic noise in the imagery, thus producing visual distractions in the corrected images. More advanced methods can be employed that include models of the image noise in the design of the inverse filter (cf. the treatment of the Wiener filter in Sect. 8.5). With the speed of today's image processors, ad hoc solutions are often developed simply by interactively applying high-frequency boost filters until a visually pleasing result is obtained. Figure 13.20 shows an image processed using a boost filter to enhance the high spatial frequencies. Because these filters are designed to enhance high frequencies, there is often no requirement for preserving radiometry or imagewide statistics. Consequently, the filters are usually implemented in the spatial domain using relatively small convolution kernels often referred to as *modulation transfer function compensation* (*MTFC*) kernels. In any situation where image restoration or boost filters are used, caution should be taken in any ensuing radiometric analysis of the image data to ascertain the impact of the filters on image radiometry. In many cases, it is advisable to restrict radiometric assessment to analysis of relatively large objects before any high-frequency restoration is attempted. By avoiding measurements near edge boundaries, the adverse effects of blurring on the radiometric values can largely be avoided. The boosted images can then be used for visual assessment or algorithms that use texture or spatial analysis.

13.2 RADIOMETRIC EFFECTS

In this section, we will look at how radiometric image fidelity can be characterized through the image chain. Like spatial image fidelity, radiometric image fidelity can also be characterized in many ways and at many points along the image chain. A major distinction involves questions of absolute versus relative radiometry. In absolute calibration, we are concerned with how closely the measured value can be matched to a set of external reproducible standards. In relative calibration, the primary interest is internal consistency (i.e., does a change of 10 units always represent the same effect even if we don't know how many absolute units those 10 units represent?). When speaking of sensors, this is often thought of in terms of radiance. A sensor calibrated in absolute units typically would have internal

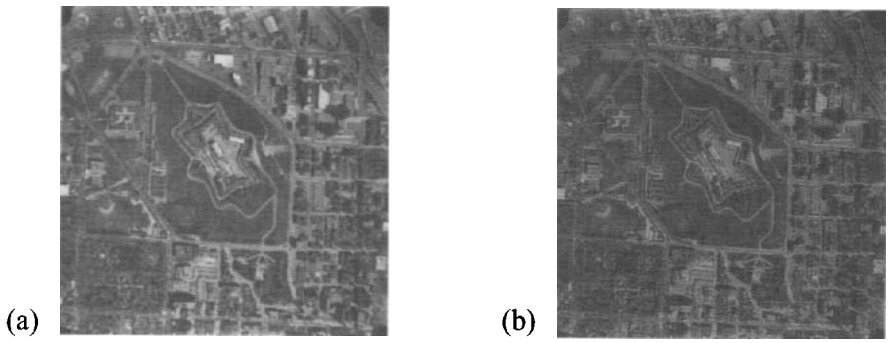


Figure 13.20 Effect of a high-frequency boost (MTFC) filter: (a) original image and (b) image after high-frequency boost.

sources of known absolute radiance, so each reading could be converted to an absolute value. A sensor with relative calibration might consistently return the same value when exposed to a specific radiance level, but the absolute value of that radiance could not be inferred. Alternatively, a system with a lesser degree of relative calibration might be able to determine only that a change of so many radiance units occurred between point A and point B, but not what the radiance level was. In general, as discussed at length in Chapter 7, the calibration of the sensor is only one step in the process, since our interest is typically rooted in how well we can measure parameters such as ground temperature or reflectance. The image chain thus includes all radiation propagation effects, sensor effects, and analytical procedures discussed in Chapters 4 through 7. In analyzing the mean-level radiometric performance along the image chain, we would have to assess all the contributions to the governing equation and their impact on the final measurements using error propagation modeling, as discussed in Section 4.6.2. We will not reiterate these issues here except to recall that the discussions in those chapters implicitly assumed a unit MTF response for all measurements. Clearly, as we saw in the last section, readings on small objects or near edges will not accurately represent the radiance that should be associated with those targets (small here is usually measured in terms of several EIFOVs).

Our discussion in previous chapters often assumed that the measured values were simply a function of radiance as characterized by the governing equations of Chapters 4 through 7. We recognize, however, that nature is not this simple, and there are variations about the mean-level values described by the governing equations. We collectively refer to these deviations as noise, and in this section we will look briefly at the propagation of noise through the image chain.

13.2.1 Noise

Noise is perhaps best characterized as our uncertainty as to whether an individual measurement represents the value of a parameter. As discussed in Section 5.3.2, it is generally characterized in terms of the deviation from a mean value. Perhaps

more important, in terms of image chain analysis, uncorrelated noise sources tend to add in quadrature such that

$$n_T = (n_1^2 + n_2^2)^{1/2} \tag{13.22}$$

where n_T is the total noise and n_1 and n_2 are the noise levels from two uncorrelated noise sources. The nature of quadrature addition is such that the largest noise source(s) drastically dominates the total noise. Thus, when trying to reduce the noise level of the entire imaging system, it is important to have identified the magnitude of the noise sources in each link so that mitigating measures are applied where they will have some impact.

When considering noise sources, we recognize that the various sources of noise that influence the signal level from the detector can be treated collectively as one source of noise (n_d). The components of detector noise were introduced previously in Section 5.4. It is useful to recall that we can think of this noise in the measured units (i.e., voltage or current), in terms of radiometric input units (i.e., noise-equivalent power), or in terms of target parameters (i.e., noise-equivalent temperature difference or noise-equivalent reflectance difference). The preamplifier and conditioning electronics can also be sources of noise (n_a). For analog systems, the record and playback systems will be additional noise sources (n_r). In digitizing the signal, the continuous input signal is quantized into bins, which results in additional uncertainty in the signal level that is characterized as quantization noise (n_q). Oppenheim and Schafer (1975) point out that, when the quantization value is at the center of a bin of width b, and the input values are uniformly distributed, the standard deviation in *quantization noise* n_q can be expressed as

$$n_q = b/\sqrt{12} \tag{13.23}$$

Depending on the image chain, additional noise sources will exist and must be included in the computation of the overall system noise (n), i.e.,

$$n = (n_d^2 + n_a^2 + n_r^2 \cdots + n_q^2)^{1/2} \tag{13.24}$$

In general, the noise values will be expressed in terms of electronic signal levels (amps or volts). The concept of noise-equivalent input units can still be used to convert overall system noise into intuitive values (i.e., change in apparent temperature or reflectance) using the principles described in Section 5.3.

These standard noise concepts are generally thought of as variations about a mean level. However, in terms of the impact of noise on sensor performance, more complex issues come into play having to do with the interaction of noise levels, the spatial structure of the noise, and the spatial response (MTF) of the system. This introduces the problem that all noise is not equal, as illustrated in Figure 13.21, where the same image is shown with two sources of noise having the same RMS noise levels. In Figure 13.21(b), the noise is random, and while the image is degraded, it is not nearly as distracting as in Figure 13.21(c), where the same RMS noise is shown as a patterned or correlated noise. Thus, we see that the distribution of the spatial frequency of the noise can be a factor in the impact of the noise. This

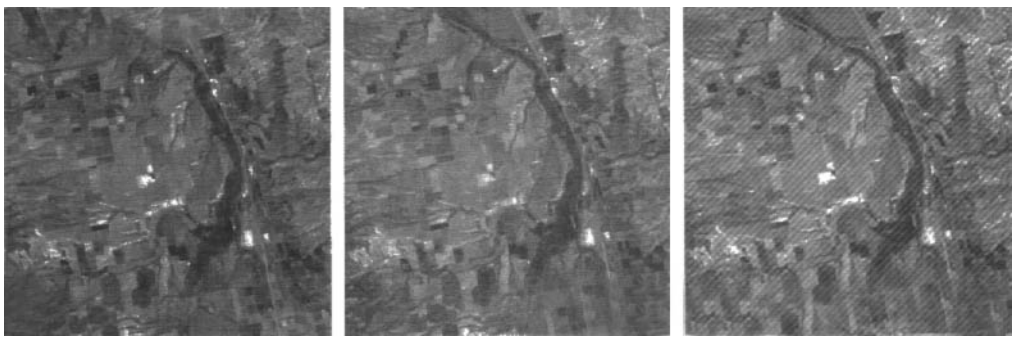
is a visual analysis problem that also impacts machine-aided exploitation where algorithms will be more sensitive to noise at certain frequencies.

The relative impact of noise levels and spatial frequency distribution on image fidelity is very often a function of the procedures to be used in the image analysis. A detailed assessment of these complex interactions is beyond the scope of our interest here. However, note that the image fidelity issues are often addressed by relying on an empirical metric. For example, in the case of thermal infrared systems, a common metric is the *minimum resolvable temperature difference* [*MRTD*; cf. Holst (1993)]. This is the temperature difference that must exist between a set of four bar targets (assumed blackbodies) for the bar pattern to be just visually discernible. The MRTD is specified under a set of test and observation conditions (e.g., target scale or viewer conditions). The MRTD of a system as a function of spatial or angular frequency will be a function of noise levels, noise structure, and the MTF of the entire system. Thus, as a figure of merit, it considers all of the important features. It has the disadvantage of not being fully modeled, so the MRTD under other conditions must be obtained by experiment, which can be a tedious process.

Other performance-based metrics have been proposed for other sensors and/or applications. In general, they tend to be application specific and can be modeled as a complex function of the noise level, noise power spectrum, and system MTF. We will restrict ourselves here to stating the need to characterize these functions as indicators of image fidelity, recognizing that in general a complete set of data is needed to characterize image fidelity and that only in restricted cases can image fidelity be simplified to a single metric.

13.2.2 Noise Artifacts

The type of periodic noise shown in Figure 13.21 is usually caused by electronic or mechanical variations (or their harmonics) in sensor systems that induce noise in the detector or preamplifier. Other sources of periodic, correlated, or patterned noise are caused by detector-to-detector calibration differences. In linear arrays



(a) (b) (c)

Figure 13.21 Impact of noise structure on the appearance of an image: (a) original image, (b) image with random noise added, (c) image with periodic noise added. The RMS noise levels of the noise added in (b) and (c) are the same.

and whisk-broom-style detectors, the digital count out of each detector is a function of the incident radiance. The radiance-to-digital-count conversion is normally computed separately for each detector. Any slight error in calibration can result in the output digital counts for that detector being slightly high or low (or having slightly more or less gain). This will result in a periodic striping in the image, which will be most apparent in uniform areas or when the contrast has been highly stretched. This effect is very pronounced when a detector fails. This striping will be in the across-track direction for whisk-broom scanners and in the along-track direction for push-broom systems.

Another type of periodic noise occurs associated with groups of lines. This often results from slight gain or bias differences in groups of detectors. For example, in a bow-tie scanner, the detectors may be calibrated at the start of each line, and then a slight bias drift (droop) can occur across the line. For example, in the Landsat TM, all 16 detectors in each reflective band are calibrated against onboard references at the start of each line. Because the mirror scans in both directions, any droop effects occur on opposite sides of the image in groups of 16 lines. This results in a banding structure in the across-track direction, which can become apparent in uniform regions when the contrast is stretched. Similarly, in a linear array made up from many elements (cf. Fig. 6.15), the readout electronics may be slightly different for each element, causing small gain or bias differences between the groups of detectors in the elements. This will result in an along-track banding pattern (cf. Fig. 13.22).

A number of other artifacts often exist in raw sensor imagery. Most of these are very sensor specific and require the development of sensor-specific correction algorithms to reduce or eliminate the effects. For example, in linear array systems it is common to see streaking or smearing after a very bright target. This occurs in the readout direction because the charge is not completely transferred out of the cell and some is left to be added in with the succeeding charge packet. This results in a bright streak behind very bright objects. Some of these artifacts can be very difficult, or even impossible, to characterize, to the point where the impact can be successfully modeled for removal from the imagery. In the next section, we will briefly treat some of the more straightforward methods for reducing noise artifacts.

13.2.3 Approaches for Correction of Noise and Periodic Structure in Images

With very large detector arrays, individual detector failures become highly probable. When failures occur, several correction options exist. The most straightforward is to simply replace the affected line with an adjacent line from a working detector. This results in some spatial artifacts but does not change the radiometric value of any of the pixels. An alternative method is to use one of the interpolators discussed in Section 12.3.1.2. These interpolators will reduce the spatial artifacts but may introduce radiometric artifacts in the interpolated line.

One method to reduce the effects of detector-to-detector variation in a whisk-broom or bow-tie system is to use the imagewide histograms for each detector (e.g., for the Landsat TM this would be the histogram made up of every 16th line of

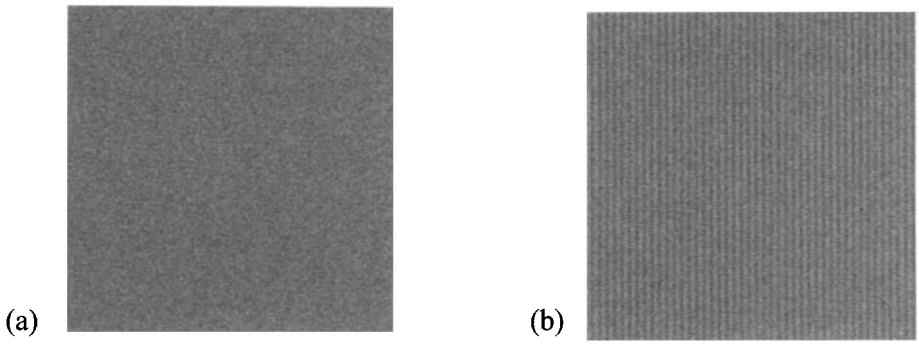


Figure 13.22 Examples of noise structure: (a) random noise and (b) structured noise.

raw data). Over many, many lines, the mean and standard deviation of these detector specific histograms should be the same. Any slight difference can be attributed to gain or bias calibration differences between the detectors. These calibration differences can be corrected by forcing the histograms to be the same, using the linear histogram matching algorithms described in Section 7.5 for the PIF scene normalization process. The values for each detector are adjusted for gain and bias differences from some norm (either the detector with maximum contrast or the average of the image-wide histograms). A similar approach can often be used for certain types of banding. Histograms of pixels from a common subelement in a linear array are matched to another element's histogram using linear histogram specification (i.e., a gain and bias correction). Because subelements can be very long, the scene structure can sometimes induce real variations in the histograms. To avoid this affect, only homogeneous regions of the scene are used in generating each element's statistics.

To this point, we have concentrated on approaches for dealing with image artifacts. There are also methods available for reducing the effects of more random noise, but they often have some negative attributes. The baseline methods rely on the fact that the random noise in a signal will be reduced by averaging multiple samples. The noise reduction being proportional to $(1/N)^{1/2}$, where N is the number of samples averaged. Thus, one way to reduce noise in an image is to simply replace every pixel with the average in an N pixel neighborhood. The resulting image should have the noise reduced by $(1/N)^{1/2}$. This process is easily implemented using a convolution kernel, as discussed in Section 8.2. However, the process acts like a low-pass filter and will blur the output image. As discussed in Section 8.2, if the noise is more of a salt-and-pepper type (random spikes and drop-outs), a nonlinear filter (e.g., median) may be more effective, although this, too, will introduce some artifacts. Rather than average spatially and suffer the resultant blur, it is sometimes possible to average multiple images of the same scene. For example, if the sensor is a rapid framing system that produces many images over a short period, multiple frames can be averaged together. The random noise will again be reduced by $(1/N)^{1/2}$. This approach can be very effective if there is little

relative motion between the scene and the sensor over a period of many frames. In this case, the successive images will be in registration, and no spatial blurring will occur. Figure 13.23 illustrates some of the tradeoffs associated with spatial and temporal averaging for noise reduction.

It is also possible to take advantage of time averaging in the design of systems that are not of the rapid framing style. This alternative method uses multiple detectors in the along-scan direction of a scanning system or in the along-track direction of a push-broom system (cf. Sec. 6.1). As the image moves across the detectors, it is sampled by the first detector, and then essentially the exact same image point on the ground is sampled a moment later as it is swept over the next sensor. By using several detectors in the image motion direction, the same point on the ground is imaged several times. These sampled values can then be averaged to reduce noise. This process is referred to as *time delay and integration* (*TDI*) because each sampled signal is held a short time (while the image advances over the detectors) and is then summed with the signal from each succeeding detector. The final integrated value should have its noise level reduced by $(1/N)^{1/2}$, where N is the number of detectors in the TDI process. This method has the advantage that with carefully controlled timing circuits, there should be little or no spatial blurring. Since the time between successive detector acquisitions is very small, even moving objects will normally have very limited blur. The negative side to this approach is the additional cost and complexity in the sensor design. The TDI process is normally implemented on board the sensor (often right on the focal plane) to avoid recording multiple images. It is possible, where recording bandwidth is available, to record images from each successive detector and shift and average the images during ground processing to achieve the TDI effect.

In many cases, external sources of electronic or mechanical noise that are induced into the detector or preamp will result in periodic noise patterns. When these patterns are identified in preflight conditions, they can often be eliminated by improved shielding and isolation of the detector, preamplifier, or noise source. Regrettably, these noise sources often develop in flight and must be removed in post-processing. If the noise is periodic, it will cause localized increases in the power spectrum of the image. The particular frequencies can be identified by noise spikes in the power spectrum and filtered in the frequency domain using a very localized band rejection filter as discussed in Section 8.5 and illustrated in Figure 8.17. This will result in loss of information at the filtered frequencies, so the filters should be kept small and used with caution.

13.3 SPECTRAL AND POLARIZATION EFFECTS

A potential problem source in any imaging chain, but one that is taking on increasing importance as we move to multispectral analysis tools, is a shift in the system spectral response after ground spectral calibration. This can be a simple shift in band centers such as might result in a slight mechanical deviation in a grating in a spectrometer, or it can be due to a change in the bandpass of filters such as that reported for narrow band interference filters by Flittner and Slater (1991). It is

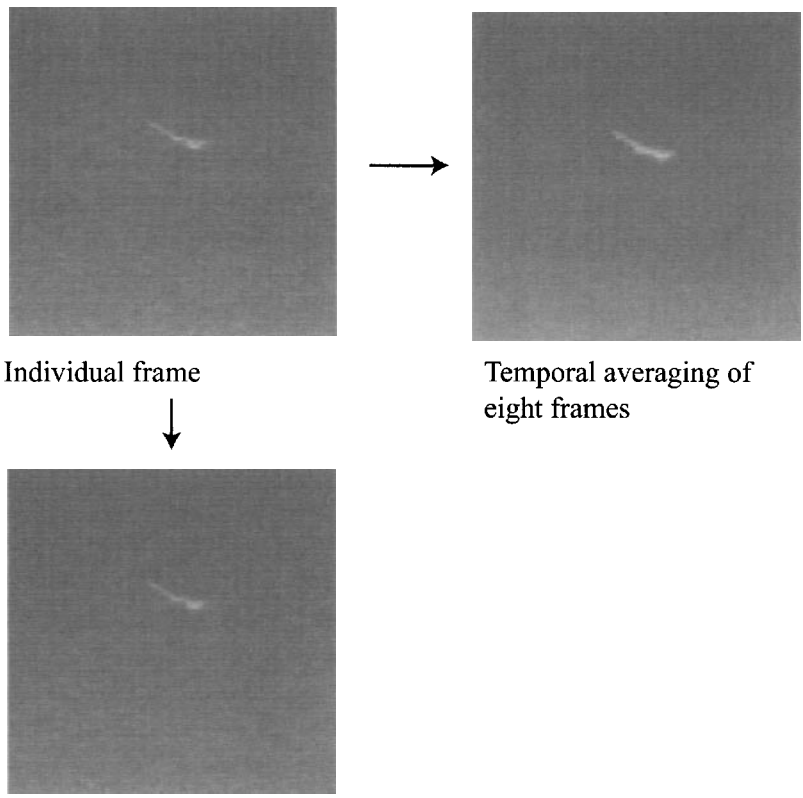


Figure 13.23 Effects of spatial and temporal averaging.

believed that the interference filters used for band selection underwent mechanical changes in the space environment that resulted in changes in the spacing of the layers in the filter. The net result of these changes was a shift in the band edges. If these spectral changes go unnoticed, they can introduce radiometric calibration errors, atmospheric correction errors, and misinterpretation of spectral signatures. The relative importance of these changes will depend a great deal on the calibration techniques used and on where the spectral shift occurs relative to spectral structure in the target or the atmosphere (i.e., a very small change at the edge of an atmospheric window could have a major impact, whereas a large change in the middle of a window or in a spectral region where the target and backgrounds were slowly varying would have limited impact). It can be very difficult to detect and characterize these changes in satellite systems. In most cases, a careful analysis of well-known targets with spectral structure near the band edges must be performed to identify and characterize the problem in instruments with bandpass filters. In the case of spectrometers, errors are often easier to detect if the edges of an atmospheric absorption feature are sampled. Many of these absorption features are well characterized (cf. Sec. 3.5). They can be used to identify errors in spectral calibration, and the spectrometer recalibrated in an iterative fashion by fitting the observed spectra to the absorption line structure [cf. Green et al., (1994)]. Qu et al. (2002)

describe an approach for spectral fitting of observed spectra to the shape of atmospheric absorption features that yields estimates of the spectral response functions (band centers and FWHM values) of the spectrometer. Spectrometers designed to study slight variations in line spectra may require more involved on-board or in-flight verification methods. For example, laser line sources or filters (typically using absorption, not interference to achieve absorption) can be introduced with strong spectral absorption features as part of the on-board calibration process. Flux from these lasers, or solar flux passed through the filter, can be introduced onto the focal plane in a fashion similar to that used for the radiometric calibration procedures discussed in Section 6.4. As early as 1973, Earth-based laser beams were imaged by space platforms [cf. Piech and Schott, (1975)], and this practice may be required in the postlaunch characterization of the spectral calibration of satellite systems if on-board spectral sources are not available.

13.3.1 Feature/Spectral Band Selection

For systems with more than a few spectral bands, problems arise associated with the high volumes of data. In many cases, this becomes a data storage or transmission problem (i.e., we are producing data faster than we can move it down the image chain). In other instances, it is an analysis problem produced by having so much data to process that the algorithms run too long. In the case where the data are to be used in multivariate classifiers, it is often not necessary, or even desirable, to process all of the data. This is particularly the case when feature bands (spectral bands) are highly correlated. In this instance, adding more bands to a classifier can actually reduce classification accuracy. Several methods have been suggested to attempt to isolate optimal or near-optimal subsets of features for use with GML classifiers. Swain (1978) describes three measures referred to as divergence, Jeffries-Matusita distance, and transformed divergence. These are all measures of class separability, which, when averaged over all classes, yield a measure of the discrimination quality of a group of spectral features. By comparing all possible combinations of subsets of the spectral features (e.g., which 4 out of 12 available bands), the one that produces the highest quality metric can be selected. Only the reduced subset of bands are then used in the overall image classification. Swain (1978) indicates that if the classes can be assumed to be normally distributed, then the *divergence* between class i and class j can be expressed as

$$Div_{ij} = \frac{1}{2} tr\left[\left(\mathbf{S}_i - \mathbf{S}_j\right)\left(\mathbf{S}_i^{-1} - \mathbf{S}_j^{-1}\right)\right] + \frac{1}{2} tr\left[\left(\mathbf{S}_i^{-1} + \mathbf{S}_j^{-1}\right)\left(\mathbf{m}_i - \mathbf{m}_j\right)\left(\mathbf{m}_i - \mathbf{m}_j\right)^{\mathrm{T}}\right] \quad (13.25)$$

where Div_{ij} is the divergence between class i and class j and tr[**A**] indicates the trace of the matrix **A**; the rest of the terms were defined in Section 9.2. The first term in the equation for divergence characterizes the difference between the covariance matrices (i.e., the difference in the shape of the distributions), and the second term is a measure of the normalized statistical distance between the means. Overall, the divergence is a measure of the separability between classes i and j. The average divergence is a weighted summation over all pairwise combinations of classes, i.e.,

$$\overline{Div} = \sum_{i=1}^{k} \sum_{j=1}^{k} \rho(i)\rho(j) Div_{ij} \tag{13.26}$$

where the weights are simply the a priori probabilities. The average divergence of each possible subset of ℓ spectral bands can then be computed, and the subset that yields the largest average divergence will tend to yield the best separability. A problem arises, however, when one combination of bands generates classes with very large divergence values for some classes and small values for others, and a second generates modest divergence values for all classes, as illustrated in Figure 13.24. Clearly, the second case represents a better overall pairwise selection of features. This indicates that increasing the pairwise divergence has a diminishing return. Swain (1978) indicates that this limitation is overcome by the *Jeffries-Matusita distance*, but at considerable computational cost. A more commonly used heuristic approach is the *transformed divergence,* expressed as

$$Div_{ij}^{\mathrm{T}} = 2\left[1 - e\left(-Div_{ij}/8\right)\right] \tag{13.27}$$

This has the characteristic of exponential saturation of the divergence measure and scales the transformed divergence over the range 0 to 2 [Mausel et al., (1990)], in evaluating separability measures, used a scaling factor of 2000 rather than 2, and this factor is widely used). The average transformed divergence can then be computed by substituting the transformed divergence in Eq. (13.26).

