1672-1-X

THE UNIVERSITY OF MICHIGAN NOTES FOR A PROGRAM OF STUDY IN REMOTE SENSING OF EARTH RESOURCES

14 February 1968—3 May 1968
Conducted at NASA, MSC, Houston, Texas

November 1968
Revised and Edited

Willow Run Laboratories
INSTITUTE OF SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF MICHIGAN
Ann Arbor, Michigan

FOREWORD

These notes are a revised and edited second draft of the notes used in the program of study in Remote Sensing of Earth Resources (14 February 1968 - 3 May 1968) presented at the NASA Manned Spaceflight Center in Houston, Texas. The original set of notes was meaningful only in conjunction with the corresponding oral presentations. A considerable number of attendees were unable to attend all sessions and others were not able to attend at all because of limitations in space and time. This second draft is intended to provide those who were unable to hear the oral presentations with a set of notes that relies less upon oral presentations for completeness. Unfortunately, much of what was presented cannot be recorded in note form. The vast number of slide projected illustrations, the pointing out of salient features of remotely sensed data from these slides, and the discussions and exercises remain unrecorded.

The notes were also reduced in quantity by the removal of redundancy. The principal redundancy was in subjects covered by Jenkins and White, "Fundamentals of Optics," second edition. Failure of this text to arrive in Houston on time forced the lecturers to compose notes in place of the text. The reader of these second draft notes is expected to own this text. New composition has been incorporated to bring the second draft into a form which approximates a continuous exposition of the subject with as near to a consistent nomenclature as possible.

For the records, a summary is given below of the lectures as presented in Houston. The second draft, however, does not accurately follow these lectures in order or content for the reasons stated above, so the lecturers cannot be held wholly accountable for the treatment of the subject in this second draft.

Summary of First Draft Notes

Each summary includes the title and author of the lectures, the purpose of the lecture, the scope and approach, and a listing of the subjects discussed.

I. Fundamentals and Physical Principles (Thomas Limperis, Chairman)

A. International Lighting Vocabulary of the International Commission on Illumination (CIE); standardized nomenclature and definitions of radiometric terms

1. Radiation
2. Visible radiation; light
3. Infrared radiation
4. Wavelength
5. Radiant energy
6. Radiant exitance (at a point of a surface)
7. Planck's law
8. Stefan-Boltzmann law
9. Radiance temperature (luminance temperature)
10. Distribution temperature
11. Color temperature
12. Correlated color temperature

B. Lecture on Electromagnetic Waves by Dr. Gwynn Suits; an elementary discussion of electric and magnetic fields:

1. Field equations
2. Poynting vector
3. Direction of polarizating
4. Radiation of electric and magnetic dipoles
5. Material-field interactions

C. Lecture on Wave Motion, by Dr. Gwynn Suits; a brief elementary discussion of simple harmonic wave motion:

1. Wavelength, frequency, and propagation velocity
2. Phase angle
3. Standing waves
4. Longitudinal and transverse waves
5. Polarization: vertical and horizontal

D. Lecture on Superposition of Waves, by Dr. Gwynn Suits; an elementary discussion of the superposition principle which is both a method for synthesizing complicated waveforms from simple harmonic waves and a mathematical tool for certain kinds of analysis:

1. Power spectrum
2. Fourier components
3. Carrier frequency
4. Spatial frequency
5. Amplitude and phase distortions
6. Fourier transform

E. A Review of Blackbody Radiation Laws, by Dr. Dwight Bornemeier; a detailed analysis and discussion of the laws which govern the electromagnetic radiation of substances and objects. Needed for defining the spectral signatures of targets and backgrounds in remote sensing applications. Topics include:

1. Thermal radiation
2. Kirchoff's law
3. Stefan-Boltzmann's law
4. Planck's law
5. Rayleigh-Jeans and Wien's radiation "laws"
6. Wien's displacement law
7. Graybodies
8. Blackbody tables
9. Blackbody curves
10. GE radiation sliderule

There is also an appendix in which is derived the relationship between the energy density and the radiant emittance.

F. Short Review of the Theory of Stops, by Dr. Dwight Bornemeier; an elementary discussion of stops, which limit the radiation entering an optical system:

1. Aperture stops: entrance and exit pupils
2. Principal (or chief) ray
3. Field stops: entrance and exit windows
4. Relative aperture and f-number
5. Depth of focus, depth of field

G. Lecture on Atomic and Molecular Spectra, by Dr. Gwynn Suits; an elementary discussion of absorption and radiation in terms of energy states:

1. Atomic models
2. Quantum-mechanical theory
3. Energy levels
4. Molecular spectra
5. Band spectra

H. Optics, Mirrors and Lenses, by Dr. I. W. Ginsberg; a detailed discussion and analysis of ideal optical components which are used in almost every form of remote sensor:

1. Classical laws of optics
   a. Reflection
   b. Refraction
2. Dispersion
3. Total reflection
4. Principal plane and principal ray
5. Optic axis
6. Objects and images
   a. Real image
   b. Virtual image
7. Focal points and focal planes
8. Magnification
9. Thin lens equation
10. Principal planes: primary and secondary

I. Notes on Aberrations in Lens Systems, by William L. Brown; a detailed discussion and analysis of deviations from the ideal in lens systems:

1. Spherical aberration
2. Coma
3. Astigmatism
4. Curvature of field
5. Distortion
6. Chromatic aberrations
7. Corrections for lens aberrations
8. Graph of resolution of typical optical systems

J. Interference and Diffraction, by Dr. I. W. Ginsberg; a detailed analysis and discussion of the principles of physical optics as they relate to remote sensing applications:

1. Electric fields
2. Phase velocity
3. Doppler effect
4. Superposition of waves
5. Interference of coherent waves
6. Coherence
7. Thin-film interference
8. Diffraction
   a. Fraunhofer
   b. Fresnel
9. Resolution limits (Rayleigh criterion)

K. Lecture on Structure and Properties of the Human Eye, by Dr. Gwynn Suits; an elementary discussion of the human eye as a remote sensor:

1. Uses of the human eye
2. "Subsystems" comprising the eye
3. Tristimulus theory of color perception
4. Chromaticity
5. Data processing of optical stimuli

L. Lecture on Search Properties of the Eye, by Dr. Gwynn Suits; an elementary discussion of the detection-lobe theory of the human eye:

1. Detection lobe in daylight
   a. Contrast
   b. Angle off foveal axis
   c. Range
2. Binocular-aided vision

M. Lecture on Visual Aids and Stereo-Vision, by Dr. Gwynn Suits; a brief discussion of aids which improve the vision of the human eye:

1. Remote-sensing aids
   a. Telescopes
   b. Binoculars
2. Close-up examination aids
   a. Magnifiers
   b. Microscopes
   c. Stereoscopes
3. Brightness limit

4. Magnification
5. Stereo vision

N. Detection Processes and Detectors ("NASA Houston Course"), by Thomas Limperis; a detailed discussion and analysis of the quantum-detection process:

1. Photoelectric effect
2. Wave-particle duality of electron
3. Atomic models
4. Energy levels
5. Quantum states and quantum numbers
6. Crystalline solids
7. Lithiom atom
8. Ideal photon detector
9. Infrared detectors

O. Lecture on Atmospheric Phenomena, by David Anding and Richard Kauth, presented by William L. Brown; a detailed analysis and discussion on the effects of atmospheric phenomena on radiation with data:

1. Atmospheric absorption, scattering, and transmission
2. Model atmospheres
3. Scattering theories
   a. Mie
   b. Rayleigh
4. Earth/atmospheric radiation
5. Downlooking spectral radiance

A detailed bibliography is appended for further study.

II. Radiometry and Spectrometry

A. Radiometry for Remote Sensing, by Dr. George Zissis; an elementary discussion of the field of radiometry with some analysis related to responsivity of radiometric systems:

1. Remote sensing symposia
2. Radiometric terms
   a. Performance
   b. Calibration and measurement
   c. Reference positions
3. Components
4. Responsivity
5. Selected papers

B. Sources of Radiant Energy, by Anthony LaRocca; an intermediate discussion of sources used to calibrate radiometric instruments:

1. Kirchoff's law
2. Isotherman enclosures
3. Blackbodies and greybodies
4. Blackbody theory
5. Real sources
6. Reference sources
7. Standards

C. Methods of Calibration ("Lecture No. 7"), by Dr. George Zissis; a comprehensive introduction to the distant and near, point and extended, source methods of calibrating a radiometer:

1. Categories of calibration types
2. Distant-point source case
3. Distant-extended source case
4. Near-extended source case
5. Near-point source case

D. Spectrometers ("Lecture No. 15"), by Dr. George Zissis; a comprehensive survey and analysis of instruments used to determine the wavelength distribution of radiation:

1. Prism spectrometers
   a. Monochromatic case
   b. Dispersion
   c. Resolving power

2. Grating spectrometers
   a. Two, three, and six equal slits
   b. n-slits (grating), with n very large
   c. Dispersion
   d. Resolving power
3. Types of prisms and prism spectrometers
4. Types of gratings and grating mounts
   a. Plane and concave gratings
   b. Blazed gratings

An appendix is included which discusses the very popular Ebert grating spectrometer, and with some consideration of the Littrow mount. A large number of problems with solutions is included.

E. Field Equipment Price List

III. Photography, Television and Photogrammetry (James Cooper, Chairman)

A. Introduction to Photogrammetry, by Paul E. Norman (NASA, MSC); detailed introduction to "The Art or Science of Obtaining Measurements from Photographs." An outline is included with the notes, which are written in both English and Spanish. The figures are packaged separately, so they can be used with either version.

B. Fundamentals of Television, by Rowland H. McLaughlin; an elementary discussion of the principles underlying a television system used as a remote sensing device:

1. Purpose of TV system
2. Characteristics of human vision
3. Visual representation
4. Picture elements
5. Image repetition and flicker
6. Image transmission and reception
7. Image analysis
   a. Aspect ratio
   b. Geometric form
   c. Resolution
      i. Vertical
      ii. Horizontal
8. Uniform linear scanning
9. Resolution vs. bandwidth
10. Color television
    a. Primary colors
    b. Encoding and transmission
    c. Compatibility

A set of exercises is included.

C. Magnetic Recording, by Rowland H. McLaughlin; an intermediate introduction to the fundamentals of magnetic tape recording, which is used for storing data collected by a remote sensing system:

1. Magnetism
2. Poles and magnetic field strength
3. Magnetic induction
4. Lines of flux
5. Electromagnetic induction
6. Permeability

7. Magnetization
8. Hysteresis
9. Magnetomotive force and reluctance
10. Demagnetization
11. Magnetic recording
12. Direct recording
13. D-C biasing
14. A-C biasing
15. Recording system components
    a. Tape recorder head
    b. Tape
    c. Tape transport
16. Data storage on magnetic tape
    a. Direct recording
    b. Avoid recording
    c. FM recording
    d. Video recording
    e. Others

A data sheet is included describing the Ampex FR-1400 instrumentation recorder.

D. A Review of Photography of the Earth from Sounding Rockets and Terrain Photography on the Gemini IV Mission: Preliminary Report, both by Dr. Paul D. Lowman (NASA, Goddard); Technical Notes.

E. Basic Film Processes, by James B. Cooper; an outline for lectures on the photographic film processes.

F. Aerial Photography, by Don Groening (Wright-Patterson AFB); an outline covering techniques and components for aerial camera systems, with data sheets on components.

G. Photointerpretation, R. Colwell,(University of California); an oral and pictorial exposition on the fundamentals of photointerpretation with illustrations, applications, and personal experience.

IV. Scanners, Target Signatures, and Ground Measurements (Phil Hasell, Chairman)

A. Introduction to Line Scanners, by Dr. Gwynn Suits; an intermediate discussion of the principles of line-scanning:

1. Introduction
   a. Image plane scanner
   b. Object plane scanner
2. Scanning modes
3. Comparison: scanners vs. photographic cameras
4. Principles of operation
   a. Scanning
   b. Signal-to-noise ratio
   c. Time-constants
5. Thermal imagery

B. Classifications of Scanning Modes (Block Diagrams: "Optical Sensing Techniques"), by William L. Brown; presentation of a system for classifying scanning devices according to platform motion, optical-mechanical scanning modes, and detector geometry.

C. Description of Scanner Components and Their Functions, by Philip Hasell; a detailed outline with helpful block diagrams and figures.