Rosenblum (1990) describes a similar metric that was used to separate texture-based image features (cf. Sec. 8.3) and to select spectral bands based on reflectance spectra in sensor design studies. This metric, referred to as thresholded *separability*, uses a ratio of the statistical distance to a user-defined maximum distance value. The ratio is then thresholded, so that any value greater than 1 is set equal to 1. The separability measure of class j from class i takes on the intuitively appealing form of

$$\mathrm{Sep}_{ij} = \frac{|\mathbf{S}_i| + (\mathbf{m}_j - \mathbf{m}_i)^T \mathbf{S}_i^{-1} (\mathbf{m}_j - \mathbf{m}_i)}{\mathrm{Sep}_{\max}} \tag{13.28}$$

where the numerator looks very much like the GML discriminant function between class means for the case of equal a priori probabilities (cf. Sec. 9.2) and the denominator is chosen to ensure very good, but not extreme, separability between class means. The threshold operation ensures that the problem of excessive separability in one class does not overwhelm the final metric (cf. Fig. 13.24). The overall separability can then be defined as

$$\overline{\mathrm{Sep}} = \sum_{i=1}^{k} \sum_{j=1}^{k} \mathrm{Sep}_{ij} \tag{13.29}$$

and the subset of features that maximizes the overall separability would be chosen as near optimal. Rosenblum (1990) found that the separability measure performed comparably to the transformed divergence with improved speed.

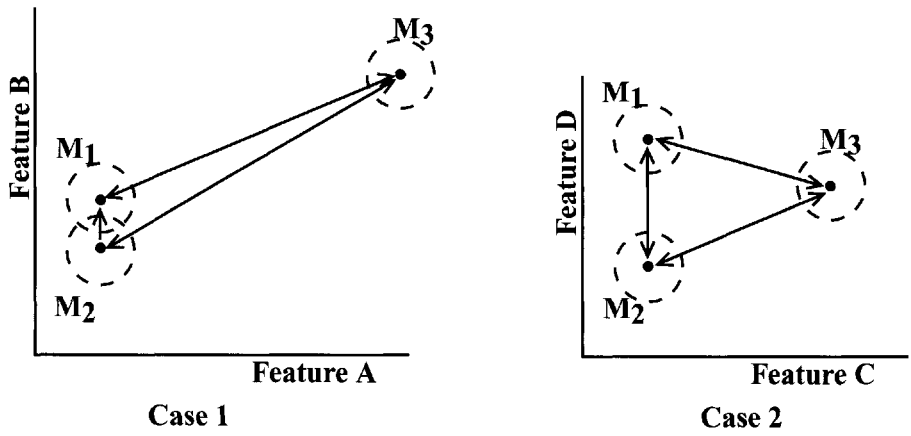


Figure 13.24 Limitation of simple divergence. A feature set that has overlap in clusters but has one cluster center very far from all the others can appear better than a set with approximately the same separation in all centers.

All of the feature selection methods presented in this section are scenario dependent. This means that the features (e.g., spectral bands) chosen are nearly optimal for classification of the target classes used in the optimization study. In order to use this approach in deciding which bands to downlink or record, robust studies of optimum band combinations would need to be performed for various scenarios (i.e., various target classes) and imaging conditions.

13.3.2 Polarization Issues

To this point, we have largely ignored the effects of polarization on the imaging process. Implicitly, we have assumed that either the flux incident on the detector was randomly polarized or the sensor was insensitive to polarization effects (i.e., flux at any polarization would yield the same signal level). These assumptions are often invalid, and in some cases may lead to substantial errors. For example, grating spectrometers tend to be very polarization sensitive, and most EO imaging systems tend to have some polarization-dependent sensitivity, unless great effort has gone into minimizing this sensitivity. As a result, if the radiance reaching the sensor is polarized, it will cause different signal levels depending on the orientation of the polarization. Simply put, a polarization-sensitive sensor calibrated with randomly polarized flux will be out of calibration for anything but randomly polarized flux. One way to deal with this problem is to design polarization-insensitive sensors and to characterize the polarization sensitivity so that any error due to residual sensitivity can be tracked. An alternative approach is to attempt to use the polarization state as a signature. In cases where the energy matter interactions at the Earth's surface induce a change in the polarization state of the incident flux, we can use this change (which will be a function of the makeup and condition of the surface) as a signature. In order to take advantage of this type of feature, our sensor must be able to discriminate variations in the polarization state of the incident flux. One simple way to do this is to orient cross-polarized filters in front of matched

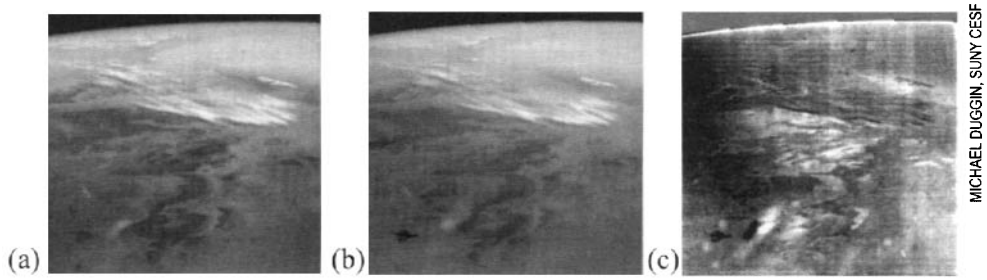


Figure 13.25 Effects of scene-induced polarization differences: (a) and (b) are space shuttle images cross polarized relative to each other, and (c) is a polarization difference image.

sensors. Ideally, orientation of the filters is varied to maximize the difference between the two images formed. Figure 13.25 shows an example of a pair of images obtained with a polarization-sensitive system. In this case, the system consisted of two conventional film cameras with polarization filters on each. The filters are oriented at right angles to each other, and the cameras are synchronized for image acquisition [cf. Duggin et al., (1989)]. Differences between the image pairs are then a function of the polarization effects of the scene elements (note that image-to-image normalization is usually performed first to account for cross-calibration effects and polarization effects in the atmospheric parameters).

By using multiple filters in conjunction with the multispectral scanner or push broom designs discussed in Chapter 6, it is possible to acquire up to a full four-element *Stokes vector* representation, if four filters are used (cf. Fig. 3.11). This allows a calculation of the *degree of polarization* (*DOP*) of the polarimetric radiance field, *degree of linear polarization* (*DOLP*), and the *degree of circular polarization (DOCP)* (cf. Sec. 3.2.3).

The degree of terrain-induced polarization difference will be a function of wavelength, surface roughness (relative to the wavelength), and sun-object-sensor angle. Rough surfaces will tend to randomly scatter the flux-inducing random polarization, while smoother surfaces will tend to more selectively scatter relative to polarization. This difference in the amount of energy reflected into perpendicular or parallel components will tend to be rather small when the incident flux is near normal and become more exaggerated at grazing angles [cf. Hecht (1987)]. The images shown in Figure 13.25 were taken at a low view angle to enhance polarization effects. Note that in the difference image, smooth surfaces will tend to have higher differences and therefore appear brighter in the image. This effect is used to advantage in Figure 13.26, where the DOLP image shows that man-made (smooth) objects tend to induce more linear polarization in the reflected radiance than do rougher background surfaces.

The extent to which polarization is an issue is very much a function of viewing geometry, target structure, and wavelength [cf. Whitehead (1992)]. Rough targets viewed from nadir will tend to have little polarization, so the impact of the

polarization sensitivity of the sensor on calibration is reduced. On the other hand, for quantitative measurements of smooth surfaces from an oblique perspective, great care must be taken to ensure that the sensor is not polarization sensitive or that the polarization effects are fully known (which is often difficult to document if the instrument was not intended for polarization studies).

Travis (1992) points out the importance of the polarization effects of the atmosphere on polarization imaging. Rayleigh scattering introduces significant polarization that is strongly a function of the scattering phase angle (cf. Sec 3.4.2.1). He suggests that polarization effects in the visible region are dominated by atmospheric effects, with terrestrial polarization becoming more important as we move into the NIR, where the mean albedo (particularly over land) is higher and Rayleigh scatter appreciably reduced. The presence of aerosols tends to reduce the polarization induced by Rayleigh scatter. Travis (1992) suggested that multispectral polarimetric imagers may be able to use this reduction in polarization to characterize the aerosol content. The French POLDER instrument flown onboard the Japanese ADEOS 1 and 2 spacecraft was designed to study both atmospheric aerosols and the coupled Earth atmosphere bidirectional reflectance using multispectral polari-

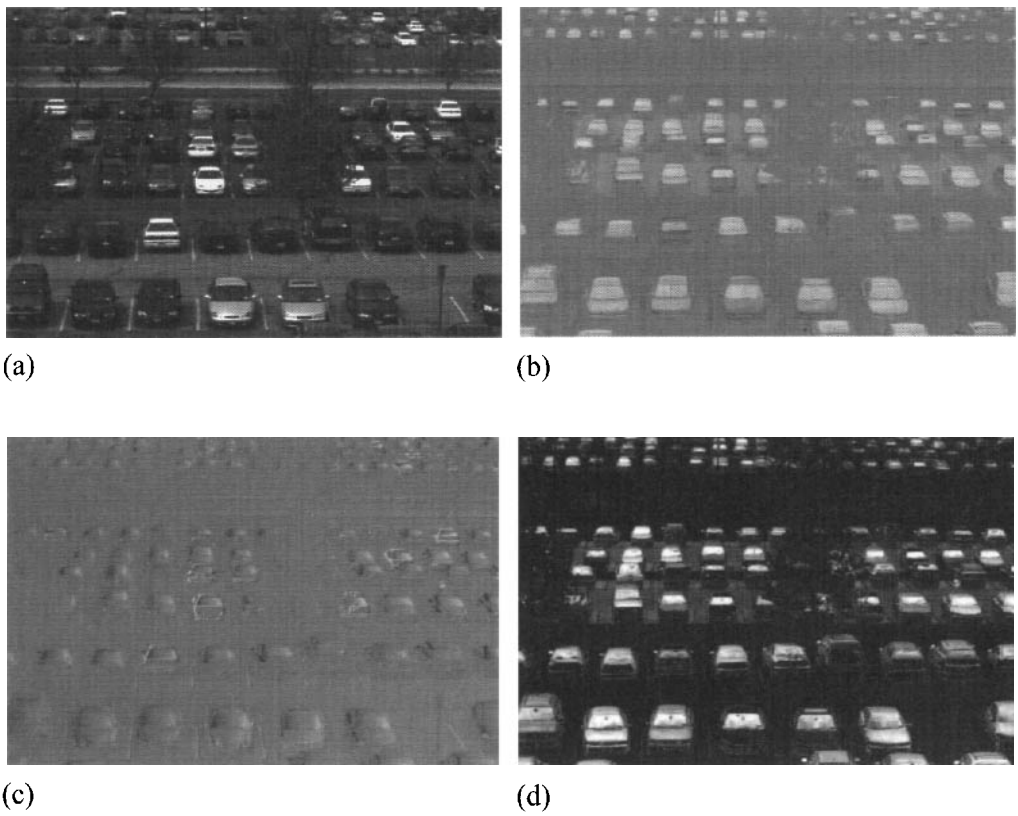


Figure 13.26 An image of a parking lot taken with a polarization-sensitive sensor (see Fig. 3.11): (a) unpolarized component (S_0), (b) S_1 component, (c) S_2 component, and (d) DOLP computed from (a), (b), and (c).

metric imaging techniques. The low-resolution (≈ 6 km) system is designed to use polarimetric analysis tools to study global aerosols [cf. Deschamps et al. (1994)]. These data are of considerable value to both climatologists interested in studying the atmosphere and to remote sensing scientists interested in reducing atmospheric effects.

13.4 SPATIAL, SPECTRAL, AND RADIOMETRIC TRADEOFFS

For the most part, we have tended to treat the radiometric and spatial resolution image chains as separate entities. While this is often a convenient simplification, it is also useful to recognize that these image chains are fully interwoven. This is most easily seen if we think about observing a scene where the radiance is varying at some spatial frequency v. Using the $G\#$ concept developed in Eq. (5.9), the variation in flux onto the detector in the absence of MTF effects should be

$$\Phi(v) = \int \frac{L_\lambda(v) A_0 d\lambda}{\mathrm{G\#}} \tag{13.30}$$

where A_0 is the detector area [cm^2]. In fact, Eq. (13.30) is only valid if the spatial MTF is unity, which for most systems would only be true at very low frequencies. The actual expected variation in flux $\Phi'(v)$ should more rigorously be expressed as

$$\Phi(v) = \Phi'(v) = \mathrm{MTF}(v) \int \frac{\mathrm{L}_\lambda(v)}{\mathrm{G\#}} A_0 d\lambda \tag{13.31}$$

where for convenience we have assumed that the system MTF is constant over the spectral bandpass of the detector, and the prime (′) is dropped for clarity. The variation in the signal (S) observed when viewing a scene whose radiance is varying at a spatial frequency v could then be expressed as

$$S(v) = \mathrm{MTF}(v) \int \frac{R(\lambda) L_\lambda(v)}{\mathrm{G\#}} A_0 d\lambda \tag{13.32}$$

where $R(\lambda)$ is the sensor's spectral responsivity.

However, Eq. (13.32) is only valid in the absence of noise [$N(v)$]. The actual expected signal S' would be

$$S(v) = S'(v) = \mathrm{MTF}(v) \int \frac{R(\lambda) L_\lambda(v)}{\mathrm{G\#}} A_0 d\lambda + N(v) \tag{13.33}$$

where the prime (′) is again dropped for clarity. The signal to noise can then be estimated as

$$\frac{S(v)}{N(v)} = \frac{\mathrm{MTF}(v) \int \frac{R(\lambda) L_\lambda(v) A_0 d\lambda}{\mathrm{G\#}}}{N(v)} + 1 \cong \frac{\mathrm{MTF}(v) \int \frac{R(\lambda) L_\lambda(v) A_0 d\lambda}{G\#}}{N(v)} \tag{13.34}$$

For convenience, we have expressed the parameters in Eqs. (13.30) to (13.34) as a function of a single spatial frequency dimension (v). Clearly, for imaging systems, these variables are often more effectively addressed in terms of the spatial frequencies (u,v) with respect to the x and y axis [i.e., $S(v)$ becomes $S(u,v)$, etc.].

In the formalism introduced in Chapter 8, Eq. (13.30) would be expressed as

$$G(u,v) = H(u,v) \bullet F(u,v) + N(u,v) \tag{13.35}$$

where $G(u,v)$ is the frequency domain representation of the observed image (i.e., the 2-D Fourier transform of the image), $H(u,v)$ is the 2-D MTF of the entire image chain leading up to the observed image, $F(u,v)$ is the frequency domain representation of the image in the absence of any MTF effects, and $N(u,v)$ is the frequency domain representation of the noise. We choose to use the representation in Eq. (13.33) at this point to show explicitly the interconnection between the strands in the image chain. Equation 13.33 shows us how completely the radiometric and spatial strands of the image chain are interwoven and reminds us that if, for convenience, we treat them separately, we must recombine the strands or subchains to understand any subtle interaction effects that can be introduced. These interactions are particularly important when we consider efforts at making precise radiometric measurements on objects with high spatial frequency (i.e., where the MTF deviates significantly from unity).

13.5 IMAGE-QUALITY METRICS

Consideration of all of the strands in the image chain leads us to the question of what makes an image good. For visual analysis, a widely used measure of image quality is the *National Imagery Interpretability Rating Scale* (*NIIRS*) [cf. Leachtenauer (1996)]. The NIIRS scale is based on a subjective assessment of an analyst's ability to perform exploitation tasks ranked in difficulty over a 0 to 9 scale. It has become a common measure of the image quality associated with reconnaissance systems [cf. Riehl and Maver (1996)]. Table 13.2 lists examples of the tasks/criteria associated with NIIRS levels. Criteria have been developed for visible, thermal infrared, radar, and multispectral imagery. A trained analyst can compare a given image to a list of NIIRS criteria and determine an appropriate ranking even if none of the specific targets are in the image. The original NIIRS rating values were set up so that a factor of 2 change in GSD would result in approximately a one unit change in NIIRS. However, as discussed in the previous section, GSD is clearly not the only metric governing image interpretability. Leachtenauer et al. (1997) describe the *general image quality equation* (*GIQE*), which was developed to relate the subjective NIIRS scale more closely to quantities measurable in an image or to be predictable based on system design parameters. The GIQE is based on regression analysis of measured parameters versus analyst-defined NIIRS values for a large number of images over the range of NIIRS values. Leachtenaur et al. (1997) express the GIQE for VNIR NIIRS as

$$NIIRS = 10.251 - aLog_{10}GSD_{GM} + bLog_{10}RER_{GM} - 0.656H_{GM} - 3.44G/SNR \tag{13.36}$$

where GSD_{GM} is the geometric mean of the GSD (in inches), RER_{GM} is the geometric mean of the *relative edge response*, H_{GM} is the geometric mean of the *height overshoot* caused by edge sharpening, G is the *noise gain* due to edge sharpening, SNR is the signal-to-noise ratio, and $a = 3.32$ and $b = 1.559$ if RER > 0.9 or $a = 3.16$ and $b = 2.817$ if RER < 0.9. The adjusted R^2 value for this relationship was found to be 0.986 with a standard error of 0.282 NIIRS units. The terms in this relationship are discussed in more detail below. An important factor to recognize is that the GSD term, because of its large range, dominates the relationship with the RER term, also playing a strong but lesser role. The two sets of coefficients were developed to account for an apparent change in the relationship for low RER values.

Because NIIRS is widely used to characterize image quality and because NIIRS values can be well represented by the GIQE, it is useful from an image-chain perspective to look at each of the terms in the GIQE equation.

The GSD is the sample distance in each direction projected onto the ground and expressed in inches with the geometric mean expressed as

$$GSD_{GM} = (GSD_x \cdot GSD_y)^{\frac{1}{2}} \tag{13.37}$$

with

$$GSD_x \approx \frac{x}{f} \cdot H_{sr} \tag{13.38}$$

and

$$GSD_y \approx \frac{\text{pixel pitch}}{f \cos\theta} \cdot H_{sr} \tag{13.39}$$

Table 13.2 Examples of NIIRS Rating Levels

NIIRS Value	Example Criteria
0	Imagery is uninterpretable
1	Distinguish between major land use classes (e.g., urban, agriculture, forest, water, barren)
2	Indentify road patterns (e.g., clover leafs on major highway systems
3	Detect individual houses in residential neighborhood
4	Count unoccupied railroad tracks along right-of-way in a rail yard
5	Identify individual rail cars by type (e.g., gandola, light box)
6	Identify automobiles as sedans or station wagons
7	Identify individual railroad ties
8	Identify windshield wipers on a vehicle
9	Detect individual spikes in a railroad

for a push-broom instrument, where x is the along track sample distance on the focal plane, H_{sr} is the slant range, f is the focal length, and θ is the view angle from nadir. Since the GSD is often a good estimate of the GSS of a system, it is understandable why it should dominate the NIIRS expression. Furthermore, the NIIRS ranking was originally set up so that a step size of one NIIRS would correspond to a factor of 2 change in GSD, which explains the $3.32 \log_{10} GSD_{GM}$ term. If we just consider the first two terms in the NIIRS equation (i.e., the bias term and the GSD term), we see that a 2-inch GSD results in an NIIRS of 9.3, a 10-inch GSD an NIIRS 6.9, a 100-inch GSD an NIIRS of 3.6, and a 1000-inch *GSD* an NIIRS of 0.3. Thus, we can see that the GSD term dominates the range of the NIIRS scale, with the remaining terms controlling the subtleties of the scale.

The RER is the slope of the edge spread function (ESF), as illustrated in Figure 13.27. It will normally take on values ranging from 0.2 to 0.95, which will introduce very small delta NIIRS changes for RER values near one and may introduce delta NIIRS adjustments of one to two NIIRS for low RER values. A NIIRS change of 0.2 or more is usually considered significant. Recall from the discussion in Chapter 8 that the Fourier transform of the ESF is the MTF, and so the RER can be computed from empirical measurements of edges in the imagery or from experimental or predicted values of the MTF. High MTF values near the Nyquist limit will result in high RER values. Thus, we see that the system MTF can play a significant role in the final NIIRS value.

The overshoot height is the magnitude of the normalized edge response induced by an MTFC filter. As illustrated in Figure 13.28, it is calculated over the range of 0 to 3 pixels from the edge. If the normalized edge response is increasing over this range, the value at 1.25 pixels is used. If a maximum occurs, the value at the maximum is used. These measurements can be made empirically or, more typically, by numerical convolution of the system PSF and the MTFC kernel with an oversampled edge. Like the GSD and the RER, this metric can vary substantially in the x and y directions, and the geometric mean (square root of the product) is used in the NIIRS calculations. Note that the overshoot value will typically be close to one, and the effect of the overshoot term in the NIIRS calculation is to lower the NIIRS value by a delta NIIRS of a fraction to about one unit. The greater the MTFC boost, the greater the correction, but since the term is negative, it will reduce the NIIRS value. This term can be thought of as correcting for the fact that an MTFC kernel will tend to increase the RER value over what just the system MTF would produce. The height overshoot term tends to compensate for the exaggeration of the sharpness in the RER calculation.