D. Scanner Applications Related to Performance, by Philip Hasell; discusses limitations to performance imposed by the constraints of the scanner applications such as:

1. Ground-based
2. Airborne
3. Spaceborne
4. Single or dual channel
5. Multiple channel

E. Scanner Design Parameters and their Interrelation and Tradeoff Design Exercise ("Lecture Outlines" and "Derivation of the Parametric Equations") by John G. N. Braithwaite; a detailed discussion and analysis of factors entering into the performance of a scanning instrument:

1. Schematic of an airborne system
2. Scanner geometry
3. Multispectral dispersive scanner
4. Performance calculations
   a. Background-limited performance
   b. Shot-noise-limited detectors

5. Trade-offs
6. Spectral resolution

F. Design Features of Scanner Optical Systems and Scanner Wavelength Band Selection, by John G. N. Braithwaite; a detailed discussion based upon the Handbook of Military Infrared Technology, Chapters 4 through 10. Outline is included in "Lecture Outlines" in notes, along with outline for E above.

G. Lecture on Infrared Detector Cooling, by William L. Brown; an elementary-to-intermediate discussion of the different facets of the subject:

1. Historical development
2. Need for cooling IR detectors
3. Types of coolers
4. Detector cell parameters which vary with temperature
   a. Band gap
   b. Absorption
   c. Refractive index
   d. Detectivity

A wide selection of charts, tables and data sheets is included to assist one in carrying out a design task related to cooling for remote-sensing applications.

H. Scanner Signal Handling, Stabilization and Magnetic Tape Recording, Calibration, and Problem Areas with Scanners, by Leo Larsen; outlines for lectures on these subjects, followed by a 100 + page collection of relevant data. The following is a detailed table of contents, with page numbers on this collection:

30-33 Mercury-doped germanium detector

34-36a Mercury-cadmium-telluride detector

37-38 Thermoelectrically cooled IR detector

39-67 Operational amplifiers

68-69 Field effect transistors

70, 71 Thin-film preamplifier

72-90 Gyros and stable platforms

73-100 Tape recorders

101-113 Calibration of imagery

114-118 Flight data records

119-120 Electronic distortions

I. Data Processing and Analysis, by Philip Hasell; a general discussion of the handling of data from airborne scanners:

1. Signal processing
2. Single-band imagery
3. Multiband imagery or data
   a. Addition and subtraction of bands
   b. Two-band discrimination
   c. Multiband discrimination
   d. Statistical analysis
   e. Spectral signatures
4. Imagery interpretation and analysis
   a. Single-band imagery
   b. Multiband processed imagery
   c. Displays

J. Mathematical Derivation of the Temperature Predictive Model, by Dr. Dwight Bornemeier (lecture delivered by William L. Brown); a detailed, advanced study of heat transfer processes related to remote sensing of environment. The lecture which was actually presented gave only the elementary facets of this paper:

1. Heat conduction
2. Heat capacity
3. Diffusion equation for heat
4. Convective heat transfer
5. Radiative heat transfer

A "Thermal Modeling Problem" is included to illustrate the principles involved.

K. Discussion of Specific Spectral Characteristics, by Fabian C. Polcyn (this and the next three lectures are represented by outlines and appended informational sheets); an introduction to spectral characteristics of terrain features:

1. Spectral properties of plants
2. Reflection of radiant energy from soils
3. Effects of altitude on imagery
4. Additional spectral curves

L. Effects of Illumination Changes, Angle Dependencies, and Thermal Contrasts by Fabian C. Polcyn; a detailed look at factors which cause scenes to appear different at different times and from different vantage points:

1. Diffuse and specular reflection
2. Aspect angle effects
3. Solar angle and cloud cover
4. Diurnal variations
5. Thermal modeling (labelled "I, Introduction")

M. Relationship of Sensor Observables to Inferred Parameters, by Fabian C. Polcyn; an introduction to the interpretation of infrared imagery:

1. Sensor-user link
2. Calibration
3. Inference relationships
4. Examples
   a. Diseased plants
   b. TVA water-discharge

N. Examples of Potential Applications of Remote Sensing, by Fabian C. Polcyn; user-related presentation of remote sensing in oceanography, agriculture, forestry and hydrology. A specific application is the detection of doubtful shoals.

O. Remote Sensing of Electromagnetic Radiation for Earth Resources Applications (Ultraviolet), by William R. Hemphill; an outline and detailed paper on the application of ultraviolet reflectance and stimulated luminescence to the remote detection of natural materials. Included is a data sheet on a nitrogen-pulsed gas laser

$$(\lambda = 3371 \pm \tfrac{1}{2}\,\text{Å})$$

P. Ground Measurements, by Robert Horvath; a comprehensive outline of the equipment needed to measure and record "ground truth" data for remote sensing programs:

1. The need for ground truth
2. Sources of ground truth data
3. Instrumentation
   a. Characteristics to be measured
      i. Target
      ii. Environment
   b. Calibration
   c. Data recording and storage
4. Analysis and correlation of data. Additional sample mission plans and data sheets are appended.

Q. Infrared Scanners Used on Meteorological Satellites, by Dr. William Nordberg (NASA, Goddard); a detailed outline of the Nimbus Meteorological Satellite System, plus a paper on "Inference of Properties of the Earth from Satellite Measurements of Infrared Emission."

1. Problem areas
2. Information gained

V. Radar, Lasers and Passive Microwaves (James M. Wolf, Chairman)

(These notes constitute a comprehensive basic introduction to the subject.)

A. Radar and Passive Microwaves, by James M. Wolf; topics discussed are:

1. The electromagnetic spectrum
2. Common names of radar frequencies
3. Microwave radiometer
   a. Resolution
   b. Antenna
   c. Receiver sensitivity
   d. Noise figure
   e. Receiver types
      i. Crystal-video
      ii. Crystal mixer superheterodyne
      iii. Maser
      iv. Parametric amplifier
      v. Tunnel diode
   f. Noise voltage
   g. Calibration
   h. Search rate
   i. Atmospheric attenuation

B. Radar, by James M. Wolf; topics include:

1. Essential elements
2. Historical development
3. Typical radar
4. Plan position indicator (PPI)
5. Advantages and disadvantages
6. The radar equation
7. Transmitters and receivers
8. Bandwidth
9. Antenna gain
10. Radar cross section
11. Corner reflectors
12. Other targets
    a. Metal sphere
    b. Dipole

c. Cylinder

d. Multiple objects

13. Shadowing

14. Polarization

C. Propagation of Short Radio Waves, by James M. Wolf; topics covered:

1. Propagation constant

2. Attenuation

3. Penetration

4. Atmospheric effects

5. Earth curvature

6. False images

a. Mirages

b. Ducting

c. Echoes

i. Ghosts

ii. Angles

iii. Second-go-round echoes

7. Improvement of radar data

8. Chirp radar

9. Angular resolution

10. Synthetic resolution

11. Moving target indication (MTI)

a. Doppler radar

b. "Butterfly" radar

c. Automatic MTI

D. Radar System Design, by James M Wolf; covers:

1. Choice of wavelength

2. Transmitter power

3. Transmitter modulation

4. Examples

E. Antennas, by James M. Wolf; topics include:

1. Beam types

a. Pencil

b. Fan

c. Cylinder

2. "Cosecant squaring"
3. Antenna rotation
4. Tolerances

F. Mapping System Parameters, by James M. Wolf; topics discussed are:

1. System design
2. Recording
3. Aircraft crab error correction
4. Measurement systems
5. Radar image distortions

G. Reflections from the Earth Surface, by Dr. Gwynn Suits; an intermediate discussion of microwave reflection from the earth and power penetration into the earth:

1. Interaction of EM waves with matter
2. Dielectrics, conductors
3. Brewster's angle

H. Lecture on Lasers, by Dr. Gwynn Suits; a brief introduction to lasers with data:

1. Physical principles
2. Principal appropriate uses of lasers
3. Data

I. Lecture on Wave Front Reconstruction, by Dr. Gwynn Suits; a brief introduction to the field of holography, which may some day have application to remote sensing:

1. Huygen's principle
2. Gratings
3. Holography

VI. Data Processing (Donald S. Lowe, Chairman)

A. Lecture on Review of Probability, by Dr. Gwynn Suits; an elementary review of the concepts of probability theory, which must be considered in any useful data processing program.

1. Mutually exclusive events
2. Wholly exhaustive events
3. Expectation ratio
4. Probability
   a. Joint probability
   b. Probability density
5. Likelihood ratio
6. Random digits (table)

B. Discrimination Lecture, by Dr. Gwynn Suits; an elementary-to-intermediate discussion of the science of sorting out items into proper categories:

1. Definition
2. Criteria
3. One- and two-attribute systems

C. Comparative Photogrammetry, by Donald S. Lowe; a brief outline of the factors which contribute to the difficulty of obtaining information from imagery:

1. Cameras
2. Optical-Mechanical Scanners
3. Radar

D. Data Storage and Retrieval, by Donald S. Lowe; a brief outline of techniques of cataloging, filing, and retrieving information from airborne imager:

1. Storage media
   a. Film
   b. Paper print
   c. Magnetic tape
2. Data correlation
3. Ground truth
4. Information to be filed

A flow chart is included as an illustration.

E. Spectral Discrimination, by Richard R. Legault; a very comprehensive advanced review of the techniques for discriminating on the basis of spectral information:

1. Implementation hardware
2. Spectral filtering
   a. Optimum contrast filter
   b. Example
3. Discrimination processor

A second paper is included which is less advanced and more specific; an inexperienced reader should perhaps begin with the second paper.

F. Contrast Ratio Study, by William L. Brown; an elementary discussion of a program which was used for specifying optimum film-filter combinations for certain discrimination situations.

1. "Normalized reflectance difference"
2. "Contrast ratio" program
3. Definition of "contrast"
4. Examples
5. Problem areas

G. Multispectral Facilities

H. Facilities, by Donald S. Lowe; a compilation of facts and figures on facilities for a complete remote sensing program, with approximate cost figures.

Reference List

References which were supplied for use in conducting the study program are listed below. Those reference numbers which are marked with an asterisk are references which would be important for a reader of the second draft notes to have to assist in a self-study program.

1.* Proceedings of the Symposia on Remote Sensing of Environment, I, II, III, and IV.

2.* Manual of Photographic Interpretation

3.* Manual of Photogrammetry, Vols. I and II, 3rd edition.

4.* Manual of Color Aerial Photography

5.* Handbook of Military Infrared Technology

6. Infrared Radiation, Simon.

7. Radar System Engineering, Ridenour.

8.* Propagation of Short Radio Waves, Kerr.

9. Magnetic Tape Recording, NASA sp-5038 (1966).

10. Selected Papers from Symposia on Remote Sensing, (ASP).

11.* A Review of Photography of the Earth from Sounding Rockets and Satellite, NASA TN D-1868 (1964).

12.* Terrain Photography on the Gemini IV Mission: Preliminary Report, NASA TN D-3982 (1967).

13.* Geologic Applications of Orbital Photography, NASA TN D-4155 (1967).

14.* Infrared and Ultraviolet Photography (M-3), Eastman Kodak Data Book.

15.* Aerial Photography, Eastman Kodak Data Book.

16. Kodak Aero-Neg. Color System, Eastman Kodak Data Book.

17. Aerial Reconnaissance Systems and Support Equipment, Hycon, (1-68-2.5C-V).

18.* Experimental Spectroscopy, Sawyer (paperback).

19.* Fundamental of Optics, Jenkins and White.

In addition to these references, a few appropriate excerpts were supplied as needed. The G. E. Radiation Slide Rule was also supplied and is invaluable for any calculation in remote sensing where blackbody radiation is important.

Editors for this second draft were G. H. Suits, chief editor, J. J. Cook, Newbern Smith, and Irvin Sattinger. New composition was supplied by the editors and some of the original lecturers.

CONTENTS

SECTION III: PHOTOGRAPHY

SECTION I

FUNDAMENTALS AND PHYSICAL PRINCIPLES

1. INTRODUCTION TO REMOTE SENSING

1.1. DEFINITION, SCOPE, AND METHODOLOGY OF REMOTE SENSING

In the past several years, interest has grown rapidly in a field which has been named "Remote Sensing of the Environment."* The term "Remote Sensing" is taken to mean the acquisition of information about objects or phenomena in our environment through the use of sensory devices at positions remote from the subject under investigation. The information is derived from data obtained by means such as the measurement of electromagnetic radiation by air-or satellite-borne cameras, television, ultraviolet and infrared detection apparatus, radar and radio frequency receivers; the measurement of acoustical energy by seismograph, sonars and microphones; the measurement of nuclear radiations; and the measurement of force fields by gravimeters and magnetometers.