The noise gain term (G) is simply the square root of the sum of the squared values in the normalized MTFC kernel, i.e.,

$$G = \left[\sum_{j=1}^{M} \sum_{i=1}^{N} k_{ij}^{2} \right]^{\frac{1}{2}} \tag{13.40}$$

where

$$\sum_{j=1}^{M} \sum_{i=1}^{N} k_{ij} = 1 \tag{13.41}$$

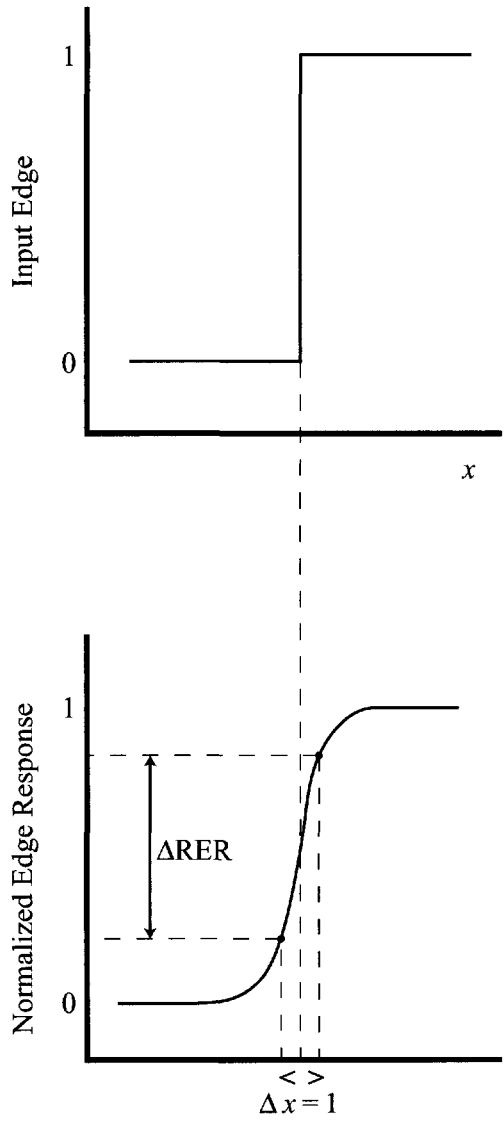


Figure 13.27 Illustration of the relative edge response, which is the slope of the normalized edge spread function evaluated at ±0.5 pixels of the nominal edge location.

and k_{ij} is the i,jth entry in the MTFC kernel. Inspection of Figure 13.29 shows that the noise gain value will increase with the boost associated with the MTFC kernel. The G value attempts to capture the negative impact of the visual amplification of noise effects by the MTFC sharpening kernel. The importance of the noise gain value is scaled by the SNR value since boosting the noise in a low-SNR image is likely to have a much more deleterious effect than a comparable noise boost in a high-SNR image. The SNR for NIIRS calculations is defined as the change in signal observed between a 7% and 15% reflector divided by the standard deviation in the signal for an extended uniform target, i.e.,

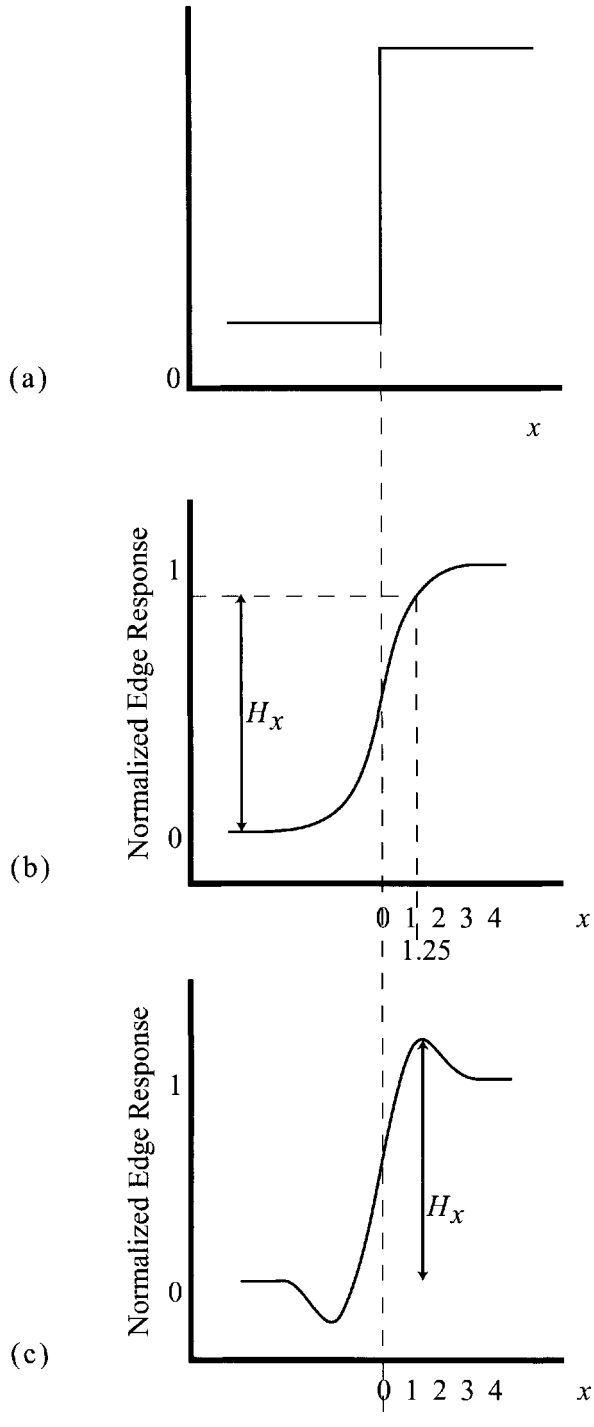


Figure 13.28 Illustration of height overshoot computation for edges from two different systems: (a) cross section of an edge, (b) shows a monotonically increasing function over the range 0 to 3 pixels that would be evaluated at 1.25 pixels and (c) a peak induced by the MTFC boost that would be evaluated at the peak (the sampling is done on 0.25 pixel centers).

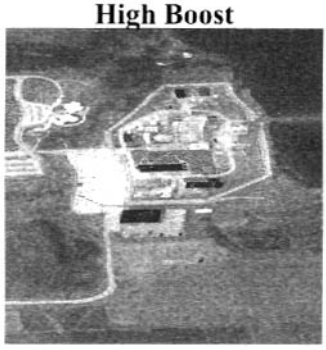
$$SNR = \frac{DC_{15} - DC_7}{\sigma_{DC}} \tag{13.42}$$

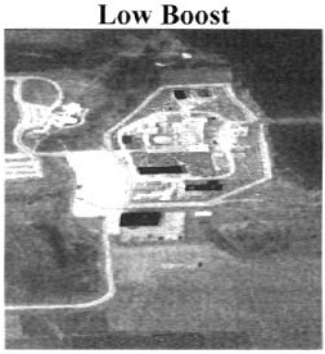
With SNR values of 50 or more being common, it is clear that this term will only become important for high-gain, low-SNR cases where the effect of exaggerated noise gain will be a reduction in NIIRS.

Figure 13.30 shows the same image at different NIIRS levels to provide the reader a more intuitive feel for the relative changes associated with NIIRS levels. The discussion above yields insight into how various components of the image chain impact image quality as perceived by an analyst charged with visual exploitation of the imagery. However, as we move to machine-based analysis, the factors that make up image quality have not been as neatly codified as they have been with the NIIRS scale and the GIQE. In fact, this is a function of the number of exploitation algorithms available. If we ask about the quality of an image, the answer may be heavily driven by how we analyze the data (i.e., what algorithm or sequence of algorithms). An image that yields good results for one analytical approach may yield poor results for another. Shen (2003) and Kerekes and Hsu (2004a) describe some initial steps towards generation of spectral image quality metrics/equations that might be applicable to hyperspectral data. These studies are task and potentially even algorithm specific. However, they demonstrate how major system parameters such as spatial resolution, spectral resolution and signal to noise can be related to performance levels using regression studies similar to those used to generate the GIQE. Shen (2003) indicates that scene-dependent parameters (e.g., scene complexity) can be important metrics driving a spectral quality equation, and Kerekes and Hsu (2004b) show that the impact of various terms (e.g., spatial resolution) can be dramatically different when spectral quality metrics are developed for different tasks (i.e., target detection vs. land cover classification).

13.6 SUMMARY OF IMAGE CHAIN CONCEPTS

In this chapter, we have looked at some of the characteristic limitations placed on imaging systems by spatial, radiometric, and spectral resolution. In an ideal system, we would like to simultaneously extract information about fine spatial structure, subtle brightness changes, and detailed spectral character. However, one of the fundamental axioms of remote sensing is that there are never enough photons to let us image everything we want. To improve signal to noise for improved radiometric precision, we need more signal (more photons). This can be accomplished with larger detectors, longer dwell times, or bigger optics. For a fixed optical system, this generally means sacrificing spatial resolution, since larger detectors and/or increased dwell time will generally degrade resolution. Similarly, when we image in narrower spectral channels to improve spectral resolution, the number of photons available to be detected is reduced, decreasing signal to noise. Again, the spatial resolution must usually be sacrificed to maintain sufficient signal to noise to discern radiometric differences in the spectral image. Thus, we can think





	Noise Gain	Kernel
High Boost	$\sqrt{29}$	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$
Medium Boost	$\sqrt{10}$	$\begin{bmatrix} 0 & -0.5 & 0 \\ -0.5 & 3 & -0.5 \\ 0 & -0.5 & 0 \end{bmatrix}$
Low Boost	$\sqrt{2}$	$\begin{bmatrix} 0 & -0.1 & 0 \\ -0.1 & 1.4 & -0.1 \\ 0 & -0.1 & 0 \end{bmatrix}$

Figure 13.29 Illustration of several simple normalized MTFC kernels and the noise gain term computed for the kernels.

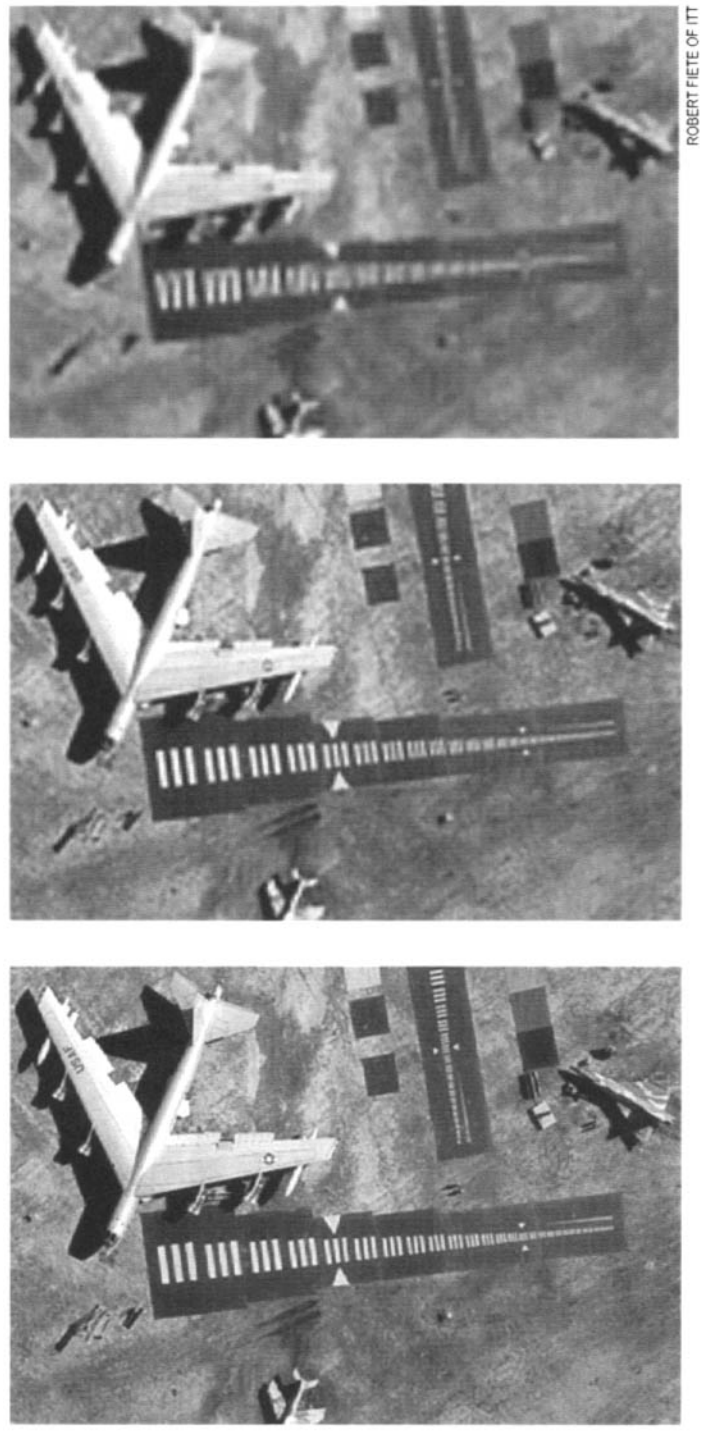


Figure 13.30 Examples of the same image at different NIIRS levels illustrating the change in image quality associated with different NIIRS values. Recognize that the NIIRS assigned to these images was for the original image and not for the version printed here.

of spatial, spectral, and radiometric resolution as comprising a three-dimensional tradeoff space.

What is not well defined is how to perform tradeoff studies in this three space. Ideally, we would like to be able to determine uniform utility surfaces in this space that would tell us how one combination of system parameters relates to another (e.g., is a three-meter, one-band data set with a signal-to-noise of approximately 150 "better or worse" than a nine-meter, three-band data set with similar signal-to-noise?). One tempting approach to addressing this question is to use classical information theory as developed by Shannon (1949). This approach, which is commonly used in channel capacity calculations, uses entropy as a measure of information. In the simplest case, the *entropy* can be expressed as

$$H = \sum p(DC) \ln_2 (p(DC)) \tag{13.43}$$

where H is the entropy expressed in bits per pixel, and p(DC) is the probability of a digital count occurring.

Pratt (2002) discusses how higher order entropy expressions can be developed that take into account the correlation effects between neighboring pixels or different spectral bands using joint probabilities. These higher order entropy calculations provide a method for comparing the minimum information storage or channel capacity. However, because they rely on intensive calculations and are not necessarily directly related to effective measures of information from an application standpoint, this approach has not been extensively pursued. At present, there is no generally accepted rigorous theoretical solution to this question. This is because the "effective information content" of an image is a function of the algorithms and analytical tools that are available for extracting the information from the image (cf. Sec. 13.5). For example, classifier-based tradeoffs have been used to select optimum spectral bands and/or texture metrics for use in land cover mapping (cf. Sec. 9.2; i.e., pick the bands that yield the best classification as described in Sec. 13.3). Because monochrome and multispectral algorithms are often very different, even application-specific comparisons can be very difficult (i.e., a change in algorithms at a later time could change the decision about the most appropriate spatial-spectral-radiometric trade). At present, we must be satisfied with recognizing the importance of these tradeoffs and accept that design decisions are still often made on the basis of application-specific algorithms, often using metrics that are still evolving.

13.7 REFERENCES

Beland, R.R. (1993). "Propagation through atmospheric optical turbulence." Chap. 2 in. Smith, F.G., ed., The IR and EO Handbook Vol. 2. SPIE, Bellingham, WA.

Coltman, J.W. (1954). The specification of imaging properties by response to a sine wave input. *Journal of the Optical Society of America*, Vol. 44, No. 6, pp. 468-471.

Deschamps, P.-Y., Breon, F.-M., Leroy, M., Podaire, A., Bricaud, A., Buriez, J.C., & Seze, G. (1994). The POLDER mission: instrument characteristics and scientific objective. *Geoscience and Remote Sensing, IEEE Transactions*, Vol. 32, Issue 3, pp. 598-615.

Duggan, M.J., Israel, S.A., Whitehead, V.S., Myors, J.S., & Robertson, D.R. (1989). Use of polarization methods in Earth resources investigations. *Proceedings SPIE*, Vol. 1166, pp. 11-22.

Fiete, R.D. (1999). Image quality and [lambda]FN/p for remote sensing systems. *Optical Engineering*, Vol. 38, No. 7, pp. 1229-1240.

Flittner, S.E., & Slater, P.N. (1991). Stability of narrow-band filter radiometers in the solar-reflective range. *Photogrammetric Engineering and Remote Sensing*, Vol. 57, No. 2, pp. 165-171.

Foos, D., & Fintel, W. (1990). Eastman Kodak Company, Personal communication.

Gaskill, J.D. (1978). Linear Systems, Fourier Transforms, and Optics. Wiley, NY.

Goodman, J.W. (1968). Introduction to Fourier Optics. McGraw-Hill, NY.

Granger, E.M., & Cupery, K.N. (1972). An optical merit function (SQF) which correlates with subjective image judgements. *Photographic Science and Engineering*, Vol. 15, No. 3, pp. 221-230.

Green, E.O., Conel, J.E., Margolis, J., & Chrien, T.G. (1994). "Spectral calibration of an imaging spectrometer inflight using solar and atmospheric absorption bands." IGARSS '94, Surface and Atmospheric Remote Sensing: Technologies, Data Analysis and Interpretation, California Institute of Technology, Pasadena, CA.

Hall, C.F., & Hall, E.L. (1977). A non-linear model of the spatial characteristics of the human visual system. *IEEE Transactions on System, Man, and Cybernetics*, Vol. SMC-7, No. 3, pp. 161-170.

Hecht, E. (1987) Optics. 2nd ed., Addison-Wesley, Reading, MA.

Holst, G.C. (1993). "Infrared imaging system testing." Chap. 4 in Dudzik, M.C. ed., Vol. 4, IR and EO System Handbook, Electro-optical system design, analysis, and testing. SPIE Press, Bellingham, WA.

Kaufman, Y.J. (1982). Solution of the equation of radiative transfer for remote sensing over nonuniform surface reflectivity. *Journal of Geophysical Research*, Vol. 87, No. C6, pp. 4137-4147.

Kerekes, J., & Hsu, S. (2004a). Special quality metrics for VNIR and SWIR hyperspectral imagery. *Proceedings SPIE*, Vol. 5425, pp. 549-557.

Kerekes, J., & Hsu, S. (2004b). Spectral quality metrics for terrain classification. *Proceedings SPIE*, Vol. 5546, pp. 382-389.

Leachtenauer, J.C. (1996). "National Imagery Interpretability Rating Scales: overview and product description." Proceedings ASPRS Annual Convention, Vol. 1, pp. 262-272.

Leachtenauer, J.C., Malila, W., Irvine, J., Colburn, L., & Salvaggio, N. (1997). General image-quality equation: GIQE. *Applied Optics*, Vol. 36, Issue 32, pp. 8322.

Mausel, P.W., Kramber, W.J., & Lee, J.K. (1990). Optimum band selection for supervised classification of multispectral data. *Photogrammetric Engineering and Remote Sensing*, Vol. 56, No. 1, pp. 55-60.

Oppenheim, A.V., & Schafer, R.W. (1975). Digital Signal Processing. Prentice-Hall, Engelwood, NJ.

Pearce, W.A. (1977). "A study of the effects of the atmosphere on thematic mapper observations." Final Report under NASA Contract NAS5-23639.

Piech, K.R., & Schott, J.R. (1975). "Evaluation of Skylab earth laser beacon imagery." Calspan Report #KL-5552-M-1, prepared for NASA.

Pratt, W.K. (2002). Digital Image Processing. 3rd ed., Wiley, NY.

Qu, Z., Kindel, B.C., & Goetz, A.F. (2002). Atmospheric correction for two classes of hyperspectral imaging sensor. *Proceedings SPIE*, Vol. 4725, pp. 83-94.

Rabbani, M., & Jones, D.W. (1991). Digital image compression techniques. SPIE Optical Engineering Press, Vol. TT7.

Rees, W.G. (1990). Physical Principles of Remote Sensing. Cambridge University Press, Cambridge, NY.

Riehl, K., Jr., & Maver, L.A. (1996). Comparison of two common aerial reconnaissance image quality measures. *Proceedings SPIE*, Vol. 2829, pp. 242-254.

Rosenblum, W. (1990). "Optimal selection of textural and spectral features for scene segmentation." Masters Thesis, Center for Imaging Science, Rochester Institute of Technology, Rochester, NY.

Schowengerdt, R.H., Archwamety, C. & Wrigley, R.C. (1985). Landsat thematic mapper image derived MTF. *Photogrammetric Engineering and Remote Sensing*, Vol. 51, No. 9, pp. 1395-1406.

Shannon, C.E. (1949). The Mathematical Theory of Communication. University of Illinois Press, Urbana, IL.

Shen, S.S. (2003). Spectral quality equation relating collection parameters to object/anomaly detection performance. *Proceedings SPIE*, Vol. 5093, pp. 29-36.

Slater, P.N. (1980). Remote Sensing, Optics, and Optical Systems. Addison-Wesley, Reading, MA.

Swain, P.H. (1978). "Fundamentals of pattern recognition in remote sensing." In P.H. Swain and S.M. Davis, eds., Remote Sensing: The Quantitative Approach. McGraw-Hill, New York, NY.

Tatian, B. (1965). Method for obtaining the transfer function from the edge response function. *Journal of the Optical Society of America*, Vol. 55, pp. 1014-1019.

Travis, L.D. (1992). Remote sensing of aerosols with the earth observing scanning polarimeter. *Proceedings SPIE*, Vol. 1747, pp. 154-163.

Warnick, J.S. (1990). "A quantitative analysis of a self-emitting thermal IR scene simulation system." M.S. Thesis, Center for Imaging Science, Rochester Institute of Technology, NY.

Watkins, W.R., Crow, S.B., & Kantrowitz, F.T. (1991). Characterizing atmospheric effects on target contrast. *Optical Engineering*, Vol. 30, No. 10, pp. 1563-1575.

Whitehead, V.S. (1992). A summary of observations performed and preliminary findings in the space shuttle polarization experiment. *Proceedings SPIE*, Vol. 1747, pp. 104-108.

This page intentionally left blank

CHAPTER 14
IMAGE MODELING

In this final chapter, we look at the process of generating synthetic images. This subject is of interest to remote sensing in its own right but is particularly interesting from the image chain perspective. The process of effectively modeling nature to generate synthetic images that mimic real images requires a detailed knowledge and representation of the image chain. In this chapter, we will look at methods for generating synthetic images and how in many ways they represent a computer simulation of actual imaging systems. From this image chain perspective, synthetic image generation (SIG) models become a powerful tool for the study of imaging systems and an aid in system and image analysis.

14.1 SIMULATION ISSUES

There are a host of reasons why synthetic image generation is rapidly becoming a powerful tool in remote sensing. The demand for this tool has come from many different sectors: sensor designers, system operators, algorithm developers, image analysts, and image chain experts.

Sensor designers have long looked to this approach to evaluate tradeoffs between the types of image fidelity parameters discussed in Section 13.4. Synthetic images can be produced over a range of spatial, spectral, and radiometric performance specifications that new sensors might produce. These images can then be evaluated in terms of application-specific performance metrics to determine the utility of the sensor in a given application. These performance metrics may range from visual assessment through performance using fully automated or manually

assisted algorithms (cf. Sec. 13.5). This can also include sensor field of view and view angle studies to determine what range of look angles can be used to satisfactorily answer questions and to perform tradeoffs between synoptic perspective, resolution, and image acquisition costs.

The second group interested in synthetic image generation (SIG) is system operators. Here the task is to simulate a specific sensor and see how it can be used most effectively to image a phenomenon of interest. For example, with a system that can image at different days of the year, times of the day, and look angles, what combination is optimal for observing the phenomenon (e.g., crop stress), what combination is tolerable, and when will this sensor fail? Having delimited these acquisition variables, we could also ask when is the next acceptable acquisition and what are we likely to see in the image (i.e., is it worth the cost and effort to acquire the image?)?