Two kinds of information are supplied by remote-sensing devices: basic scientific information providing knowledge about the earth and its environment, and applied scientific information useful for the exploitation or conservation of natural resources, the solution of engineering problems, and the promotion of national defense.

Photography is a classic and well-developed example of remote sensing. However, many environmental phenomena do not produce effects which can be recorded by aerial cameras or which can be analyzed by photogrammetric procedures. Many new types of sensing devices have been developed which can receive and record important signals in the non-visual portion of the electromagnetic spectrum. The analysis of these signals, possibly in conjunction with photographic records, can provide information which otherwise is impracticable to obtain.

The term "Remote Sensing" is taken to include the experimental plan and choice of sensors, the reception, recording, and processing of the signal data from these sensors, and, equally important, the analysis of the resultant data so that the observer can recognize the significance of his observations.

The basic logic of remote sensing is the logic of inference. If cause C_1 exists, then effect E_1 will be observed, and if effect E_1 is observed, then cause C_1 must exist. Research in remote sensing is required to establish useful and unique inferential relationships of this kind from new

* Remote Sensing was formally recognized in January 1961 by a subcommittee of the National Academy of Sciences-National Research Council and representatives from the Office of Naval Research, Geography Branch, who met to discuss the needs for more advanced and efficient data acquisition techniques in the earth sciences. It is the title of contract NONR 1224 (44) jointly sponsored by ONR, AFCRL, and ARO under which The University of Michigan initiated a series of symposia.

and unfamiliar observed data obtained with new and unfamiliar sensing systems. Remote sensing is of little value if only observations of various effects are accomplished without knowledge of the useful inferential relationships which can be used to achieve economic or scientific objectives.

The construction and use of inferential relationships using environmental data received remotely by normal human eyesight is so common an occurrence that one is usually not aware that the inferential process has taken place. When a magician demonstrates that what is seen remotely from the audience and what can be verified by other tests are not necessarily consistent, then the fact that an inferential process has occurred in the minds of the audience becomes quite evident. Similarly, an Indian tracker and a city bred man can view the same scene and receive by remote observation the same physical signals but the Indian tracker can make useful inferences from the data concerning the presence of unseen game while the city man is unable to make such inferences. The famous fiction stories of Sherlock Holmes illustrate (at least in principle) the construction and use of inferential relations between present data and past occurrences in the same way as the Indian tracker's way of seeking game. Poor old Watson represents the "city man" who can not make those inferences from the same data. The object of remote sensing as defined here is to extend the capacity of specialists to make in a similar manner useful inferential connections from the vastly greater amounts and kinds of data that can be made available by use of special instruments.

In order that the useful inferences may be made, it is necessary to know the salient properties of the sensing devices and the physical factors which give rise to the signals.

The state of the art of sensors, sensor-carrying platforms, and data processing tools is now such that vast quantities of a variety of data about our environment can be acquired at costs that are low relative to the many potential benefits. However, there is a serious lack of that multidisciplinary knowledge and interaction which is essential to the use and exploitation of this enormous remote sensing capacity for more than the rather obvious and direct uses for which these sensing systems were primarily developed.

The engineers and physicists who have developed and initially used the sensing devices understand their operation but are not generally in a position to appreciate their full significance in providing the solution to problems faced by the natural scientist. Many natural scientists are still unaware of what these new sensors can observe and measure, and in some cases are unaware of the existence of the sensors themselves.

The successful exploitation of remote sensing techniques meets a common need of a large number of agencies, private and public, military and civilian. Large-scale search and surveillance problems face the oil and mining industries, agriculturalists, archaeologists, geographers, geologists, oceanographers and the fishing industries, the military services and arms control experts, to name a few.

1.2. COURSE OBJECTIVES

It is assumed that the student using these course noted will be oriented toward natural science, natural resources or conservation problems. He should be able to visualize the manner in which special types of remote sensing information could be adapted for use in his particular field of interest and to evaluate the benefits resulting from the use of this information.

The course notes are aimed at providing the information needed by such a person to do the following:

1. Specify a number of important physical quantities pertinent to his field of interest whose remotely sensed attributes might conceivably lead to recognition and location of those quantities by remote sensing means.

2. Be able to show by modeling or analysis an estimate of the expected distinction between remotely sensed attributes of his target quantities and possible interfering background quantities.

3. Be able to cite or locate in the literature experiments comparable to those he needs to perform.

4. Specify an experimental plan, including instrumentation, calibration, and type and accuracy of data required to test the model or analysis.

5. Specify the data format and the method of processing the data.

6. Estimate the probable success and cost of an operational remote sensing system to locate and recognize the target quantities.

1.3. PHYSICAL PRINCIPLES

Because of the emphasis in these notes on the use of underlying physical principles for application of remote sensing systems, the first section is devoted to a presentation of those fundamental principles of radiation theory, optics, atomic physics, electronics, and thermodynamics which particularly apply to the remote sensing field.

Many treatments of the fundamental principles of optics apply directly to the design of instruments. These treatments are found in standard texts on optics, such as Fundamentals of Optics by Jenkins and White, for Physicists and Engineers. The presentation here is intended to give the natural scientist sufficient knowledge of wave motion and of the interaction of radiation with matter so that the common features of microwave, infrared, visible, and ultraviolet devices can be seen, the implications of device specifications can be understood, and the basis for constructing inferential relations between device signal output and the physical causative factors can be visualized. In addition, a general knowledge of the characteristics of the human eye is of importance for remote sensing applications and is treated in a later section.

The study of microwave systems, such as radar and passive microwave devices, requires some specialized knowledge of the elements of microwave theory and practice, which is covered in the Section on microwaves.

1.4. TYPES AND CAPABILITIES OF SENSORS

Following the presentation of fundamental principles, each type of remote sensor is discussed. Its theory of operation and major features are described. The form and content of the output data produced by the sensor are explained, with particular emphasis on the relationships between the data and the targets and background being observed. Typical uses for the sensor are then presented as examples of the application of remote sensing principles.