The third group with a great deal of interest in synthetic images is algorithm developers. Here the interest is in developing and testing algorithms on scenes that contain the target of interest in a variety of forms and over a range of acquisition conditions. This is motivated in some cases by a lack of real imagery, because the algorithms are for a sensor under development or to help motivate the development of a new sensor. However, far more often it is to supplement data from an actual sensor where a robust data set does not exist. Another group of algorithm developers are looking to synthetic image modeling, not just for the images but to incorporate the models into algorithms to assist in machine-aided analysis. In many advanced algorithms, there is a process of hypothesizing about what is present at a location in an image or about the condition of something that is presumed to be at a location in an image (e.g., is there a ship located at this dock, is it a freighter or a war ship, is it under power?). The algorithm could have the model produce images of several types of ships at the dock from the view angle of the image, under the tidal conditions at the time of image acquisition, and then extract features from the synthetic image for comparison with features from the actual image. The features would then be used in the hypothesis testing to determine if a ship is present and, if so, what it is and what its status is. For clarity, we have simplified this process, which would often involve many tests of the feature extraction algorithm over slightly different models to develop a statistical representation of the features for comparison with the actual image features. In many cases, some or all of these model runs are performed in advance to improve algorithm performance times.

This concept of hypothesis testing leads us to the fourth group of individuals interested in synthetic image generation. These are image analysts, whose interest is also in hypothesis testing. An analyst might speculate about what he or she believes is in an image. By simulating what an image would look like based on that speculation and comparing the synthetic to the actual image, the analyst can often accept or reject the hypothesis. Alternatively, the analyst can use synthetic images to determine under what imaging conditions a hypothesis could be more definitively addressed (i.e., I can't tell if the crop in question is stressed, but if the image were acquired at 10 a.m. looking at 25° to 35° to the west, the probability of isolating the stressed vegetation will be much greater both in terms of spectral signatures and texture metrics).

The fifth group interested in synthetic imaging wants to use the images and the computer modeling processes associated with SIG for training purposes. Under the training umbrella, there are at least two distinct groups. One group wants to use the images as a backdrop or context for training, where the images provide an environment but are not of fundamental interest in and of themselves. The use of SIG in flight simulators would be a good example of this category. The image fidelity required for these applications can range from very crude terrain structure for general purpose training to very location-specific detailed scenes where a landing in a particularly difficult location under adverse conditions is of interest. The other group interested in training are those individuals charged with educating image analysts or applications specialists about what information is contained in images and how various conditions impact the appearance and information content of an image. For this group, the ability of the SIG approach to reveal the impact of various phenomena on the final image is often as important as the image itself, since understanding these phenomena and their interplay in image formation is critical for effective image analysis.

This brings us to the final group interested in SIG and, in the context of this volume, the group that would use the SIG process to study the image chain. These are the systems engineers who, either in designing a new system or analyzing an existing system's potentials and limitations for a particular task, are interested in using the SIG approach to combine the many complex interactions that go into the final image formation. If these interactions are well modeled by the SIG process, then the weak links in the image chain can be identified, including those links where the weakness is the result of some combination of interactions that might be difficult to isolate without an end-to-end simulation of the entire image formation process. The SIG approach also allows for the testing of alternative designs or improvements to the system before such changes are actually implemented.

These various interest groups place a broad range of demands on the SIG process. There are differing levels of fidelity, speed, interaction, etc., required for different groups and even the definition of fidelity is often different depending on the application. For our purposes here, we are interested in approaches that can be an aid to remote sensing image and systems analysts. This would include algorithm developers. This pushes us toward an interest in higher fidelity models with a reduced concern about the types of real-time or near real-time issues that flight simulation might require. Furthermore, from the standpoint of the underlying science of remote sensing and image analysis, we are interested in models that are based on the physical principles we have discussed throughout this book. SIG models based on these principles can help us assess the extent to which we understand the imaging process and point to where our understanding is flawed or oversimplified. If we simulate an image and compare it to an actual image in a controlled experiment, the discrepancies between the images can point us to limitations in our understanding of the imaging process. They may also point us to limitations in the implementation of the process in the SIG model, and care must be taken to differentiate between these two cases (i.e., when is the model inadequate and when is our understanding inadequate?). From this standpoint, the SIG process becomes a very powerful tool for the remote sensing scientist in pointing out where our understanding is inad-

equate and where more research and development effort is warranted. It helps us ascertain whether simplifying assumptions that were made are acceptable, or if we have cut too many corners. It can also point out the importance of a parameter or interaction mechanism we had completely overlooked. Conversely, a mechanism we could have spent a great deal of time trying to understand and include in our image analysis may be shown to be of little importance.

14.2 APPROACHES TO SIMULATION

In this section, we will briefly review some of the approaches to SIG and some of the methods that have been employed to simulate various phenomena that impact final image formation. In Section 14.3, we will review the basic elements of a SIG model and present some examples of how SIG can be used to analyze and understand certain phenomena.

14.2.1 Physical Models

One approach to SIG is to start with three-dimensional scale models of the scene of interest. These physical models can include terrain, land cover, structures, vehicles, etc., with the degree of detail a function of the resolution of the sensor to be simulated. This approach is described by Francis et al. (1993) for simulations in the reflective region. The model builder in this case includes all the texture and structure into the model that the eye would discern. The scene is illuminated with a collimated beam to simulate the sun and many diffuse sources to simulate the sky (cf. Fig. 14.1). The sunlight-to-skylight ratio and solar location are controlled to simulate the desired conditions. Path radiance is added either through the use of a beam splitter in front of the sensor or in postprocessing for digital simulations. The sensors are located above the target model and the optics adjusted to simulate the desired field of view. The resolution of the imaging system in the simulation facility is usually constrained to be significantly better than the system to be simulated. The PSF of the system to be simulated can then be cascaded with the digital image acquired by the test system. The image in this case is designed to represent the radiance field at the sensor. This approach is attractive in that many of the target–background interactions (e.g., shape factor effects) and within-target texture are taken care of by the physical model. It is also easy to change camera and sun angles for multiple images of the same scene. It has the disadvantage that a skilled model builder is required to construct and shade the targets such that the right reflectance and reflectance variations are included in the scene. This problem becomes acute when this approach is used to simulate multispectral scenes. This can be done to some extent by color balancing the sunlight and skylight sources and carefully selecting paints for the scene elements. However, the process quickly becomes intractable over wide wavelength ranges. Maver and Scarff (1993) describe a hybrid approach to simulate multispectral scenes where the physical models and illumination system are used to generate templates that are then processed using computer models. For example, by just illuminating the scenes with the

skylight lamps, an image can be obtained that can be segmented into material types (this can be simplified by the selection of paint shades in the model construction). The variations within a class can be used to generate maps of skylight irradiance levels, including shape factor effects. By using just the sunlight lamp, an image can be obtained that maps just the solar irradiance levels on each target, including slope-azimuth effects and shadows. The material maps can then be used to select reflectance spectra from a spectral reflectance database and the scene radiometry reconstructed using radiation propagation models, such as MODTRAN. The latter stages of this process resemble the steps used in the fully synthetic process discussed in the next subsection (14.2.2) and will not be treated here. The advantage of this approach is that some of the spatial variation can still be included by the model builder, and certain material interactions are also automatically included [e.g., shape factor (F)]. The disadvantage is that the scene segmentation of com-

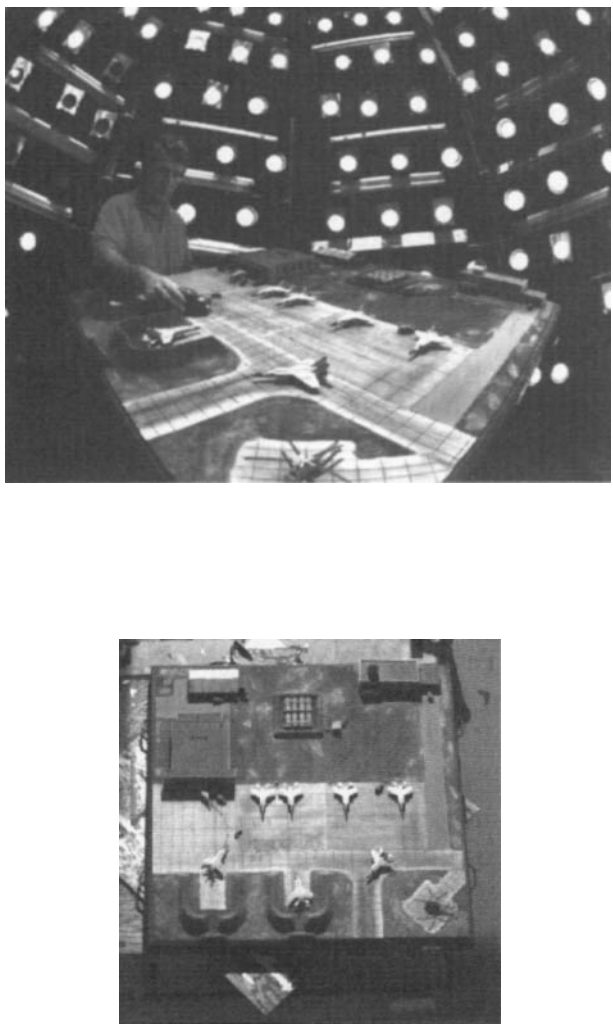


Figure 14.1 Simulation facility and synthetic image produced using physical models.

plex scenes can become difficult, requiring considerable manual editing. Also, specular effects (BRDF) are very difficult to simulate. Model elements need to be quite diffuse in order to avoid specular glints from one of the many skylight illumination sources. If a more truly diffuse sky were produced, there would still be the problem of having the model builder properly building BRDF effects into the model. Finally, this method cannot effectively simulate thermal phenomena, so alternative methods are required at longer wavelengths. While the physical-model approach has significant drawbacks, it has a great advantage when within-class spatial detail is important to approximate the visual appearance of an image. This can be seen in Figure 14.1, where a test chamber and model board are shown along with a synthetic image produced by this approach.

14.2.2 Fully Computerized Models

An alternative to the use of physical models and the method that is almost universally used for SIG that includes the thermal infrared spectral region, or imaging spectroscopy, is a fully synthetic approach. The scene elements, radiation propagation, and sensor effects are simulated using computer models. This approach is attractive and widely used because it allows for essentially endless variation in the scene and interaction processes. On the other hand, the computational complexities of this approach in terms of the scientific issues, the coding, and the run time are disadvantages. Because we are interested in a method that covers the EM spectrum from 0.4 to 20 μm, and where all of the interaction mechanisms can be addressed independently, we will concentrate on the fully synthetic approach. Many of these SIG models rely heavily on what is referred to as a first principles approach. To the extent practical, they attempt to accurately model the physical processes taking place in the image chain. The net result is a high-fidelity model that can provide insight into the process of image formation, as well as the synthetic image itself.

The fully synthetic SIG models are characterized by at least the following components, which are treated in more detail in the next section. They all have some means of creating and materially attributing objects and land cover, which we will refer to as the object data, and of assembling these data into a three-dimensional representation of the world, which we will call the synthetic scene. Both the object database and the synthetic scene are usually created using some form of computer-aided design (CAD) software. The objects may be facetized surface models, or they may be comprised of combinations of solids. The solids are described by mathematical functions combined using Boolean logic [this process is referred to as *constructive solid geometry, (CSG)*; cf. Foley et al. (1990)]. In some cases, the objects will have internal structure, and the surface and internal structural elements will be connected according to thermodynamic linkages. In all cases, every element making up the scene must have associated with it a set of material properties that we will refer to as the material database. This database contains information on the optical and, where applicable, the thermodynamic properties of the materials.

The models operating in the thermal region also have to have knowledge of the temperature of each scene element. In some cases, these temperatures are as-

signed externally; however, in the more comprehensive models, the temperatures are computed by the SIG model itself. In this instance, a meteorological database must exist to assist in temperature calculations performed by a thermodynamic model capable of computing the surface temperature of all scene elements. At least three types of thermodynamic models are in use: slab models that compute the one-dimensional heat flow perpendicular to a surface and may include internal heat sources or sinks; two-dimensional models that include lateral conduction between adjacent surface facets; and three-dimensional thermal models that include thermal interactions of surface facets with each other and with internal structures. All three types of models may or may not contain radiational heat-exchange interactions with other elements in the scene, depending on their level of sophistication.

A radiometry model is used to generate an estimate of the radiance field toward the sensor. This radiometry model is often coupled with a radiation propagation model such as MODTRAN (cf. Sec. 7.3.3 and 7.4.3) to compute the level of irradiance onto the surface. The radiation propagation model also computes the effect of the atmosphere along the target sensor path to yield the effective radiance field at the sensor. The atmospheric propagation models will often have a database of atmospheric conditions that are required as input. Finally, a sensor model must be available to characterize the sensor location, view geometry, field of view, resolution, and spectral and radiometric response.

Figure 14.2 shows a block diagram of the flow and interactions associated with a generic SIG model. Any given SIG model might not be divided along the exact lines shown in Figure 14.2 and may contain more or fewer interactions than shown. Because of varying applications, computational approaches, and simplification schemes, the various SIG models include a rather diverse set of approaches. A detailed treatment of this diversity is far beyond the scope of this text. Rather, in the next subsection we will look in more depth at one particular model to see how the image chain is modeled within the SIG process. However, before beginning this more in-depth treatment, we will examine some alternative models and highlight some particularly interesting features of the models. The most advanced of these models are under active development, so we will not explore any detailed information, since changes are occurring too rapidly for it to be relevant.

Many of the models use rather primitive background models to simplify calculations. Kornfeld and Penn (1993), on the other hand, discuss the use of extremely detailed background objects (trees in particular) for use in the modeling done at the Army Research Lab (cf. Fig. 14.3). This work, which emphasizes ground vehicle combat scenes, is also characterized by detailed modeling of the impact of specific sensor distortion effects on FLIR imagers. Similarly, many SIG models use simplified approximations to many of the thermodynamic computations. In part, this avoids the complexity of rigorous computations, but perhaps more important, it avoids the difficulties associated with laying out and assembling the detailed CAD figures needed to support higher order thermodynamic calculations. However, Sanders et al. (2000) describe the *Multi-service Electro-optical Signature* (*MuSES*) code designed to support very extensive thermodynamic models of targets, including detailed internal thermodynamic interactions. This level of detail is emphasized because one major application is to study the thermal signatures

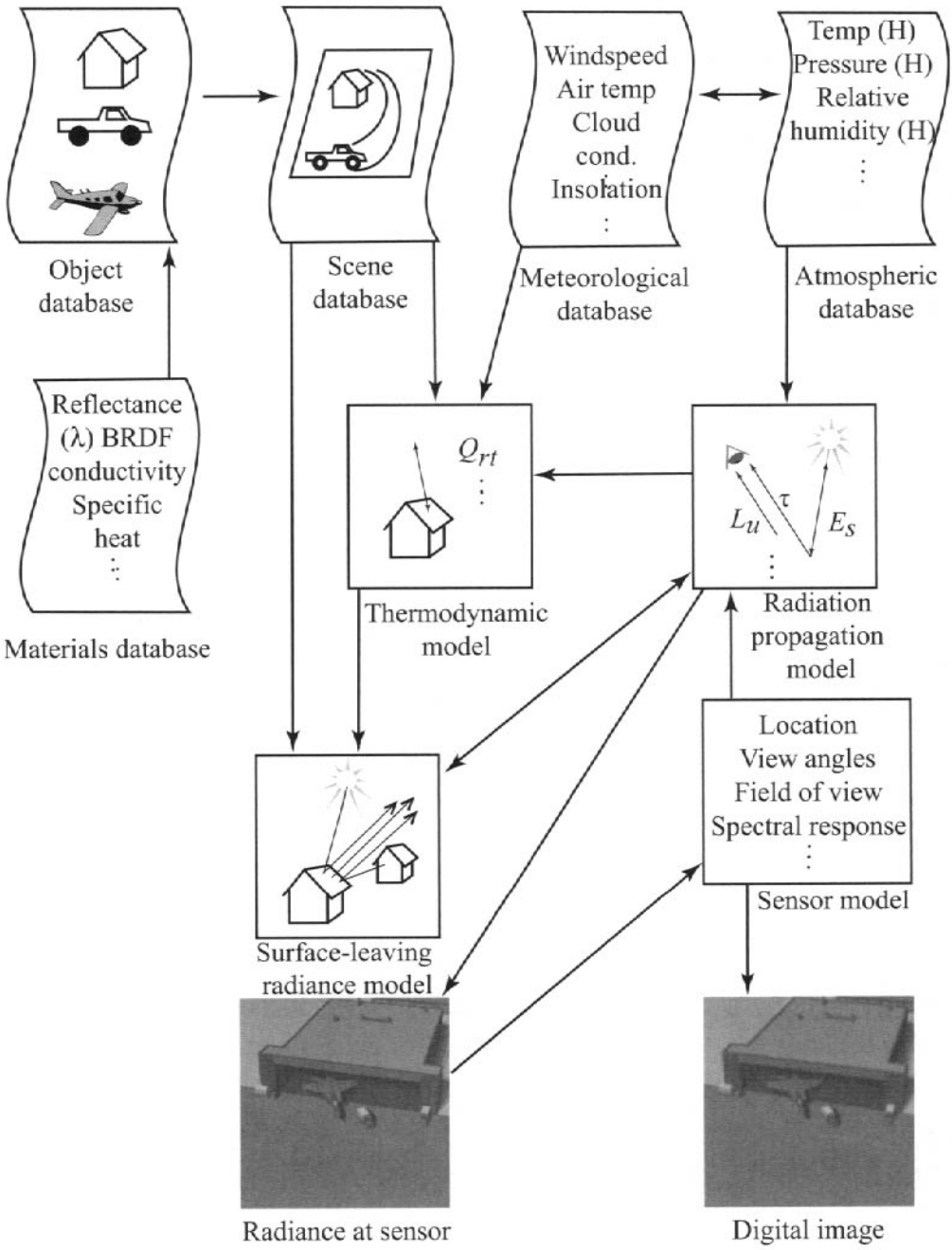


Figure 14.2 Diagram of conceptual data flow and interaction mechanisms in a generic SIG model.

from ground vehicles imaged with very high-resolution sensors as described by Johnson et al. (1993), (cf. Fig. 14.4). Stewart et al. describe another interesting approach to the thermodynamic modeling problem used in the *simulated infrared image model* (*SIRIM*) code. This model divides the objects up into volume elements (voxels) and performs thermodynamic calculations on the interactions between voxels to provide a method to approximate three-dimensional heat flow within ob-

Figure 14.3 LWIR synthetic image of targets in forested area.

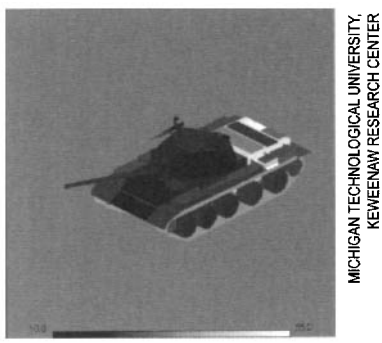


Figure 14.4 Synthetic LWIR image of a tank.

jects. These thermal calculations can then be coupled to background simulations (e.g., sea surface structure) to produce final image representations, including target and background interactions (cf. Fig. 14.5). Cathcart et al. (1993) and Sheffer et al. (1993) point out that background modeling can be as important as target modeling and discuss procedures that have been developed for use in the generation of land and sea backgrounds, respectively (cf. Fig. 14.6). Savage et al. (2006) describe the Irma model which is designed to support multiple sensing modalities (passive vis-LWIR, millimeter-wave, LADAR and SAR.) Moorhead et al. (2001) describe the CAMEO-SIM model that has been developed to support studies of camouflage, concealment, and deception methodologies and Curran and Curlee (2006) describe imbedding MuSES target models into CAMEO-SIM scenes. All of the SIG models have particular strengths and weaknesses, and the user is cautioned to investigate their capabilities carefully to determine which approach is best suited to the user's particular requirements.

Figure 14.5 Synthetic LWIR image of a ship on modeled ocean.

Figure 14.6 Synthetic LWIR image showing a helicopter over simulated water.

14.3 A MODELING EXAMPLE

In this section, we will look in more detail at the components of a SIG model. In particular, since the SIG process is essentially an effort to produce a computer model of the image chain, we want to look at how SIG can be used to break out elements along the image chain that would help us to understand and analyze images. To do this, we will use one SIG model as a point of reference to see in a common framework how the image chain can be modeled. As discussed in the previous section, there are significant variations in the approach to SIG from one model to another. In addition, since even the model we will use as a baseline is under constant upgrade, we will tend to avoid the details that are dynamic and emphasize the conceptual issues. The model we will use as a baseline is the *Digital Imaging and Remote Sensing Image Generation* (*DIRSIG*) model described by Schott et al. (1992, 1999). It was obviously chosen in part because of the author's familiarity with the model, but more important, because it was developed along the image chain principles described throughout this treatment. Furthermore, it uses the same governing radiometry developed in Chapter 4, which will simplify the presentation.

The potential user of a modeling tool should recognize that some of the phenomenology modeled in the examples presented here may not be handled in other models, or it may be treated differently. As shown in the figures in this section, most modern scene simulation tools generate reasonably attractive imagery from a first-look visual standpoint. It is incumbent on the user to ensure that all the phenomenology needed for a specific application is adequately treated in the model chosen.

In order to understand how the SIG process can help us to understand the image chain and, conversely, how the image chain approach can be used in generating SIG images, we will walk through the components of the SIG process illustrated in Figure 14.2. We will look in more detail at how each step represents links and interaction mechanisms along the image chain and at how these links are implemented in at least one SIG model.

The first step in the SIG process is to create a scene. In order to do this, the ground coverage of the sensor and the flight path of interest must be determined (i.e., are we looking at a single frame, sequential frames along a track, an approach sequence, a strip beneath a push broom, etc.?). We also need to know the approximate resolution of the sensor in terms of GIFOV and/or the approximate level of subpixel image exploitation to be applied to the final image product. This helps determine the degree of detail needed in the CAD models. For example, to simulate a sensor with an 80-m GIFOV, it doesn't make sense to show the detailed structure on the roof of a building. On the other hand, to simulate an aerial system with a 0.1-m GIFOV, this level of detail may be important. The level of detail will also often be a function of the particular application. For example, if we are interested in studying stress in a particular crop, we might have very detailed models of the structure of that crop and yet include only the most primitive structure in adjacent crops. (Note that, depending on the approach taken and the resolution, this could be a 3-D spatial model of the structure with optical properties of each element or a detailed BRDF model of the canopy treated as a unit.)

The objects in a scene are created using CAD models such as shown in Figure 14.7. DIRSIG currently uses facetized surface models that may contain one or more internal heat sources (sinks) for use in simulations where self-emitted flux is relevant. Each facet is assigned a material ID, which is associated with the material database. The terrain can also be created using the CAD approach. However, if digital elevation models (DEMs) are available, they can be converted into facetized surface models in the same format as the object data (the DEM data can also be interpolated to finer sample centers and facetized structure added if the data are too course for the required resolution). In addition, a level of surface variation can be added using a bump mapping approach that will deviate the local facet normal by an amount proportional to the gradient in a gray-scale bump map projected onto the surface. Material types can then be assigned to the facetized terrain using the CAD process. For large regions, this can be a very tedious process. An alternative approach that has been implemented involves using image data that is geometrically registered to the DEM. If this image data can be classified by land cover or material type, the material ID associated with that land cover can be automatically assigned to the corresponding location on the facets derived from the DEM. In

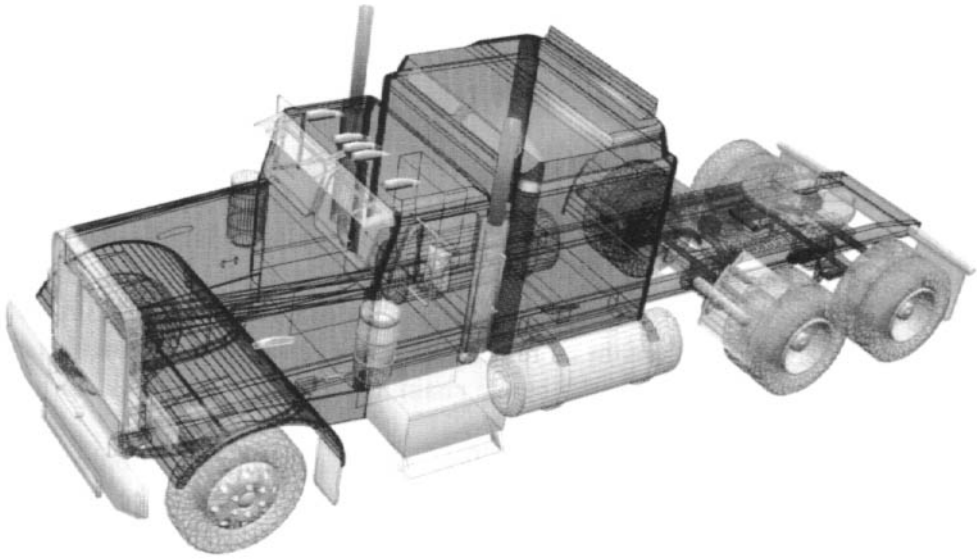


Figure 14.7 Wire frame of an object used in the SIG process. The object is produced using CAD software, and material types are assigned to each facet during the construction process.

either case, the final scene is produced by using the CAD process to place objects where required in the scene, and a geographic location is assigned to the scene coordinates. Specialized objects or objects with particular features are also included at this point using the CAD process and control functions. For example, objects that move as a function of time (clouds, vehicles) are located at various points in the scene and the time sequencing assigned (i.e., when is the object where?) as shown in Figure 14.8.

Very closely linked to the scene database, whose construction we just described, is the material database. Each material in the scene is a pointer to this database. The database contains the optical and thermodynamic properties of each material type. The optical properties consist primarily of reflectivity (emissivity) as a function of wavelength and view angle (from the target normal) over the wavelength range of interest or spectral BRDF data where appropriate (note that if polarimetric phenomena are being modeled, the BRDF data are in the form of models or lookup tables that allow computation of spectral Mueller matrix representations of the BRDF). For translucent objects, such as clouds or tree leaves, this will also include the spectral extinction coefficient. The thermodynamic parameters include the solar absorption coefficient, specific heat, conductivity, thickness, and broad-band emissivity and may include information concerning the magnitude, operating times, and operating levels of internal heat sources (sinks).

In order to access the scene data, the DIRSIG model turns to the sensor model. The sensor model includes, as a function of time, the location of the sensor in three-dimensional space, the orientation of the primary optical axis, the sensor field of view, and sampling information. For example, for a simple framing sen-

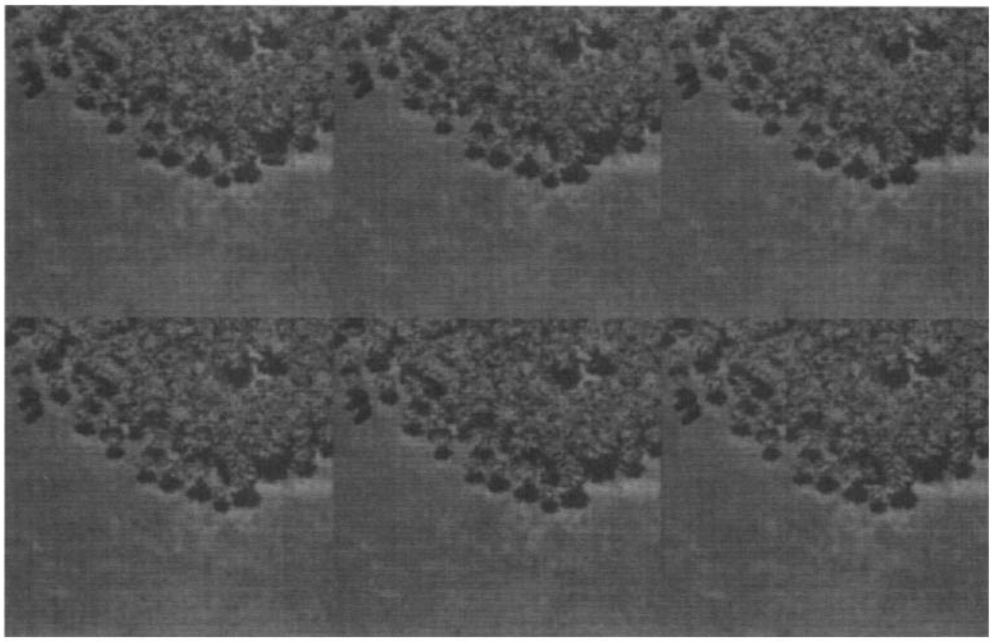


Figure 14.8 DIRSIG false color infrared image sequence showing a vehicle passing under overhanging trees. These images could be used to predict the vehicle signature under a range of adjacency and obscuration conditions. See color plate 14.8.

sor, the sensor model would include the FOV with respect to the sensor's x and y directions, the number of pixels in x and y, and the point spread function associated with the sensor or with each pixel if available. The sensor model also includes the number of spectral bands and the spectral response function of each band. In order to produce a final digital image, we want to first generate a representation of the image that contains spatial, spectral, and radiometric resolution, equal to or greater than the final image so that any degradation effects can be properly treated. This is accomplished in DIRSIG by producing a radiance image that represents the radiance in each band (oversampled spectral radiance weighted by the sensor response) oversampled for each pixel at higher spatial resolution than the sensor and before any sensor-induced radiometric noise or quantization. To generate the required spatial resolution in the radiance image, the oversampled data for the pixel are convolved with the point spread function to yield the radiance sampled for the pixel.

A ray tracer is used to access the scene data and integrate the entire scene generation process. To simplify our treatment, we will assume we are imaging with a pinhole camera and that for each "pixel" in the radiance image we trace a ray from the focal point of the optical system through the pixel center and into the scene, as illustrated in Figure 14.9. The ray tracer's function is to trace backward along the photon paths to determine what interactions took place that would affect the flux on the sensor at the point of interest. The ray tracer locates the point on a facet where the ray hits the scene and begins to gather data. First, the material database is accessed to determine the material type, as well as the slope and aspect of the material. If the material is translucent (e.g., a cloud), the ray trace is continued until an opaque object or the sky is hit. For translucent objects, the distance

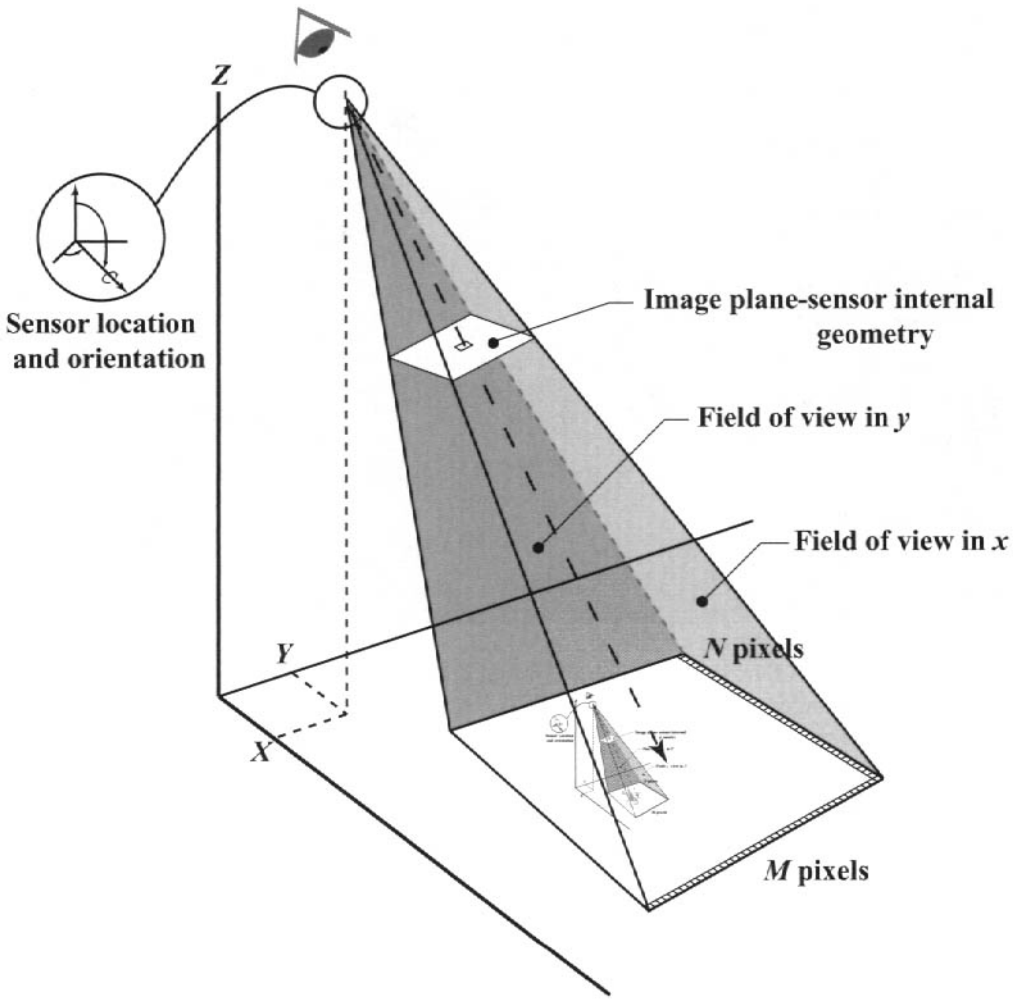


Figure 14.9 Illustration of the ray-tracing process for a simple framing camera. To generate an $N \times M$ radiance array, rays are traced from the focal point through each pixel center in an $N \times M$ image plane. Note the $N \times M$ array is denser than the final image array to allow convolution and resampling with the instrument PSF.

the ray travels through the object must be computed so that the extinction can be determined. The ray tracer also computes a number of geometric factors, such as the view angle (θ), the slant path to the sensor (for transmission and path radiance calculations), the angle from the surface normal to the sensor, and the direction of specular reflection. It then casts out rays to ascertain three additional sets of data. First, it casts a ray toward the sun to see if the target is in sun or shade (assuming it's a daytime image), and, if running in the thermal region, it casts rays to where the sun has been every k minutes for the previous N hours (N is usually 12 or more hours; k is usually 15 or fewer minutes) to generate a sun-shadow history for the target (cf. Fig. 14.10). Second, rays are cast into a hemisphere above the target plane to determine the shape factor and the radiation environment of the target surround (cf. Fig. 14.11). In cases where full BRDF data are available, these rays can

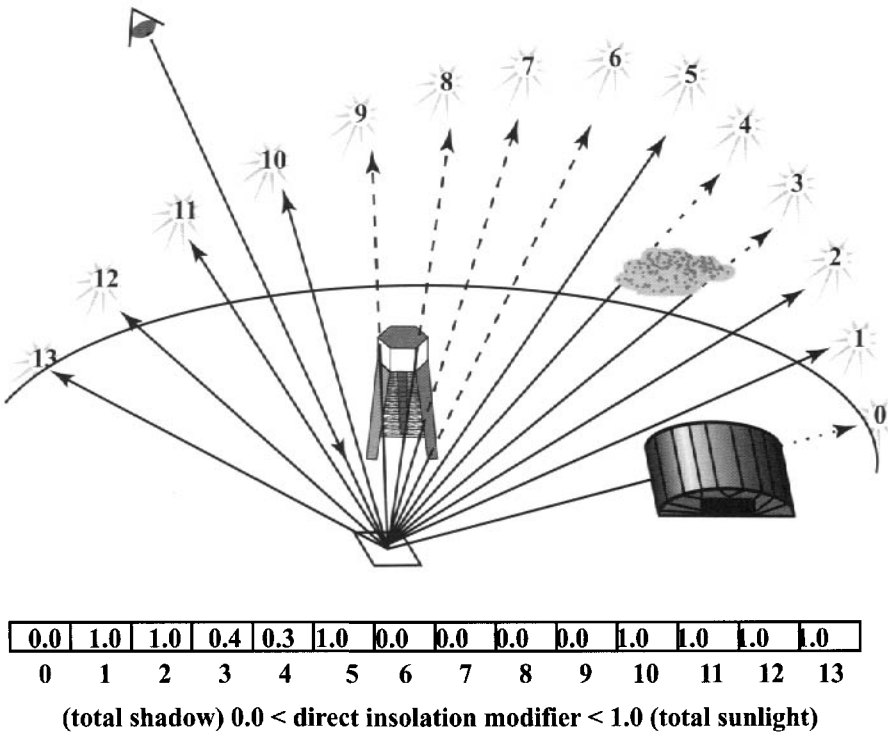


Figure 14.10 Sun shadow history, including partial obscuration by transmissive

gather the temperature, reflectance, and sun-shade condition of adjacent surfaces to determine directional radiance loads for BRDF calculations. Third, a ray is cast in the specular bounce direction to determine if an object (including cloud) is hit (if so, its temperate, reflectance, and sun-shadow condition are determined), or if no object is hit, the ray is cast to the sky to determine where in the sky the specular ray came from. This is important because the sky is not isotropic, and the variations in the sky can greatly impact the appearance of specular objects.

The data the ray tracer has gathered is sent to the thermal model if self-emitted photons are relevant in the bands being simulated. The temperature of most objects is a function of the object's thermal inertia and environmental driving factors. The SIG model acquires data about thermal inertia from the materials database (specific heat, conductivity, density, thickness), but a meteorological history must also be available. This history consists of direct and diffuse insolation levels, air temperature, relative humidity, cloud type and coverage, precipitation rates (and temperature), air pressure, and wind speed as a function of time. This history may be available from experimental data or may be predicted based on standard meteorological forecast data along with geographic data and a forecast date. The insolation levels are adjusted based on the sun-shadow history information, and the radiational exchange is modified by the object's solar absorptivity, the shape factor term, and the temperature of adjacent objects. The temperature is calculated by a thermal model that uses all the available data (including information on internal heat sources) to solve a forward chaining differential model. DIRSIG uses a modi-

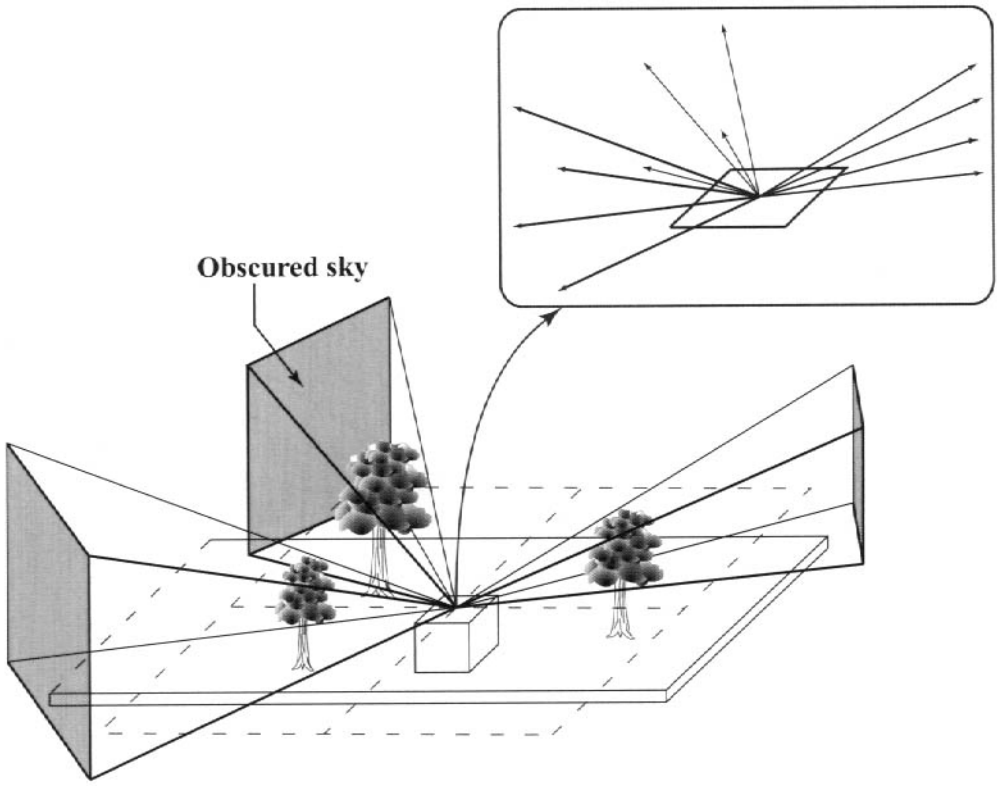


Figure 14.11 Rays are cast into the hemisphere above the plane of the target pixel to compute the shape factor for use in the thermodynamic modeling of radiational exchange and in the radiometric image formation model.

fied version of the THERM thermal model [DCS Corporation (1991)]. This type of model is very effective for passive objects; however, more exotic models may be used where internal heating effects become complex. For example, Figure 14.12 shows an example of a MuSES [cf. Sanders et al. (2000)] thermal model of a tank included in a DIRSIG scene where the background temperatures were computed by THERM. The target temperature is then added to the other data the ray tracer has collected, and the data are sent to the radiometry model.

The DIRSIG radiometry model essentially consists of the spectral version of the big equation of Chapter 4 [cf. Eq. (4.63)] without most of the simplifying assumptions other than the use of numerical approximations for most of the integrations. The radiation propagation terms E'_s, τ_1, τ_2, L_u, and L_d are all derived from MODTRAN [cf. Berk et al. (1989)] or FASCODE when very high spectral resolution is required (e.g., for LIDAR simulations) and vary as a function of wavelength, slant range, target elevation, view angle, and azimuthal angle. The atmospheric makeup is controlled by the standard user-supplied inputs to the radiosonde profiles in MODTRAN, with the lowest layers modified by the surface meteorology from the DIRSIG weather data files. The radiometry model solves for the spectral radiance reaching the sensor and then computes the effective radiance in the band using the spectral responsivity for each of the sensor's spectral bands. At this point, the radiance values are inserted into the radiance image arrays for each band.

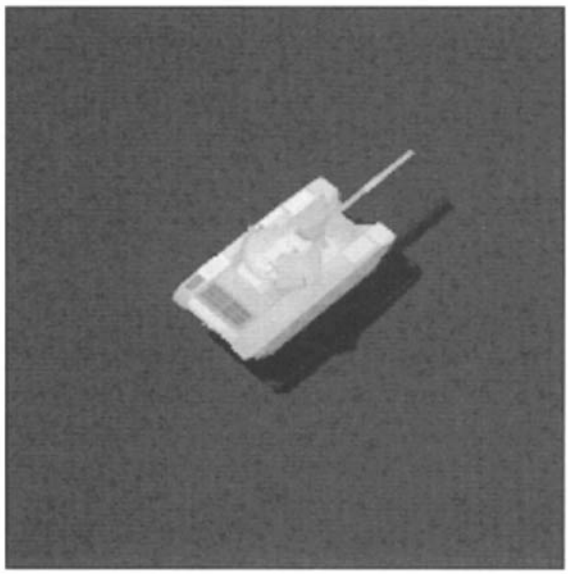


Figure 14.12 DIRSIG rendering of an image where the vehicle temperatures were computed by the MuSES model and the background temperatures were computed by THERM. Vehicle model courtesy of Signature Research.

In addition, essentially all of the data that went into the computation of the final radiance can also be inserted into an image array for training, diagnostic, or advanced algorithm development purposes. From our perspective, we can think of these diagnostic images as a history of the radiometric image chain for each pixel in the radiance image. For example, the within-band transmission image consists of the τ_{2avg} value for each pixel in the radiance image. The radiance image can be thought of as the result of sampling the radiance field reaching the sensor with an array of delta functions located at the radiance image sample centers. The radiance field must then be convolved with the PSF of the sensor system to generate the effective radiance values associated with the final image pixels. This is accomplished using the procedures discussed in Chapter 13 and illustrated in Figure 14.13. Note that DIRSIG supports modeling shift variant systems where the PSF may vary depending on the location of each detector element. The radiance values are then converted to electronic values according to the response function of the system being simulated. Both random and structured noise, as discussed in Section 13.2, can be added to the signal at this point. Finally, the image is quantized with a quantizer that simulates the quantization levels in the sensor. Note that for simplicity we have neglected sensor platform motion effects, treatment of more exotic sensors, and most of the details of the SIG process. The reader is referred to the referenced literature and the model developers for more detailed information.

A radiance field image is shown in Figure 14.14 along with some of the diagnostic images associated with its image chain. In Figure 14.15, a radiance image is shown along with some of the images associated with the latter stages of the image chain.

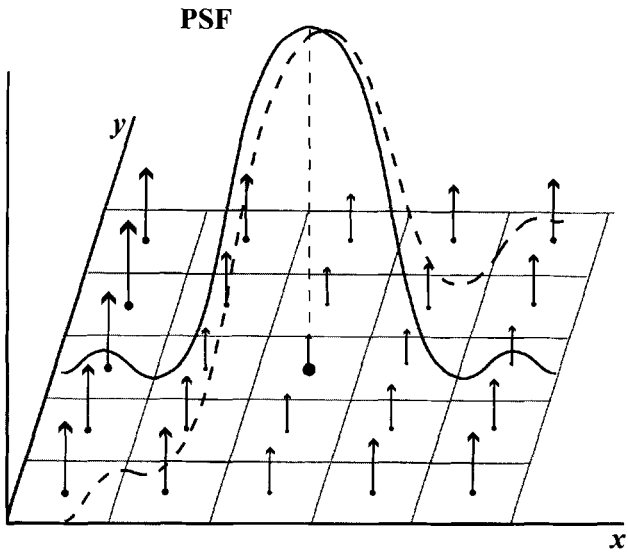


Figure 14.13 The final synthetic image is obtained by convolving the oversampled radiance field (indicated by the heights of the vectors in the figure) with the point spread function (PSF) of the system to calculate the radiance of each pixel.

The SIG images that result from the process described thus far are generally too flat because of the lack of naturally varying texture in many material classes. For example, the model builder will often not change the density or structure of grass or include wear marks in pavement unless these are considered critical elements (e.g., crop structure might be built into the canopy of a particular crop under study). It is, of course, possible to build models to the level of detail where structure is included automatically in the SIG process. However, the level of effort involved in the building of the models and the negative impact on run time are often not justified. An alternative procedure incorporates spatial structure into the SIG process using one of several methods of texture generation. This procedure leaves the mean-level radiance values in a region unchanged and introduces variations about the mean by spatially varying the reflectance as a function of material type. Schott et al. (1995) describe how this approach can be implemented to preserve the spatial and spectral correlation within and between spectral bands for generation of multispectral SIG images. Improved methods also available as options in DIRSIG, including subpixel mixture modeling to achieve texture, are described by Scanlan et al. (2004). A limitation of some of these approaches is that the spatial structure of the land cover or material in one or more spectral band(s) must be known at the scale of the radiance field image. This structure data can be generated mathematically as described by Haruyama and Barsky (1984) and Pentland (1984) or extracted from actual imagery. The resultant SIG images contain a good approximation of the spatial, spectral, and radiometric characteristics of actual images [cf. Ientilucci and Brown (2003)]. Figure 14.16 shows an example of an image where texture has been included.