Remote sensing devices may be differentiated with respect to the type of physical phenomenon sensed. Gravimeters and magnetometers are able to sense static force fields (e.g., the earth's gravitational or magnetic field). Other devices sense dynamic fields, such as acoustic or seismic vibrations or electromagnetic radiation.

Devices which sense electromagnetic radiation may be further characterized in terms of the region of the electromagnetic spectrum in which each operates (ranging from gamma rays through radio waves). Some devices operate over a sizable spectral band (e.g., from ultraviolet to infrared), and may simultaneously record many channels of data at different wavelengths. Individual types of sensors may be active (providing their own source of illumination), semi-active (depending on illumination from a separate element of the system), or passive (depending on illumination from natural sources, such as solar radiation, or on thermal emission). It is also necessary to distinguish between non-imaging devices, such as radiometers and spectrometers, and imaging devices, such as cameras and radars.

Information can be derived from several different characteristics inherent in the electromagnetic radiation. The total power of the received signal, the distribution of this power through the spectral regions sensed by the instrument, polarization characteristics of the returned radiation, and variation of the return from a single object with time of day, season, weather conditions, etc. all contribute information which can be used to identify the object sensed and determine its physical or biological condition.

1.4.1. NON-IMAGING ELECTROMAGNETIC SENSORS.

Non-imaging sensors which respond to electromagnetic radiation generally serve the purpose of instrumentation devices capable of measuring specific physical quantities associated with patches of the earth's surface or paths through the atmosphere falling within the instantaneous field of view of the instrument. The data rates of non-imaging devices are low compared with those of imaging devices, but high accuracy is usually obtained by the use of appropriate instrument design and calibration procedures.

Individual types of instruments fall into several general classes. Radiometric instruments measure total power radiated in specified spectral bands in the direction of the instrument from a patch of earth or atmosphere whose dimensions can be accurately determined. Spectrometers are used to determine the distribution of received energy through the spectral regions of interest. In the microwave region, the radar scatterometer is a specialized device which measures power returned from the earth's surface as an indication of the reflective or electrical characteristics of the surface. Other types of instruments are under development; for example, devices using lower radio frequencies to obtain greater penetration of the ground and devices based on principles of absorption spectroscopy.

1.4.2. IMAGING ELECTROMAGNETIC SENSORS. Remote-sensing systems for acquisition of pictorial data include the following basic imaging techniques: photography, for the region from the ultraviolet at 300 mμ to the near infrared at 1μ: television, for use in the same region; optical-mechanical scanning for the infrared region between 1 and 40μ (and, for some of the work, between 0.3 and 1μ); and passive microwave or radar for discrete regions between 1 mm and several centimeters or meters in wavelength. The state of the art is such that all three regions can be adequately instrumented.

Photographic imagery yields data concerning the amount of solar energy reflected from selected objects on the earth and its cloud cover as a function of wavelength, in selected narrow spectral regions. Since World War I, aerial photography using the visible portion of the spectrum has been applied to a steadily increasing variety of tasks. At the close of World War II, with the release of many trained and experienced interpreters as well as a great quantity of applicable equipment, the use of aerial photography suddenly accelerated in quantity and scope. In addition to its classic military uses, it has been found virtually indispensable in political, economic, and scientific applications. In reference 1 extensive application is indicated in geology, soils mapping, wildlife, range and watershed management, agriculture, urban analysis and planning, archaeology, geography, etc.

In spite of extensive and increasing utilization of aerial photography, many of its techniques remain relatively modest extensions of the capabilities of the human eye. Numerous methods of analysis and inference have been developed and the use of high-speed aircraft has become common, yet the range of practical usefulness for aerial photography remains more restricted than is desirable. The principal restrictions of such systems are (1) the inability to sense some very important physical parameters, such as temperature, (2) the relative slowness of interpretation as compared to the rate at which information is desired and can be obtained, and (3) the costs of acquiring imagery over large areas.

Further study may prove that electro-optical or optical-mechanical scanning techniques should be employed in this spectral region in order to acquire data of optimum usefulness. These methods, since they produce data in the form of an electrical video output, permit special processing before final recording on photographic film or magnetic tape; thus, the information from one spectral region may be combined additively with that from another region, for example, and a single picture produced which may represent an optimum image for certain studies.

Television systems generally cover the same spectral regions as photographic cameras, but have relatively low resolution for a given size of projected image. They may be designed to operate at very low light levels, a capability which may have special value for certain types of applications. The television camera differs basically from the photographic camera in that the image may be directly viewed on a television viewing screen. Since the data output is in electrical form, the imagery can be transmitted by data link or can be recorded either locally or at a distance on video tape or photographic film. The use of electrical data output and magnetic tape recording permits the use of signal adaptation or measurement methods common to other types of scanning sensors.

Infrared imaging devices will produce recordings of the thermal structure and behavior of the terrestrial and meteorological environment. Experience has shown that terrestrial data in at least two spectral regions (e.g., 4.5-5.5μ and 8.5-13.5μ) are often much more useful than either one alone. In particular, at the longer wavelengths the contrast of objects seems to be dependent on emissivity differences in the objects; at shorter wavelengths contrast seems more dependent on temperature differences of the objects. The surface condition of the object often affects the relative emittance of the object. Since measurement is of radiation emitted from the objects, the sensing system can be used day and night. Other wavelengths should be used for studies of clouds, wind, and air pollution (wavelengths corresponding to absorption regions in the atmosphere).

Optical-mechanical scanning devices and other techniques now available are capable of producing imagery over a very wide spectral range which includes ultraviolet, visible, long-and-short infrared, and microwave wavelengths. Imagery made in spectral bands heretofore little used and comparison of simultaneous imagery made in a variety of spectral bands promise to increase the utility and applicability of pictorial sensing in two ways. First, new types of information not obtainable with conventional photography can be provided. For instance, the use of infrared can furnish indications of the operation of man-made power sources and can produce thermal maps. Furthermore, the ability to sense data in the ultraviolet, visible, infrared, and microwave bands may make it possible to find a spectral region in which discrimination may be accomplished between objects which exhibit no discernible differences in the

"photographic" region. Second, a comparison of imagery from a wide variety of spectral regions may permit increased use of pictorial tone differences and lessen the need for strong reliance on the use of the fine details of shape information, one of the principal reasons for the slowness of the normal interpretive processes.