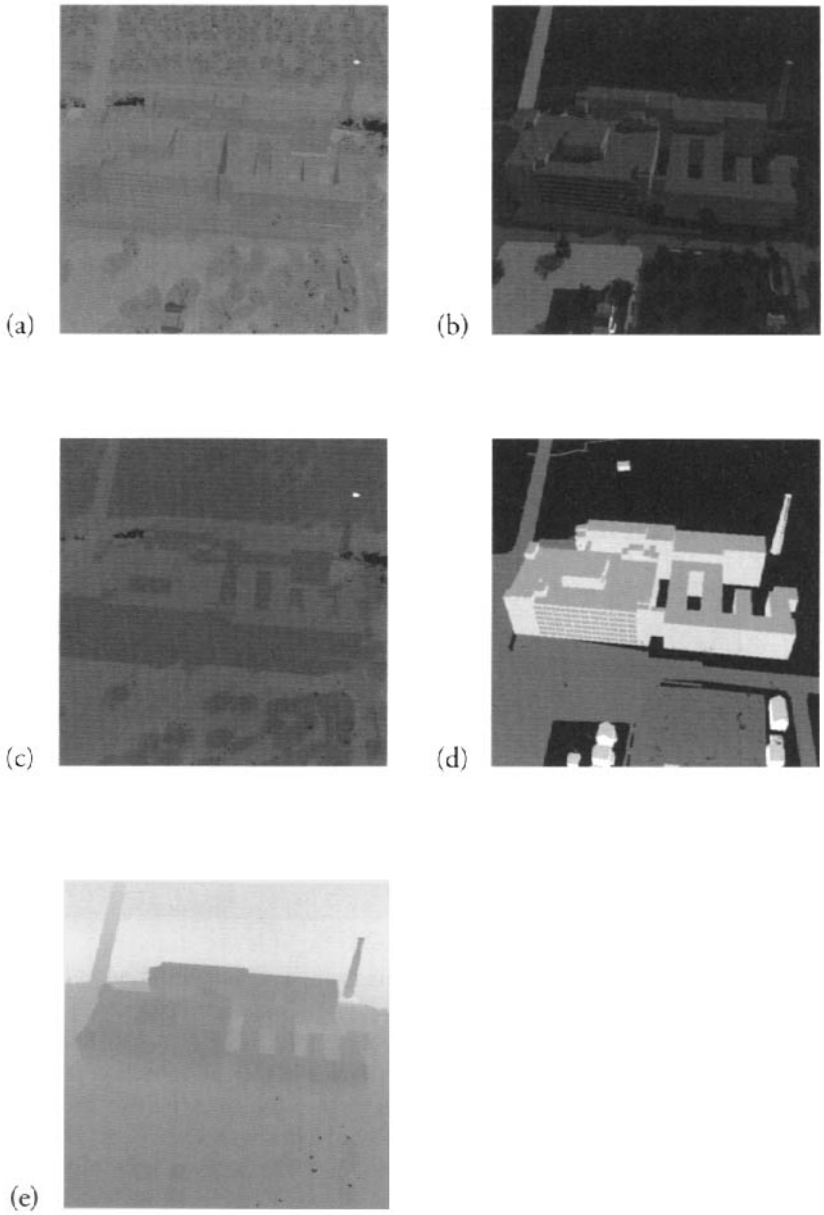


Figure 14.14 DIRSIG (a) LWIR and (b) visible radiance image and several truth images, (c) temperature, (d) material ID, and (e) path radiance.

14.4 APPLICATION OF SIG MODELS

In this subsection, we will examine just a few of the ways in which SIG can be used to help us better understand and analyze certain remote sensing phenomena. These examples are not intended to be comprehensive, as they merely scratch the surface of what SIG is currently capable of doing. Ongoing improvements to SIG will make it even more valuable in understanding and extracting information from images.

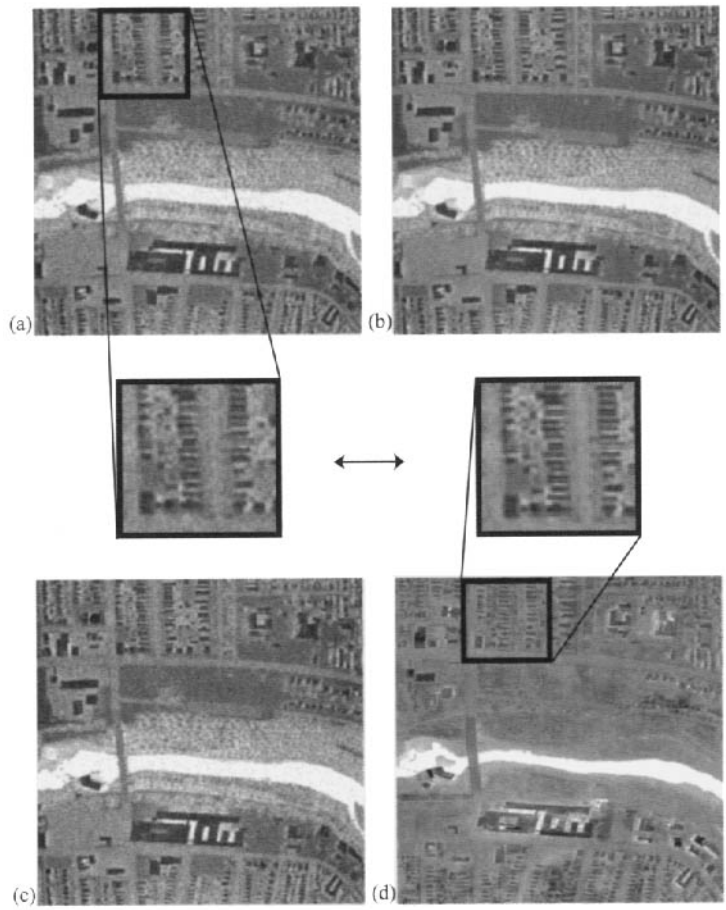


Figure 14.15 A final DIRSIG LWIR image (a) and images showing later stages of the image chain: (b) radiance image, (c) radiance image with noise effects, and (d) an actual LWIR image acquired with RIT's LWIR imaging system. Final DIRSIG image includes noise, MTF, and sampling effects. Note the synthetic scene has fully foliated trees in the gorge on either side of the river. The flight line for both real and synthetic sensors was from bottom to top, allowing the sensors to see warm side walls in these cold nightime simulations.

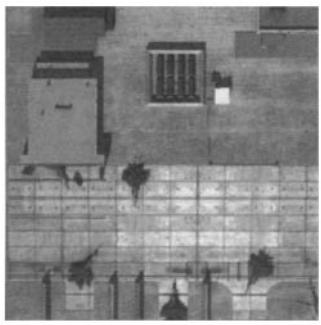


Figure 14.16 A true-color synthetic scene produced by DIRSIG showing texture effects. Note this is the same scene shown in the model board in Figure 14.1. See color plate 14.16.

Figure 14.17 shows an example of how radiometrically correct SIG can be used to characterize the impact of atmospheric effects on remotely sensed images. In this case, the scene was assumed to be a uniform, flat Lambertian reflector with a mean reflectance equal to the local albedo. The sensor modeled is a line scanner flying a north–south heading. Variations in the resultant image show how the atmosphere would be expected to influence a wide-FOV sensor. An actual image has variations due to both atmospheric and BDRF variations in scene elements. These combined effects are shown in the *advanced solid-state array spectroradiometer* (*ASAS*) images shown in Figure 14.18. The ASAS sensor is designed to study BRDF effects by taking images of the same spot at several angles ahead of and behind a sensor. The images show the dramatic variations in how the same site appears when viewed at 19 different geometries [cf. Ranson et al., (1994)]. Figure 14.19(a) shows how DIRSIG treats individual leaves as having both spectral reflectance and transmission and how the ray tracer can properly model the multiple interaction in a stacked canopy. Figure 14.19(b) shows an example of a wire frame model of an individual tree whose leaves are assigned optical properties (spectral transmission and reflectance) appropriate to the species and condition of the tree. Finally, 14.19(c) shows rendered DIRSIG scenes of a forest canopy imaged in a fashion similar to that used with ASAS. Using the "truth" images, the phenomenology causing the brightness variation in these images can be deconstructed into its various sources (e.g., atmosphere, illumination/shadowing, leaf BRDF, and canopy BRDF). The SIG models can be used to assist in understanding both the relative shape and magnitude of these phenomena. If sufficient atmospheric data are available, the SIG data can even be used in generating spatially dependent atmospheric correction values for use in radiometric image analysis (i.e., atmospheric bias and gain "images" could be made and applied to the actual images through subtraction and division). This is just one example of how SIG can be used both to understand phenomena and potentially to generate analytical corrections to assist in image analysis.

An application of the ability of SIG to model thermal effects as a function of time is shown in Figure 14.20(a)–(c). These LWIR images show an airfield from which aircraft have departed every 30 minutes for the past hour (two aircraft have departed). The differential temperature resulting from the length of time shadows have been cast on the high-thermal-inertia parking apron provides a potential signature to assess these departure (or arrival times). However, Figure 14.20(c) shows the sensitivity of this type of signature to environmental conditions. In this image, the same scene is modeled, but the wind speed, which was assumed to have been constant at 3 mph in the first scene, has now been raised to 10 mph (wind speed in both scenes was held at a constant level for the previous 12 hours). A comparison of the images shows how thermal signatures are a function of environmental variables. In this case, the increased wind speed tends to drive all object temperatures toward ambient air temperature and "scrubs" thermal contrast from the scene. For comparison, Figure 14.20(d) shows an actual image taken in the visible and a simultaneously acquired image in the LWIR [Fig. 14.20(f)] of a parking area from which an aircraft has recently departed. The complex interaction of optical and

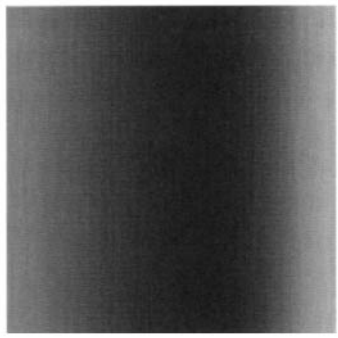


Figure 14.17 Contrast-enhanced blue spectral band DIRSIG line scanner image of a uniform Lambertian ground model. Plane is flying north (up) with the sun east (right). The brightness variations in the image are strictly due to atmospheric transmission and scattering effects (see Fig. 4.11).

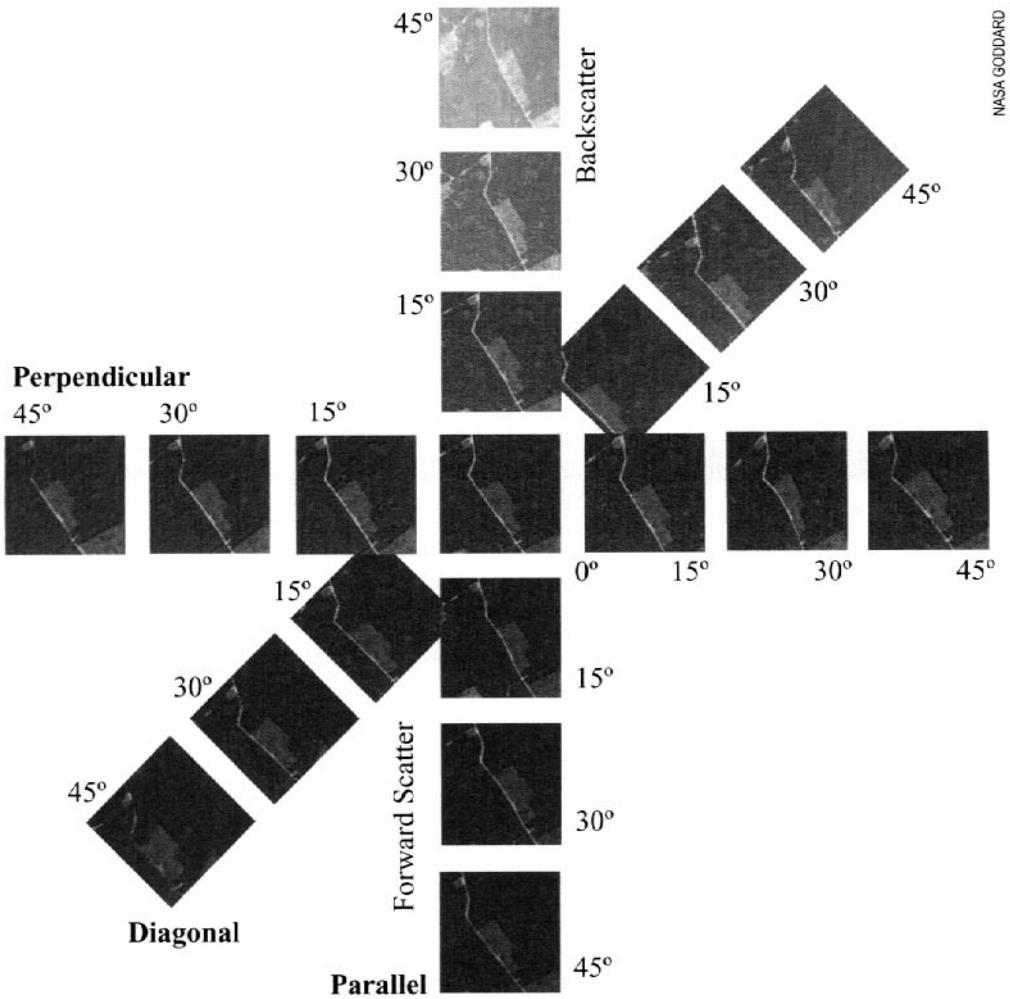


Figure 14.18 Color IR version of ASAS images from three flight lines showing combined effects of atmospheric and bidirectional reflectance variation. See color plate 14.18.

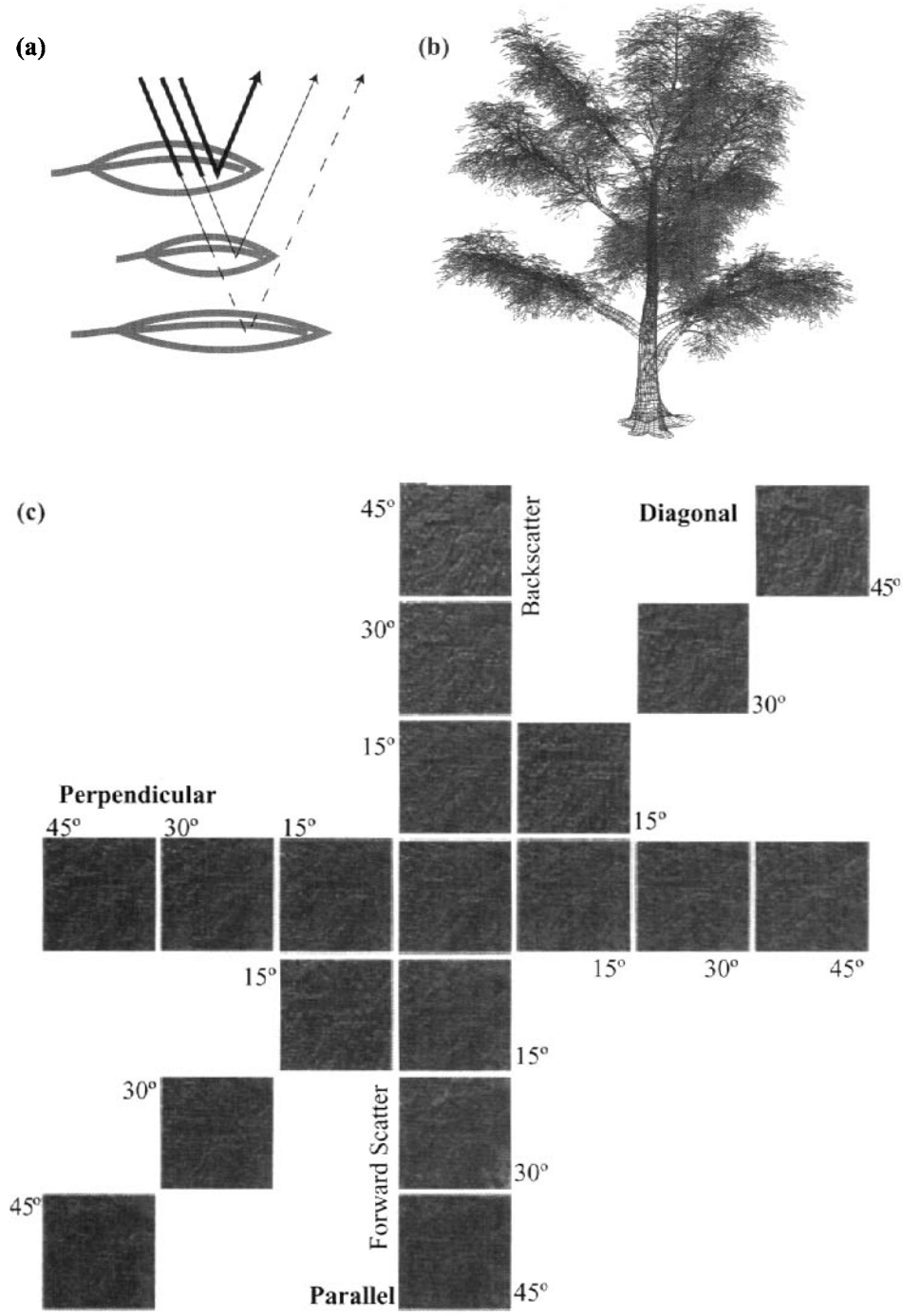


Figure 14.19 DIRSIG simulation of combined atmospheric and bidirectional canopy reflectance: (a) illustration of how the DIRSIG ray tracer includes reflectance and transmission of a stacked leaf canopy, (b) wire frame model of a tree used in the DIRSIG simulations (generated by TreePro), and (c) DIRSIG images of a canopy taken using acquisition geometry similar to that used with ASAS to acquire Figure 14.18. See color plate 14.19.

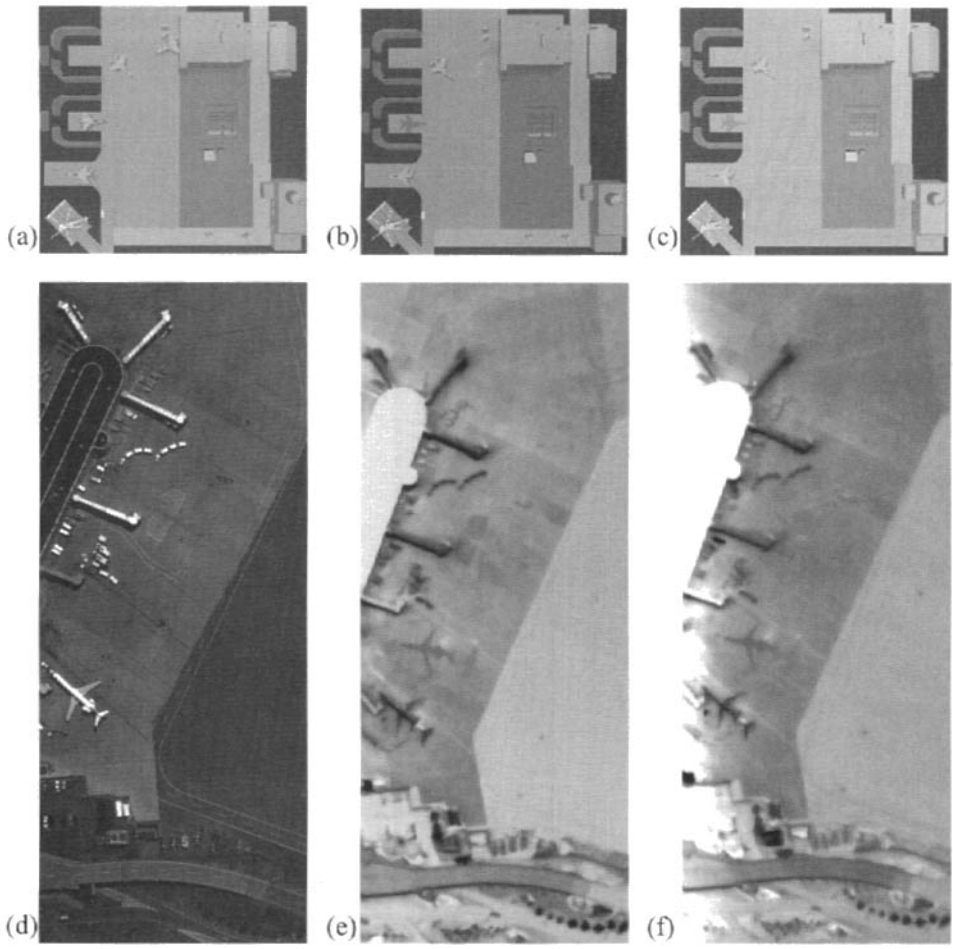
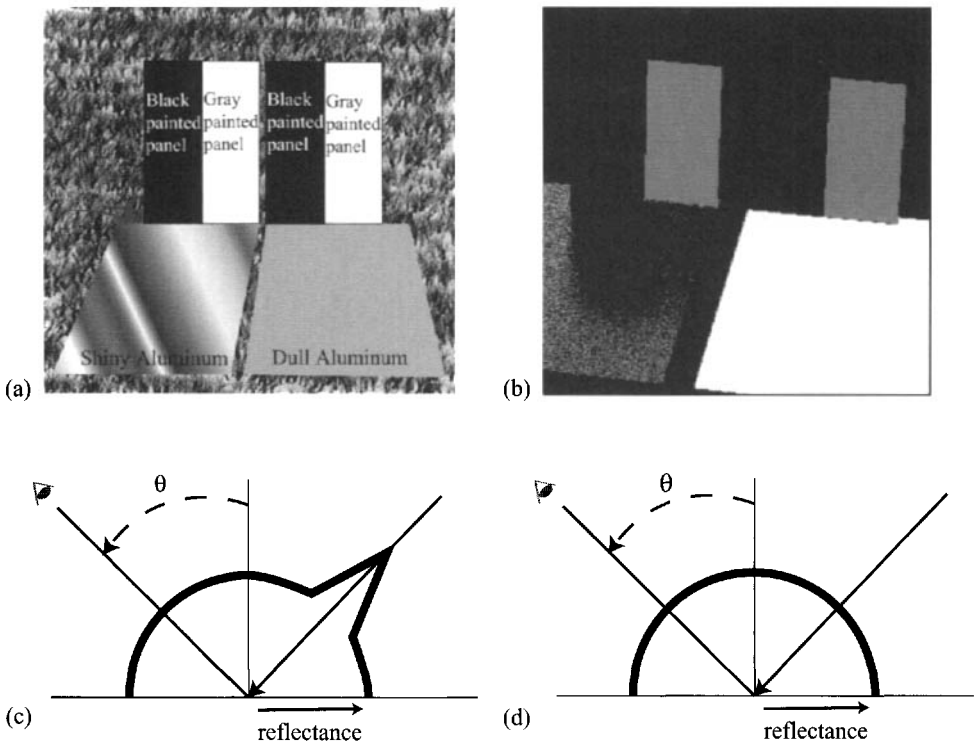


Figure 14.20 LWIR DIRSIG image showing airfield at different wind speeds and real images: (a) primitive DIRSIG image with wind at 3 mph before aircraft departures, (b) same image one hour later after two aircraft have departed, (c) same scene as (b) with wind speed at 10 mph, (d) WASP visible, (e) MWIR, and (f) LWIR image of airfield taken at same time as (d). Note the residual (shadow) from a departed aircraft.

thermal effects in MWIR and LWIR images makes detailed interpretation of the images very difficult without the aid of tools such as SIG.