Radar imagery provides a comparative measure of the reflection from various components of the earth. Reflected power (radar return) is affected by the aspect of the terrain relative to the beam direction, by the dielectric properties (at the radar frequency) of the material, and, for elements smaller than the resolution limit, by element size. Scanning at small angles (near grazing incidence) yields reflected power variations which are a sensitive function of the local slope of complex landforms; it is therefore a singularly powerful potential technique for topographic mapping. Extensive research may be required to determine optimum techniques, frequencies, scan angles, and power requirements.

Passive microwave scanning radiometers have not been developed as extensively as the other sensors under consideration here, but new techniques for improving thermal sensitivity and speed of operation are available. Although comparatively few informative pictures have been produced, polarization effects and the fact that radiation in the passive microwave region originates from beneath the surface of certain terrestrial materials (e.g., ice and snow) offer promising possibilities for geologic and arctic exploration. Passive microwave signals can also provide information about the roughness or other characteristics of land or water surfaces, which would supplement data obtained from other spectral regions.

Both radar and passive-microwave data-collection techniques have the advantage that they can be used both day and night and in the presence of cloud cover. This greatly extends their capability for obtaining continuous information concerning the earth's surface, regardless of time of day or weather conditions. It may also supply data on meteorological conditions in areas of cloud cover that are not obtainable in other spectral regions.

1.5. METHODS OF INTERPRETATION AND ANALYSIS

Interpretation of conventional aerial photography is already a well-developed art. Skilled photo-interpreters can detect and identify man-made and natural features of the earth's surface by means of shape, color, texture, pattern, and shadow. Objects can also be detected by indirect means, for example, by inferring their presence from the detection of other objects with which they are normally associated. Photographic imagery can also be used to map the spatial relationships of cultural and natural features. This can be done very accurately by the methods of photogrammetry, using specialized equipment for measuring horizontal and vertical distances, and for preparing contour and profile data from the photographs.

These methods, however, constitute only a small part of the possibilities for extracting information from remote sensing records, and much of the present research in this field is directed toward conceiving and developing new ways of enhancing our ability to analyze and interpret remote sensing data and imagery. Many of these techniques are based on the ability to make inferences about the nature and extent of remotely sensed objects from a thorough understanding of the physical principles involved. In particular, a knowledge of the reflective, emissive, thermal or dielectric characteristics of various types of materials makes it possible to detect and identify such materials and to make quantitative measurements on them. In order to be able to extend the capabilities of remote sensing in this manner, it is important for the experimenter to be able to adapt the underlying physical principles to the design of his experiment, the design or selection of the sensing equipment and recording method, and the processing and analysis of the resulting data. Following sections of these notes will provide detailed information on ways of achieving these results.

Other present day research is concerned with finding automatic methods of interpretation and analysis which can be substituted for the slow and costly methods of photointerpretation by human beings. One example of such automatic methods is the use of multi-channel sensors to record a scene simultaneously in a number of spectral regions and the processing of the recorded data by spectrum-matching equipment which can identify specific varieties of vegetation, soil or rocks on the basis of their spectral signatures. Other methods of automatic interpretation are based on the use of pattern recognition techniques for detection and identification. Recent work on such automatic methods is described in later sections of these notes.

2. PHYSICAL PRINCIPLES

2.1. THE CONCEPT OF FORCE FIELDS

The central object of remote sensing is to use what is called a force field to establish the nature and position of remotely located conditions of interest to the natural scientist.

2.1.1. STATIC FORCE FIELDS. If an excess electric charge is captured on a piece of glass, one will find that matter in the neighborhood of the glass may experience a force. If a small test body containing a small excess of electric charge is held near the glass, a force of attraction or repulsion on the test body may be observed. The magnitude and direction of the force on the test body will depend upon the spatial position of the test body relative to the glass. Each place in space could be assigned a force magnitude and direction by such a procedure. The visualization of the ensemble of all such force measurements around such a piece of glass constitutes the concept of an electric force field. The cause of the force field is attributed to the excess charge on the glass. However, an increase in test charge will result in a proportional increase in force field. In order to make the concept of the electric force field independent of the magnitude of the test charge, one visualizes that each force magnitude is divided by the amount of charge contained in the test body. In that case the electric force field should be independent of the particular test body which is used. Hence, the electric field is the visualized ensemble of force magnitude and directions per unit of test body charge in all spatial positions. The concept of the magnetic field is similar except that a test magnet is to be used instead of a test charge. The gravitational force field is correspondingly the force per unit of test mass.

The use of the earth's static electric field has not been used much for remote sensing purposes. The static magnetic and gravitational fields, however, are frequently used for exploring the subsurface structure of the earth.

It is fairly easy to show how the gravitational field is altered near the surface of the earth by a material below the surface having a different density. Newton's law of gravitational attraction states that two mass particles attract each other with a force proportional to the product of their masses and inversely proportional to the square of the distance between them:

$$F = \frac{GM_1M_2}{r_{12}^2}$$

where F is the force of attraction, G is the proportionally constant, M_1 and M_2 are the masses of the particles, r_{12} is the distance between M_1 and M_2.

It is easy to show that the force on a small test particle, M_2, due to a large homogeneous spherically symmetrical earth is given by the same formula as long as the mass, M_2, is outside the surface of the earth. The force on M_2 is directed toward the center of the earth as if all of the mass of the earth were concentrated at the center. This result is a consequence of the contribution of each small particle of earth adding together in accordance with Newton's Law. Now consider an earth which is spherically symmetrical except for a small void near the surface. Assume that the small void should contain a mass, M_3, in order to make a spherically symmetric earth. One can then calculate the effect of this void by subtracting the gravitational force of M_3 from the force due to a spherically symmetric earth as shown in figure 2-1.