The importance of understanding the directional reflectance properties of surfaces is shown in Figure 14.21. The two aluminum panels are modeled as having the same nominal reflectance spectrum, but in one case the surface is quite specular, and in the other it is quite diffuse. In the diffuse case, the sunlight is diffusely scattered and the radiance from the surround is equally weighted by the fraction of the hemisphere represented by each object, with the net result that any small object (i.e., small fraction of hemisphere) has a real but relatively limited impact on the observed radiance. In the specular case, the sunlight is forward scattered away from the sensor and the radiance from the specular direction is weighted quite



Cross section of specular BRDF function in polar coordinates for the view angle θ.

Cross section of diffuse BRDF function in polar coordinates for the view angle θ.

Figure 14.21 DIRSIG image (b) showing the impact of variations in the BRDF on the observed image. (a) A "cartoon" of the image simulated in (b), (c) a simplified illustration of a somewhat specular BRDF (similiar to the shiny surface), and (d) a Lambertian surface (see Fig. 3.2).

heavily, resulting in the same relatively small object having a more pronounced impact on the observed radiance.

The SIG modeling process can be used to study the impact of various sensor designs on imagery. One of the more fundamental differences is shown in Figure 14.22, where a scene is imaged as it would be seen by a framing camera and a line scanner. Note, in particular, how the frame camera image exhibits radial relief displacement, while the relief displacement in the scanner image is only in the across-track direction. Sensor platform effects are also important, since the scanner will not be stable over the period of image acquisition. This is shown in Figure 14.22(c), where aircraft platform motions, particularly roll, have been introduced. Careful analysis of these images would also show that the scanner images have geometric distortions due to angular sampling effects (i.e., tangent distortion and resolution degradation with angle). It is important to realize that these variations in the image characteristics are critical not only in design considerations. The performance of image processing and analysis algorithms will vary significantly with sensor characteristics, and SIG can play a crucial role in helping to develop and evaluate image analysis algorithms appropriate for different sensors or acquisition conditions. For example, Figure 14.23 shows an example of a synthetic scene con-

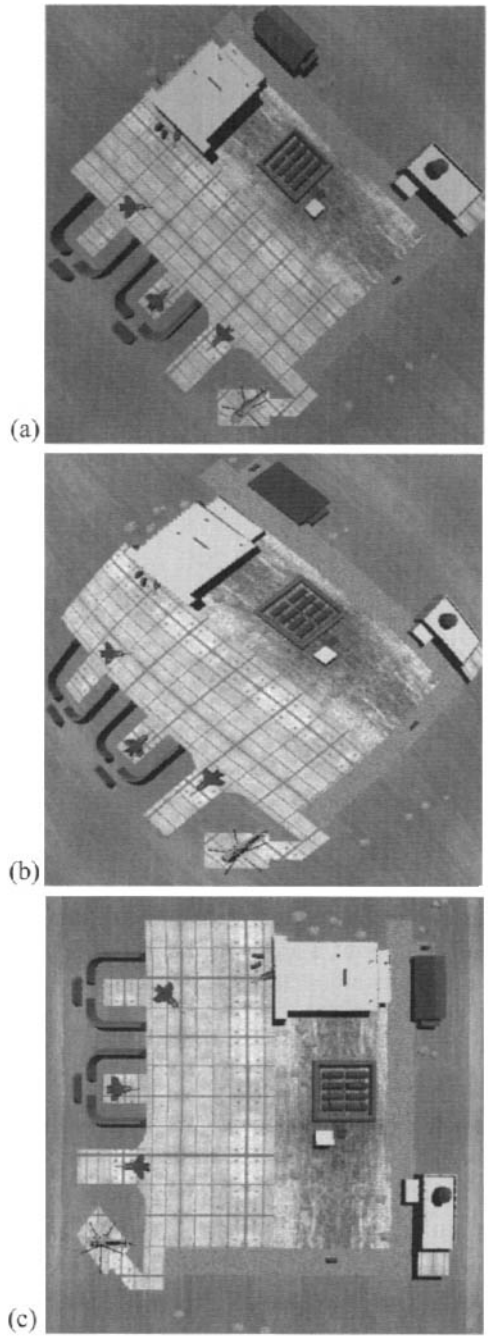


Figure 14.22 DIRSIG images showing sensor modeling: (a) frame camera, (b) line scanner model flown bottom to top including tangent errors, and (c) line scanner model flown bottom to top including low-frequency roll (note the subtle curvature in lines in the parking area) and line start jitter (this is most apparent in the parapet of the building at bottom right).

structed to evaluate hyperspectral algorithms for detection of concealed vehicles. The complexity of the illumination field on the concealed vehicle is shown in Figure 14.23, (e) and (f), which were generated by placing the real and "synthetic" camera under the camouflage net and photographing the Humvee. Note that the illumination field on the vehicle is caused by direct and diffuse illumination through holes in the net, as well as partial spectral transmission through the net material. The standoff image of the netted vehicle has to incorporate both the reflectance of the net, transmission of radiance through the net structure, the complex reflected illumination field illustrated on the vehicle, and transmission back through the net and the atmosphere. Figure 14.23(a) includes a real image of the study site for comparison.

Figure 14.24 shows three bands from a synthetic 210-band imaging spectrometer image of a large complex synthetic image developed for testing and evaluation of advanced hyperspectral image analysis algorithms. Barcomb et al. (2004) and Peterson et al. (2004) discuss a comparison of the results obtained when various algorithms are run on real and synthetic data indicating that synthetic data can reasonably mimic many features of real data.

One of the most powerful uses of SIG is to help us visualize and evaluate the importance of various phenomena on the remotely sensed image. Throughout the earlier chapters, we have repeatedly emphasized the importance of the atmosphere on the radiance levels reaching the sensor. If the atmosphere is properly modeled, the SIG process can be an extremely powerful tool in evaluating the impact of the atmosphere on image characteristics. It can also be used in reverse engineering studies to evaluate how well an approach works in removing atmospheric effects. Figure 14.25 shows a dramatic example of how the radiance from the sky will vary as a function of location and wavelength. These variations are reflected (literally and figuratively) in a muted form by the specular component of the reflectance of objects in the scene. Analysis of synthetic images can help us determine how important any particular phenomenon is to our image analysis, as well as how to remove or take advantage of those phenomenon that are significant.

One important issue regarding SIG has to do with the quality or fidelity of the models. How good are the SIG representations? Of course, the answer is usually dependent on the application. Some applications need only to determine the relative contrast between a target and a background. Others need to properly reproduce mean radiance levels and all the within-object variations. Some need to reproduce only the statistical behavior of backgrounds and certain features of a critical target set. One way to put this is that a SIG image that correctly mimics a real image for those parameters we look at, or measure, is acceptable. Clearly, a measure based on these criteria would not be robust, since a slight change in what we look at (e.g., addition of a new segmentation feature) could cause a catastrophic failure in the effective quality of the SIG representation. To date, there has been only limited development of procedures and metrics for assessment of the quality of SIG images. The state of the art has been that reasonableness was sufficient (i.e., if the image looked reasonably like an actual image or at least mimicked those parameters of interest, that was sufficient). With improvements in the SIG modeling process, especially in the area of radiometric modeling, we are asking more and more of the SIG tools and therefore

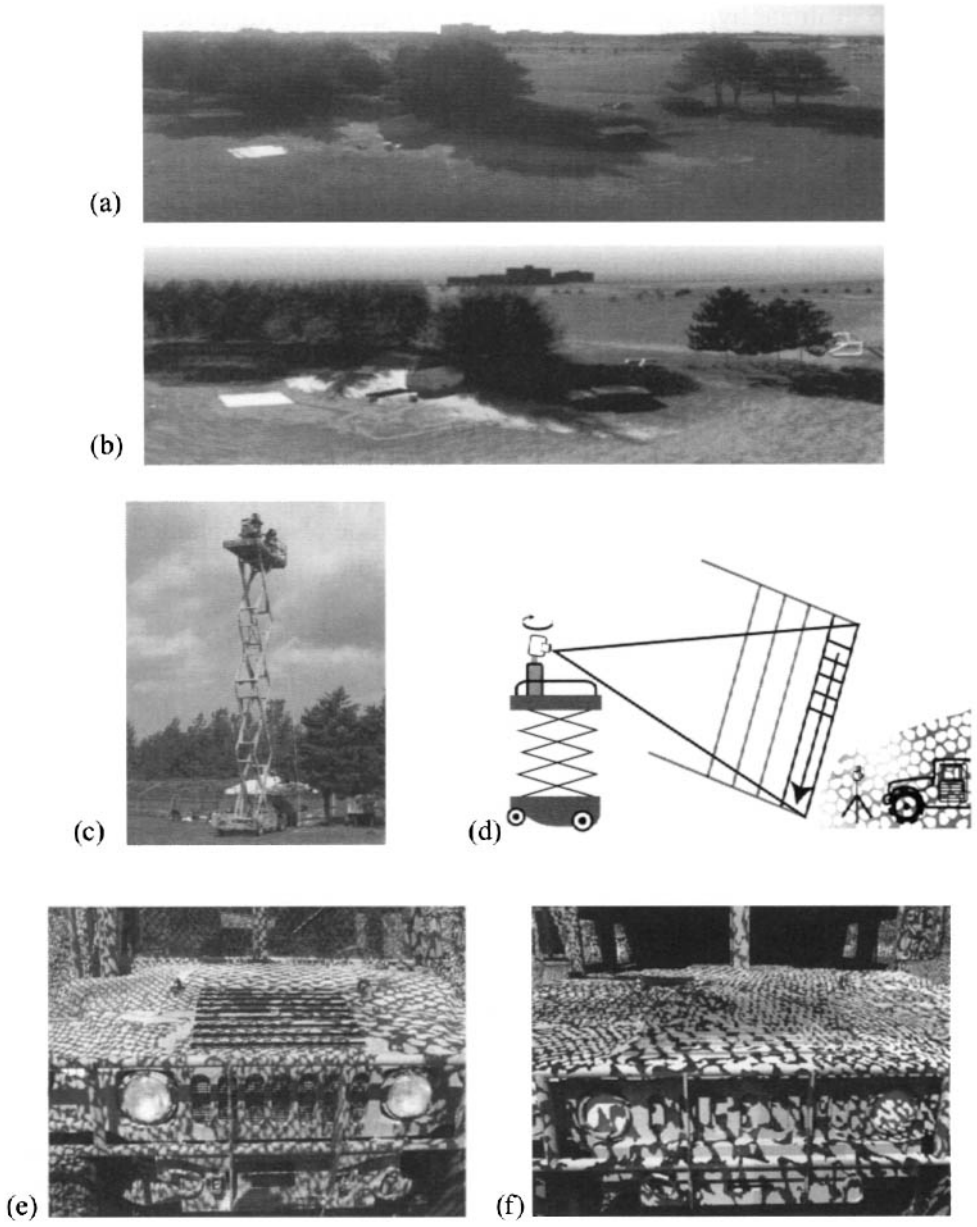


Figure 14.23 Illustration of one band from a real (a) and synthetic (b) hyperspectral line scanner image used for algorithm development. (c) RIT's MISI line scanner used to acquire the real image operating from a rotating table on the scissor truck, as illustrated in (d). (e) and (f) Real and synthetic images from a framing camera on a tripod under the net as illustrated in (d). See color plate 14.23.

need more rigorous methods for evaluating their capabilities and limitations. Rankin et al. (1992) describe an experimental design and test metrics for evaluation of the mean-level thermodynamic and radiometric aspects of thermal IR SIG images. Mason et al. (1994a,b) provide some additional metrics for relative contrast assessment using rank order statistics. Scanlan et al. (2004) address texture fidelity, Ientilucci and Brown (2003) address spectral variability, and Barcomb et al. (2004) and Peterson et al. (2004) compare target/anomaly detection algorithms on real and synthetic images. However, rather than answering fully the question of how to evaluate the total SIG process, these authors only offer some possible metrics for addressing certain issues. The more fundamental question of SIG image quality will undoubtedly be as ubiquitous as the general question of how to characterize image quality.

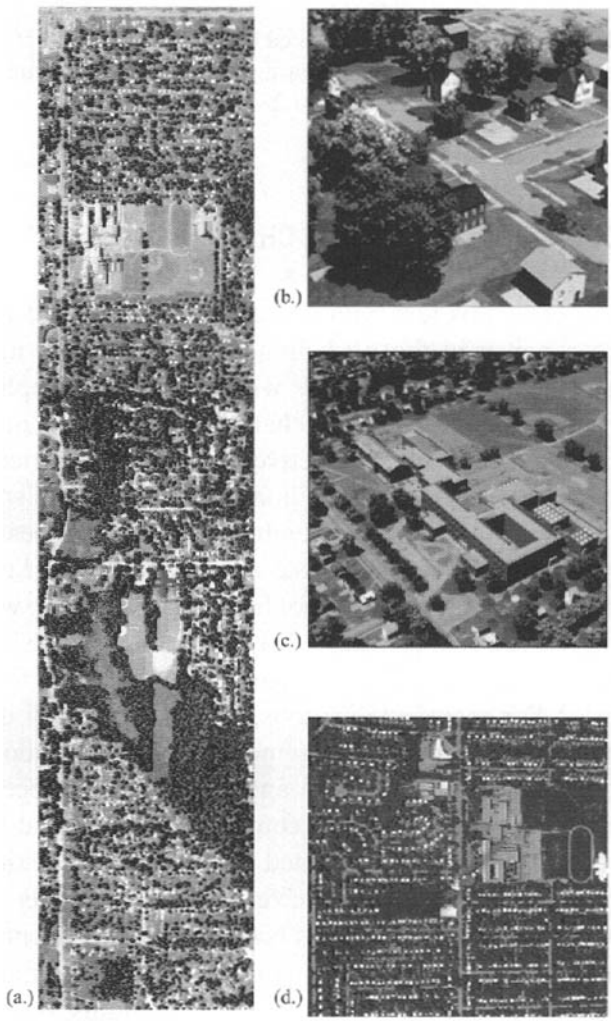


Figure 14.24 (a) True color representation of a portion of a 210-band hyperspectral image of a large synthetic scene built to support hyperspectral algorithm testing and instrument design trades (b) and (c) zooms of a subsection of the scene showing spatial detail available, and (d) portion of the same synthetic scene showing a single band of a synthetic LWIR spectrometer. See color plate 14.24.

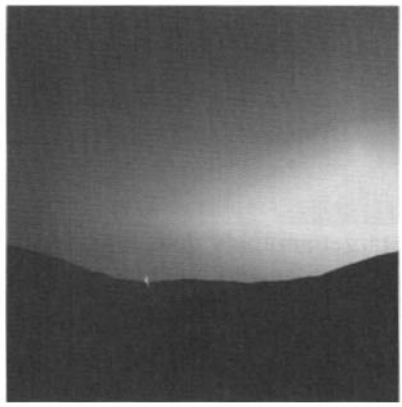


Figure 14.25 A synthetic sunset produced by DIRSIG. The effect is possible due to the extensive spectral modeling that DIRSIG incorporates with the help of MODTRAN. See color plate 14.23.

14.5 SIG MODELING AND THE IMAGE CHAIN APPROACH

We have chosen to close this text with a treatment of synthetic-image generation because it can be a powerful tool in helping to visualize the image chain. We emphasized sensor modeling, since that's where the initial emphasis on SIG has been, and modeling the rest of the image chain (processing and output) is a logical extension. From the image chain perspective, SIG offers a particular value in that if properly implemented, it merges the radiometric, spatial, and spectral strands of the image chain in a way that closely resembles the actual process. The subtleties of the interactions within and between these strands of the image chain can be very difficult to study and analyze without a tool like SIG. However, with all its pluses, it should not be perceived as a panacea. The ultimate goal of SIG is to mirror all the relevant aspects of the world being imaged. This is, for the foreseeable future, an impossible task. For many applications, SIG may already do an acceptable job and may fulfill more and more requirements as modeling tools are improved. However, the complexities of the world and the imaging process are such that many fine points are currently, and will continue to be, missed by the modeling process. Thus, the models should be treated as tools that approximate the process but do not fully represent it. Such tools are very valuable in understanding and analyzing images, but at best they are only as good as our fundamental understanding of the process we are studying. SIG can help us in designing, analyzing, and often reducing the extent of field studies and collection programs, but it cannot replace them. In short, if you want to know what the world looks like, don't forget to look out the window.

14.6 REFERENCES

Barcomb, K.E. Brown, S.D., Hattenberger, T.J., & Schott, J.R. (2004). High-resolution slant-angle scene generation and validation of concealed targets in DIRSIG. *Proceedings SPIE*, Vol. 5546, No. 38, pp. 300-311.

Berk, A., Bernstein, L.S., & Robertson, D.C. (1989). "MODTRAN: a moderate resolution model for LOWTRAN 7." GL-TR-89-0122, Spectral Sciences Inc., Burlington, MA.

Cathcart, J.M., Faust, N.L., Sheffer, A.D., Jr., & Rodriquez, L.J. (1993). Background clutter models for scene simulation. *Proceedings SPIE*, Vol. 1938, No. 37, pp. 325-336.

Curran, A.R., and Curlee, J.S. (2006). "Integrating CameoSim and MuSES to Support Vehicle-Terrain Interaction in an IR Synthetic Scene." Proceedings SPIE, Vol. 6239-14.

DCS Corporation. (1991). "AIRSIM thermal signature prediction and analysis tool model assumptions and analytical foundations." DCS Technical Note 9090-002-001.

Foley, J.D., vanDam, A., Feiner, S.K., & Hughes, J.F. (1990). Computing Graphics: Principles and Practices. 2nd. ed., Addison-Wesley, Reading, MA.

Francis, J., Maver, L., & Schott, J.R. (1993). Comparison of physically and computer generated imagery. *Proceedings SPIE*, Vol. 1904, pp. 20-23.

Haruyama, S., & Barsky, B.A. (1984). Using stochastic modeling for texture generation. *IEEE Computer Graphics and Applications*, Vol. 4, pp. 7-19.

Ientilucci, E.J., & Brown, S.D. (2003). Advances in wide-area hyperspectral image simulation. *Proceedings SPIE*, Vol. 5075, pp. 110-121.

Johnson, K.R., Curran, A.R., & Gondo, T.G. (1993). Development of a signature super code. *Proceedings SPIE*, Vol. 1938, No. 36, pp. 317-304.

Kornfeld, G.H., & Penn, J. (1993). Various FLIR sensor effects applied to synthetic thermal imagery. *Proceedings SPIE*, Vol. 38, No. 39, pp. 350-367.

Mason, J.E., Schott, J.R., & Rankin-Parobek, D. (1994a) Validation analysis of the thermal and radiometric integrity of RIT's synthetic image generation model, DIRSIG. *Proceedings SPIE*, Vol. 2223, pp. 474-487.

Mason, J.E., Schott, J.R., Salvaggio, C., & Sirianni, J.D. (1994b). Validation of contrast and phenomenology in the digital imaging and remote sensing (DIRS) lab's image generation (DIRSIG) model. *Proceedings SPIE*, Vol. 2269, pp. 622-633.

Maver, L., & Scarff, L. (1993). Multispectral image simulation. *Proceedings SPIE*, Vol. 1904, pp. 144-160.

Moorhead, I.R., Gilmore, M.A., Houlbrook, A.W., Oxford, D.E., Filbee, D.R., Stroud, C.A., Hutchings, G., & Kirk, A. (2001). CAMEO-SIM: a physics-based broadband scene simulation tool for assessment of camouflage, concealment, and deception methodologies. *Optical Engineering*, Vol. 40(09), pp. 1896-1905.

Pentland, A.P. (1984). Fractal-based description of natural scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-6, No. 6, pp. 661-674.

Peterson, E.D., Brown, S.D., Hattenberger, T.J., & Schott, J.R. (2004). Surface and burled landmine scene generation and validation using digital imaging and remote sensing image generation (DIRSIG) model. *Proceedings SPIE*, Vol. 5546, No. 39, pp. 312-323

Rankin, D., Salvaggio, C., Gallagher, T., & Schott, J.R. (1992). Instrumentation and procedures for validation of synthetic infrared image generation (SIG) models. *Proceedings SPIE*, Vol. 1762, pp. 584-600.

Ranson, K.J., Irons, J.R., & Williams, D.L. (1994). Multispectral bidirectional reflectance of northern forest canopies with the advanced solid-state array spectroradiometer (ASAS). *Remote Sensing of Environment*, Vol. 47, pp. 276-289.

Sanders, J.S., Johnson, K.R., Curran, A.R., & Rynes, P.L. (2000). Ground target infrared signature modeling with the multiservice electro-optic signature (MuSES) code. *Proceedings SPIE*, Vol. 4029, pp. 197-204.

Savage, J., Coker, C., and Edwards, D. (2006). Irma 5.1 multi-sensor signature prediction model. *Proceedings SPIE*, Vol. 6239-12.

Scanlan, N.W., Schott, J.R., & Brown, S.D. (2004). Performance analysis of improved methodology for incorporation of spatial/spectral variability in synthetic hyperspectral imagery. *Proceedings SPIE*, Vol. 5159, pp. 319-330.

Schott, J.R., Raqueño, R., & Salvaggio, C. (1992). Incorporation of time-dependent thermodynamic model and a radiation propagation model into infrared three-dimensional synthetic image generation. *Optical Engineering*, Vol. 31, No. 7, pp. 1505-1516.

Schott, J.R., Salvaggio, C., Brown, S.D., & Rose, R. (1995). "Incorporation of texture in multispectral synthetic image generation models." Proceedings SPIE, Vol. 2469, pp. 189-196.

Schott, J.R., Brown, S.D., Raqueño, R.V., Gross, H.N., & Robinson, G. (1999). An advanced synthetic image generation model and its application to multi/hyperspectral algorithm development. *Canadian Journal of Remote Sensing*, Vol. 25, No. 2, pp. 99-111.

Sheffer, A.D., Jr., Cathcart, J.M., & Stewart, J.M. (1993). Ocean background model for scene simulation. *Proceedings SPIE*, Vol. 1938, No. 38, pp. 337-349.

Stewart, S.R., Lyons, J.T., & Horvath, R. Simulated infrared imaging (SIRIM): "user's tool for simulating target signatures (U)." Environmental Research Institute of Michigan.

APPENDIX A
BLACKBODY CALCULATIONS

Key	
Spectral Radiance	$L_{T\lambda} = \frac{2hc^2}{\lambda^5}\left(e^{hc/\lambda kT} - 1\right)^{-1}$
Cumulative Radiance	$\int_0^{\lambda} L_{T\lambda} d\lambda$
Photon Spectral Radiance	$L_{p\lambda} = L_{T\lambda} / \left(hc/\lambda\right)$
Cumulative Spectral Radiance	$\int_0^{\lambda} L_{p\lambda} d\lambda$

Target Temperature = 77 K				
Wavelength λ[μm]	Spectral Radiance [$Wm^{-2}sr^{-1}\mu m^{-1}$]	Cumulative Radiance $L_{p\lambda}$ [$Wm^{-2}sr^{-1}$]	Photon Spectral Radiance $L_{p\lambda}$ [photons/$secm^2srμm$]	Cumulative Photon Radiance [photons/sec m^2sr]
3	4.367623E-22	4.513051E-23	6.596032E-03	6.800893E-04
3.2	1.551428E-20	1.640235E-21	2.499183E-01	2.633152E-02
3.4	3.554836E-19	3.874002E-20	6.084358E+00	6.597592E-01
3.6	5.658301E-18	6.401185E-19	1.025427E+02	1.152247E+01
3.8	6.632660E-17	7.838608E-18	1.268784E+03	1.486551E+02
4	5.998908E-16	7.446034E-17	1.207949E+04	1.483482E+03
4.2	4.347104E-15	5.691599E-16	9.191062E+04	1.188249E+04
4.4	2.602751E-14	3.606621E-15	5.765031E+05	7.872386E+04
4.6	1.320647E-13	1.941584E-14	3.058165E+06	4.421907E+05
4.8	5.799977E-13	9.062209E-14	1.401470E+07	2.149477E+06
5	2.244060E-12	3.730279E-13	5.648341E+07	9.199289E+06
8	2.610377E-07	1.039095E-07	1.051259E+13	4.006759E+12
8.2	4.078432E-07	1.707976E-07	1.683540E+13	6.741558E+12
8.4	6.220052E-07	2.737824E-07	2.630205E+13	1.105530E+13
8.6	9.276126E-07	4.287442E-07	4.015887E+13	1.770139E+13
8.8	1.354883E-06	6.569938E-07	6.002067E+13	2.771935E+13
9	1.941009E-06	9.865829E-07	8.793997E+13	4.251541E+13
9.2	2.730964E-06	1.453780E-06	1.264795E+14	6.395736E+13
9.4	3.778243E-06	2.104701E-06	1.787862E+14	9.448393E+13
9.6	5.145496E-06	2.997075E-06	2.486651E+14	1.372291E+14
9.8	6.905042E-06	4.202128E-06	3.406503E+14	1.961606E+14
10	9.139227E-06	5.806555E-06	4.600720E+14	2.762328E+14
10.2	1.194062E-05	7.914540E-06	6.131168E+14	3.835517E+14
10.4	1.541200E-05	1.064980E-05	8.068797E+14	5.255514E+14
10.6	1.966623E-05	1.415762E-05	1.049406E+15	7.111799E+14
10.8	2.482582E-05	1.860683E-05	1.349720E+15	9.510925E+14
11	3.102241E-05	2.419165E-05	1.717848E+15	1.257849E+15
11.2	3.839603E-05	3.113350E-05	2.164814E+15	1.646115E+15
11.4	4.709418E-05	3.968252E-05	2.702641E+15	2.132861E+15
11.6	5.727081E-05	5.011902E-05	3.344319E+15	2.737557E+15
11.8	6.908513E-05	6.275461E-05	4.103770E+15	3.482366E+15
12	8.270037E-05	7.793316E-05	4.995800E+15	4.392323E+15
12.2	9.828235E-05	9.603143E-05	6.036034E+15	5.495506E+15
12.4	1.159982E-04	1.174595E-04	7.240843E+15	6.823194E+15
12.6	1.360147E-04	1.426608E-04	8.627257E+15	8.410004E+15
12.8	1.584972E-04	1.721120E-04	1.021287E+16	1.029402E+16
13	1.836077E-04	2.063224E-04	1.201574E+16	1.251688E+16
13.2	2.115041E-04	2.458336E-04	1.405430E+16	1.512388E+16
13.4	2.423383E-04	2.912179E-04	1.634719E+16	1.816403E+16
13.6	2.762550E-04	3.430772E-04	1.891322E+16	2.169007E+16
13.8	3.133911E-04	4.020418E-04	2.177119E+16	2.575851E+16
14	3.538739E-04	4.687683E-04	2.493980E+16	3.042961E+16

Target Temperature = 273 K				
Wavelengthλ[μm]	Spectral Radiance [$Wm^{-2}sr^{-1}\mu m^{-1}$]	Cumulative Radiance $L_{p\lambda}$ [$Wm^{-2}sr^{-1}$]	Photon Spectral Radiance $L_{p\lambda}$ [photons/$secm^2srμm$]	Cumulative Photon Radiance [photons/sec m^2sr]
3	1.150317E-02	2.499939E-03	1.737220E+17	3.583590E+16
3.2	2.497574E-02	6.147829E-03	4.023320E+17	9.344130E+16
3.4	4.859804E-02	1.350521E-02	8.317910E+17	2.168540E+17
3.6	8.639658E-02	2.700467E-02	1.565720E+18	4.566050E+17
3.8	1.424685E-01	4.989118E-02	2.725330E+18	8.857100E+17
4	2.205476E-01	8.619279E-02	4.440980E+18	1.602340E+18
4.2	3.236227E-01	1.406098E-01	6.842340E+18	2.730670E+18
4.4	4.536643E-01	2.183385E-01	1.004860E+19	4.419760E+18
4.6	6.114796E-01	3.248529E-01	1.415980E+19	6.840600E+18
4.8	7.966938E-01	4.656703E-01	1.925080E+19	1.018170E+19
5	1.007835E+00	6.461231E-01	2.536740E+19	1.464350E+19
8	5.012061E+00	9.822368E+00	2.018470E+20	3.307830E+20
8.2	5.203356E+00	1.084391E+01	2.147900E+20	3.724470E+20
8.4	5.376899E+00	1.190193E+01	2.273670E+20	4.166630E+20
8.6	5.532646E+00	1.299289E+01	2.395230E+20	4.633520E+20
8.8	5.670760E+00	1.411323E+01	2.512120E+20	5.124250E+20
9	5.791576E+00	1.525946E+01	2.623950E+20	5.637860E+20
9.2	5.895570E+00	1.642818E+01	2.730420E+20	6.173300E+20
9.4	5.983332E+00	1.761607E+01	2.831310E+20	6.729470E+20
9.6	6.055538E+00	1.881996E+01	2.926440E+20	7.305250E+20
9.8	6.112933E+00	2.003680E+01	3.015730E+20	7.899460E+20
10	6.156305E+00	2.126373E+01	3.099110E+20	8.510950E+20
10.2	6.186474E+00	2.249800E+01	3.176580E+20	9.138520E+20
10.4	6.204274E+00	2.373708E+01	3.248180E+20	9.780990E+20
10.6	6.210542E+00	2.497856E+01	3.313990E+20	1.043720E+21
10.8	6.206108E+00	2.622023E+01	3.374110E+20	1.110600E+21
11	6.191790E+00	2.746002E+01	3.428670E+20	1.178630E+21
11.2	6.168380E+00	2.869603E+01	3.477810E+20	1.247690E+21
11.4	6.136648E+00	2.992653E+01	3.521700E+20	1.317690E+21
11.6	6.097331E+00	3.114993E+01	3.560530E+20	1.388510E+21
11.8	6.051135E+00	3.236478E+01	3.594470E+20	1.460060E+21
12	5.998730E+00	3.356977E+01	3.623740E+20	1.532240E+21
12.2	5.940752E+00	3.476371E+01	3.648530E+20	1.604970E+21
12.4	5.877798E+00	3.594557E+01	3.669040E+20	1.678140E+21
12.6	5.810430E+00	3.711439E+01	3.685490E+20	1.751690E+21
12.8	5.739174E+00	3.826935E+01	3.698080E+20	1.825520E+21
13	5.664521E+00	3.940972E+01	3.707000E+20	1.899570E+21
13.2	5.586927E+00	4.053487E+01	3.712470E+20	1.973770E+21
13.4	5.506816E+00	4.164424E+01	3.714680E+20	2.048040E+21
13.6	5.424578E+00	4.273738E+01	3.713820E+20	2.122330E+21
13.8	5.340575E+00	4.381390E+01	3.710080E+20	2.196560E+21
14	5.255140E+00	4.487347E+01	3.703640E+20	2.270700E+21

Target Temperature = 290 K				
Wavelength λ[μm]	Spectral Radiance [$Wm^{-2}sr^{-1}\mu m^{-1}$]	Cumulative Radiance $L_{p\lambda}$ [$Wm^{-2}sr^{-1}$]	Photon Spectral Radiance $L_{p\lambda}$ [photons/$secm^2srμm$]	Cumulative Photon Radiance [photons/sec m^2sr]
3	3.221549E-02	7.451313E-03	4.865220E+17	1.063140E+17
3.2	6.558622E-02	1.723148E-02	1.056520E+18	2.606180E+17
3.4	1.205726E-01	3.584737E-02	2.063690E+18	5.726400E+17
3.6	2.037993E-01	6.828456E-02	3.693360E+18	1.148340E+18
3.8	3.212250E-01	1.207870E-01	6.144820E+18	2.132160E+18
4	4.774618E-01	2.006557E-01	9.614240E+18	3.708070E+18
4.2	6.753096E-01	3.159328E-01	1.427800E+19	6.097300E+18
4.4	9.155428E-01	4.750181E-01	2.027910E+19	9.553010E+18
4.6	1.196932E+00	6.862655E-01	2.771680E+19	1.435260E+19
4.8	1.516448E+00	9.576035E-01	3.664250E+19	2.078850E+19
5	1.869595E+00	1.296208E+00	4.705800E+19	2.915860E+19
8	7.379321E+00	1.583242E+01	2.971820E+20	5.265860E+20
8.2	7.589838E+00	1.732934E+01	3.133020E+20	5.876340E+20
8.4	7.773697E+00	1.886569E+01	3.287180E+20	6.518360E+20
8.6	7.931580E+00	2.043622E+01	3.433800E+20	7.190460E+20
8.8	8.064382E+00	2.203582E+01	3.572480E+20	7.891090E+20
9	8.173163E+00	2.365957E+01	3.702960E+20	8.618630E+20
9.2	8.259106E+00	2.530280E+01	3.825050E+20	9.371430E+20
9.4	8.323481E+00	2.696106E+01	3.938670E+20	1.014780E+21
9.6	8.367611E+00	2.863017E+01	4.043790E+20	1.094610E+21
9.8	8.392848E+00	3.030621E+01	4.140490E+20	1.176450E+21
10	8.400548E+00	3.198555E+01	4.228870E+20	1.260140E+21
10.2	8.392057E+00	3.366481E+01	4.309080E+20	1.345520E+21
10.4	8.368693E+00	3.534089E+01	4.381340E+20	1.432430E+21
10.6	8.331737E+00	3.701093E+01	4.445880E+20	1.520700E+21
10.8	8.282420E+00	3.867235E+01	4.502950E+20	1.610190E+21
11	8.221922E+00	4.032278E+01	4.552840E+20	1.700740E+21
11.2	8.151366E+00	4.196011E+01	4.595840E+20	1.792230E+21
11.4	8.071811E+00	4.358243E+01	4.632250E+20	1.884510E+21
11.6	7.984257E+00	4.518803E+01	4.662390E+20	1.977460E+21
11.8	7.889639E+00	4.677542E+01	4.686570E+20	2.070950E+21
12	7.788830E+00	4.834327E+01	4.705110E+20	2.164860E+21
12.2	7.682644E+00	4.989042E+01	4.718310E+20	2.259100E+21
12.4	7.571832E+00	5.141587E+01	4.726490E+20	2.353550E+21
12.6	7.457089E+00	5.291876E+01	4.729950E+20	2.448110E+21
12.8	7.339057E+00	5.439837E+01	4.728970E+20	2.542700E+21
13	7.218322E+00	5.585411E+01	4.723850E+20	2.637230E+21
13.2	7.095423E+00	5.728548E+01	4.714860E+20	2.731620E+21
13.4	6.970851E+00	5.869211E+01	4.702260E+20	2.825790E+21
13.6	6.845053E+00	6.007370E+01	4.686320E+20	2.919670E+21
13.8	6.718434E+00	6.143005E+01	4.667280E+20	3.013210E+21
14	6.591361E+00	6.276103E+01	4.645360E+20	3.106340E+21

Target Temperature = 297 K				
Wavelength λ[μm]	Spectral Radiance [$Wm^{-2}sr^{-1}\mu m^{-1}$]	Cumulative Radiance $L_{p\lambda}$ [$Wm^{-2}sr^{-1}$]	Photon Spectral Radiance $L_{p\lambda}$ [photons/$secm^2srμm$]	Cumulative Photon Radiance [photons/sec m^2sr]
3	4.757115E-02	1.128275E-02	7.184250E+17	1.606710E+17
3.2	9.451737E-02	2.549161E-02	1.522570E+18	3.847710E+17
3.4	1.700641E-01	5.194975E-02	2.910770E+18	8.281050E+17
3.6	2.820127E-01	9.715743E-02	5.110780E+18	1.630260E+18
3.8	4.369693E-01	1.690556E-01	8.358930E+18	2.977230E+18
4	6.395852E-01	2.767111E-01	1.287880E+19	5.101000E+18
4.2	8.921081E-01	4.298804E-01	1.886180E+19	8.275060E+18
4.4	1.194257E+00	6.385169E-01	2.645250E+19	1.280650E+19
4.6	1.543374E+00	9.122800E-01	3.573920E+19	1.902570E+19
4.8	1.934777E+00	1.260095E+00	4.675070E+19	2.727470E+19
5	2.362221E+00	1.689795E+00	5.945750E+19	3.789550E+19
8	8.543480E+00	1.900672E+01	3.440660E+20	6.289180E+20
8.2	8.756310E+00	2.073669E+01	3.614530E+20	6.994700E+20
8.4	8.938435E+00	2.250617E+01	3.779700E+20	7.734120E+20
8.6	9.090938E+00	2.430911E+01	3.935710E+20	8.505660E+20
8.8	9.215108E+00	2.613971E+01	4.082250E+20	9.307460E+20
9	9.312380E+00	2.799246E+01	4.219100E+20	1.013760E+21
9.2	9.384298E+00	2.986213E+01	4.346160E+20	1.099410E+21
9.4	9.432468E+00	3.174380E+01	4.463440E+20	1.187510E+21
9.6	9.458527E+00	3.363290E+01	4.571000E+20	1.277850E+21
9.8	9.464118E+00	3.552517E+01	4.668990E+20	1.370250E+21
10	9.450862E+00	3.741667E+01	4.757600E+20	1.464520E+21
10.2	9.420348E+00	3.930379E+01	4.837080E+20	1.560460E+21
10.4	9.374112E+00	4.118323E+01	4.907720E+20	1.657910E+21
10.6	9.313632E+00	4.305201E+01	4.969830E+20	1.756690E+21
10.8	9.240317E+00	4.490740E+01	5.023740E+20	1.856620E+21
11	9.155503E+00	4.674698E+01	5.069800E+20	1.957560E+21
11.2	9.060450E+00	4.856858E+01	5.108390E+20	2.059340E+21
11.4	8.956341E+00	5.037026E+01	5.139870E+20	2.161820E+21
11.6	8.844281E+00	5.215032E+01	5.164600E+20	2.264870E+21
11.8	8.725298E+00	5.390728E+01	5.182970E+20	2.368340E+21
12	8.600345E+00	5.563984E+01	5.195330E+20	2.472130E+21
12.2	8.470302E+00	5.734691E+01	5.202060E+20	2.576100E+21
12.4	8.335980E+00	5.902754E+01	5.203490E+20	2.680160E+21
12.6	8.198123E+00	6.068095E+01	5.199980E+20	2.784190E+21
12.8	8.057410E+00	6.230650E+01	5.191850E+20	2.888110E+21
13	7.914462E+00	6.390369E+01	5.179420E+20	2.991820E+21
13.2	7.769843E+00	6.547212E+01	5.163000E+20	3.095250E+21
13.4	7.624065E+00	6.701151E+01	5.142900E+20	3.198310E+21
13.6	7.477590E+00	6.852167E+01	5.119370E+20	3.300930E+21
13.8	7.330833E+00	7.000252E+01	5.092710E+20	3.403050E+21
14	7.184170E+00	7.145402E+01	5.063150E+20	3.504610E+21

Target Temperature = 300 K				
Wavelengthλ[μm]	Spectral Radiance [$Wm^{-2}sr^{-1}\mu m^{-1}$]	Cumulative Radiance $L_{p\lambda}$ [$Wm^{-2}sr^{-1}$]	Photon Spectral Radiance $L_{p\lambda}$ [photons/$secm^2sr\mu m$]	Cumulative Photon Radiance [photons/sec m^2sr]
3	5.590794E-02	1.340220E-02	8.443280E+17	1.906960E+17
3.2	1.099660E-01	2.998959E-02	1.771430E+18	4.522720E+17
3.4	1.961064E-01	6.059683E-02	3.356500E+18	9.650660E+17
3.6	3.226342E-01	1.124709E-01	5.846950E+18	1.885410E+18
3.8	4.963828E-01	1.943726E-01	9.495480E+18	3.419650E+18
4	7.219317E-01	3.162040E-01	1.453690E+19	5.822890E+18
4.2	1.001177E+00	4.885149E-01	2.116780E+19	9.393370E+18
4.4	1.333259E+00	7.219584E-01	2.953140E+19	1.446330E+19
4.6	1.714784E+00	1.026763E+00	3.970850E+19	2.138730E+19
4.8	2.140247E+00	1.412266E+00	5.171560E+19	3.052970E+19
5	2.602563E+00	1.886547E+00	6.550700E+19	4.225190E+19
8	9.078147E+00	2.050875E+01	3.655980E+20	6.771220E+20
8.2	9.290742E+00	2.234564E+01	3.835130E+20	7.520330E+20
8.4	9.470846E+00	2.422179E+01	4.004830E+20	8.304320E+20
8.6	9.619732E+00	2.613085E+01	4.164640E+20	9.121270E+20
8.8	9.738872E+00	2.806671E+01	4.314270E+20	9.969160E+20
9	9.829880E+00	3.002359E+01	4.453560E+20	1.084590E+21
9.2	9.894462E+00	3.199602E+01	4.582440E+20	1.174950E+21
9.4	9.934379E+00	3.397891E+01	4.700940E+20	1.267790E+21
9.6	9.951411E+00	3.596749E+01	4.809190E+20	1.362890E+21
9.8	9.947331E+00	3.795736E+01	4.907370E+20	1.460060E+21
10	9.923879E+00	3.994448E+01	4.995720E+20	1.559090E+21
10.2	9.882751E+00	4.192514E+01	5.074510E+20	1.659790E+21
10.4	9.825581E+00	4.389598E+01	5.144080E+20	1.761970E+21
10.6	9.753932E+00	4.585393E+01	5.204770E+20	1.865460E+21
10.8	9.669289E+00	4.779625E+01	5.256960E+20	1.970080E+21
11	9.573057E+00	4.972048E+01	5.301020E+20	2.075660E+21
11.2	9.466556E+00	5.162445E+01	5.337360E+20	2.182040E+21
11.4	9.351018E+00	5.350620E+01	5.366360E+20	2.289080E+21
11.6	9.227595E+00	5.536406E+01	5.388440E+20	2.396630E+21
11.8	9.097353E+00	5.719656E+01	5.403980E+20	2.504550E+21
12	8.961277E+00	5.900242E+01	5.413370E+20	2.612730E+21
12.2	8.820277E+00	6.078058E+01	5.416990E+20	2.721030E+21
12.4	8.675185E+00	6.253012E+01	5.415230E+20	2.829350E+21
12.6	8.526764E+00	6.425032E+01	5.408430E+20	2.937590E+21
12.8	8.375711E+00	6.594057E+01	5.396950E+20	3.045640E+21
13	8.222657E+00	6.760040E+01	5.381110E+20	3.153420E+21
13.2	8.068178E+00	6.922949E+01	5.361250E+20	3.260850E+21
13.4	7.912791E+00	7.082758E+01	5.337660E+20	3.367840E+21
13.6	7.756963E+00	7.239456E+01	5.310640E+20	3.474320E+21
13.8	7.601113E+00	7.393037E+01	5.280470E+20	3.580230E+21
14	7.445618E+00	7.543504E+01	5.247410E+20	3.685510E+21

Target Temperature = 5800 K				
Wavelengthλ[μm]	Spectral Radiance [$Wm^{-2}sr^{-1}\mu m^{-1}$]	Cumulative Radiance $L_{\rho\lambda}$ [$Wm^{-2}sr^{-1}$]	Photon Spectral Radiance $L_{p\lambda}$ [photons/secm^2srμm]	Cumulative Photon Radiance [photons/sec m^2sr]
0.1	2.006962E+02	0.000000E+00	1.010310E+20	0.000000E+00
0.2	1.527878E+06	7.640392E+04	1.538280E+24	7.691900E+22
0.3	1.256920E+07	7.812579E+05	1.898210E+25	1.102940E+24
0.4	2.361383E+07	2.590409E+06	4.754910E+25	4.429500E+24
0.5	2.688175E+07	5.115188E+06	6.766190E+25	1.019010E+25
0.6	2.492426E+07	7.705489E+06	7.528180E+25	1.733720E+25
0.7	2.109236E+07	1.000632E+07	7.432580E+25	2.481760E+25
0.8	1.713215E+07	1.191754E+07	6.899510E+25	3.198370E+25
0.9	1.368304E+07	1.345830E+07	6.199280E+25	3.853310E+25
1	1.087805E+07	1.468636E+07	5.476050E+25	4.437070E+25
1.1	8.663218E+06	1.566342E+07	4.797210E+25	4.950730E+25
1.2	6.934267E+06	1.644330E+07	4.188880E+25	5.400040E+25
1.3	5.587685E+06	1.706939E+07	3.656720E+25	5.792320E+25
1.4	4.536203E+06	1.757559E+07	3.196960E+25	6.135000E+25
1.5	3.710890E+06	1.798794E+07	2.802110E+25	6.434960E+25
1.6	3.058856E+06	1.832643E+07	2.463740E+25	6.698250E+25
1.7	2.540014E+06	1.860637E+07	2.173710E+25	6.930120E+25
1.8	2.124097E+06	1.883958E+07	1.924700E+25	7.135040E+25
1.9	1.788223E+06	1.903520E+07	1.710380E+25	7.316800E+25
2	1.515029E+06	1.920036E+07	1.525340E+25	7.478580E+25
2.1	1.291267E+06	1.934067E+07	1.365060E+25	7.623100E+25
2.2	1.106769E+06	1.946057E+07	1.225730E+25	7.752640E+25
2.3	9.536757E+05	1.956360E+07	1.104190E+25	7.869140E+25
2.4	8.258708E+05	1.965257E+07	9.977910E+24	7.974240E+25
2.5	7.185637E+05	1.972980E+07	9.043190E+24	8.069340E+25
2.6	6.279752E+05	1.979712E+07	8.219250E+24	8.155660E+25
2.7	5.511054E+05	1.985608E+07	7.490570E+24	8.234200E+25
2.8	4.855570E+05	1.990791E+07	6.844070E+24	8.305880E+25
2.9	4.294028E+05	1.995366E+07	6.268730E+24	8.371440E+25
3	3.810846E+05	1.999418E+07	5.755180E+24	8.431560E+25
3.1	3.393350E+05	2.003020E+07	5.295500E+24	8.486810E+25
3.2	3.031177E+05	2.006233E+07	4.882900E+24	8.537710E+25
3.3	2.715810E+05	2.009106E+07	4.511590E+24	8.584680E+25
3.4	2.440213E+05	2.011684E+07	4.176600E+24	8.628120E+25
3.5	2.198547E+05	2.014004E+07	3.873650E+24	8.668370E+25
3.6	1.985944E+05	2.016096E+07	3.599030E+24	8.705730E+25
3.7	1.798328E+05	2.017988E+07	3.349550E+24	8.740480E+25
3.8	1.632270E+05	2.019703E+07	3.122430E+24	8.772840E+25
3.9	1.484877E+05	2.021262E+07	2.915220E+24	8.803030E+25
4	1.353696E+05	2.022681E+07	2.725820E+24	8.831230E+25
4.1	1.236641E+05	2.023976E+07	2.552370E+24	8.857620E+25
4.2	1.131931E+05	2.025161E+07	2.393240E+24	8.882350E+25
4.3	1.038041E+05	2.026245E+07	2.246980E+24	8.905550E+25
4.4	9.536604E+04	2.027241E+07	2.112330E+24	8.927350E+25
4.5	8.776603E+04	2.028157E+07	1.988180E+24	8.947850E+25
4.6	8.090638E+04	2.029000E+07	1.873510E+24	8.967160E+25

This page intentionally left blank

INDEX

C

D

E

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z