$$F_{net} = \frac{GM_1M_2}{r_{12}^2} - \frac{GM_3M_2}{r_{23}^2}$$

Hence, the net force is less than that for a solid earth for two reasons. 1. the total mass is reduced and 2. the distance to the void is much shorter than the distance to the center of the earth. The second reason is much more important than the first because the distance to the void from the test mass may be only a fraction of a kilometer while the distance to the center of the earth is many thousands of kilometers. A "gravitational low" will be measured over such a void using a test mass close to the void. One may treat an inclusion of different density than that required by a spherically symmetrical earth in a similar manner. If a small region which would have a mass, M_3, for spherical symmetry has, instead, a mass M_3', then the gravitational force may be calculated by subtracting the effect of a void and adding back in the effect of M_3' at that same position. Thus,

$$F_{net} = \frac{GM_1M_2}{r_{12}^2} - \frac{GM_1M_3}{r_{23}^2} + \frac{GM_2M_3'}{r_{23}^2}$$

$$\text{or } F_{net} = \frac{GM_1M_2}{r_{12}^2} + \frac{GM_2(M_3'-M_3)}{r_{23}^2}.$$

When $M_3' > M_3$, then a "gravitational high" will be found near that region. The force per unit test mass, g, is then

$$g = \frac{F_{net}}{M_2} = \frac{GM_1}{r_{12}^2} + \frac{G(M_3'-M_3)}{r_{23}^2}$$

A very sensitive measure of g can be made by the use of a pendulum. The period of a simple pendulum is

$$T = 2\pi\sqrt{\frac{\ell}{g}}$$

where ℓ is the length of the pendulum wire

so a careful measure of the change in pendulum period at different regions near the earth's surface will reveal changes in the value of g.

The measurement of changes in magnetic field due to local changes in magnetic anomalies are classical tools for geological exploration and will not be treated further here.

The static fields must be mapped by moving the test object from point to point throughout a region close to the surface of the earth. At greater distances from the surface, the static field anomalies not only decrease sharply in magnitude but also become more diffuse in positional definition. There appears to be no way to "focus" an image of static field anomaly by remote means the way one does with light.

2.1.2. RADIATED FORCE FIELDS. Abrupt time rates of change of static force fields can induce radiated forms of force fields which produce measurable effects at great distances from the initial disturbances. Electric and magnetic disturbances both produce electromagnetic radiation. Abrupt time changes in gravitational fields presumably produce gravitational waves but no convincing evidence is available to confirm their existence. Sudden changes in pressure or stress in a material medium produce radiated waves of stress through a medium. Sound waves occur as pressure waves through liquids, solids, and gases. Seismic waves denote "sound" waves in the earth although they may be of such frequencies that they cannot be heard. Because these wave disturbances are large at considerable distances and because these waves carry important information concerning the nature of the originating disturbance, the use of wave motion in force fields has become one of the most valuable sources of remotely sensed information. Certain properties of wave motion are common to all radiated force fields so the understanding of wave motion is central to the understanding of the function and utility of a large number of remote sensing devices.

2.1.3. WAVE MOTION. Wave motion is familiar to all when visualized as ripples on the surface of water. Since electromagnetic radiation is also described as wave motion, the common features of wave motion will be briefly reviewed for the purpose of defining the important terms to be used in later lectures.

The simplest types of waves are simple harmonic waves whose amplitude varies as the sine or cosine of position and time. Such waves have a fixed wave length defined as the distance between two crests of the wave at any instant of time, a fixed frequency defined as the inverse of the period of the wave action at a fixed position in space and a velocity of propagation which is defined as the rate at which a particular wave crest moves. As a consequence of these definitions, the following relation holds for simple harmonic waves in homogeneous media

$$\lambda \nu = v$$

where λ is the wave length

ν is the frequency

v is the velocity of propagation.

The time required for a wave to travel a distance D is then proportional to the time of travel.

$$D = vt.$$

The measurement of distance by timing a wave is fundamental to the range measurement properties of radar, sonar, and seismic systems.

Harmonic wave profiles are those which can be described by simple trigonometric functions such as

$$S = A \sin\left(\frac{2\pi x}{\lambda} - 2\pi \nu t\right).$$

The quantity $\left(\frac{2\pi x}{\lambda} - 2\pi \nu t\right)$ is defined as the phase angle of the wave. In wave motion occupying more than one dimension in space, all adjacent points having the same wave phase angle constitute a wave front. The wave moves in a direction normal to the wave front.

A wave which propagates across a boundary between two media such that the velocity of propagation is different in the two may be partially or totally reflected back into the first medium. A standing wave is one which is composed of two oppositely traveling waves having the same frequency. A standing wave can be generated from a single traveling wave with two reflecting boundaries at suitable spacings. The reflected wave from a boundary may have a phase which is different from the phase of the incident wave. If a change in phase occurs, the phase is said to be "shifted" upon reflection.

Waves are classified in accordance with the direction of the wave displacement relative to the direction of propagation. When the displacement occurs in the direction of propagation, the wave is classed as a longitudinal wave. When the displacement is at right angles to the direction of propagation, the wave is classed as a transverse wave. Water waves are transverse waves; sound waves are longitudinal waves.

Electromagnetic waves in free space are purely transverse waves; however, when electromagnetic waves are confined within metallic boundaries, they may be a mixture of longitudinal and transverse. Transverse waves may have any number of directions of displacement in the plane perpendicular to the direction of propagation. Some principal direction is chosen against which to measure the displacement direction or polarization of the wave. The terms "vertical" polarization and "horizontal" polarization refer to some understood vertical or horizontal direction specified by the context of the discussion.

All polarizations of transverse simple harmonic waves can be constructed by combining two waves having the same frequency and direction of propagation but with one vertically polarized and one horizontally polarized and with relative phase angle and amplitudes selected appropriately. Right circular or elliptical and left circular or elliptical polarizations can be constructed.

The changes that can occur to a wave upon reflection from a boundary depend upon the physical properties of the material at the boundary. If the character of the material is such that it reflects wave motion differently for different wave lengths and/or changes the type of polarization, then those material properties can be detected by remote means through the measurement of the incident and reflected waves to and from the material.

A low cost demonstration of wave motion can be made by suspending a rope between two supports. Waves can be generated at one end by plucking the rope. Transverse waves having any polarization can be generated.

The energy required to generate a wave is propagated by the wave motion toward a distant receiver. The average power per unit area carried by a wave in three dimensions past any point and in the direction of propagation is proportional to the time average of the square of the wave displacement at that point. Since it is the power of the wave intercepted by the receiver that is to be used for detection of a wave source, the study of the generation, transmission, and absorption of power in wave motion is of central concern in remote sensing.

2.1.4. SUPERPOSITION OF WAVES. The discussion of wave motion dealt with simple harmonic waves with a distinct wavelength, frequency, and velocity of propagation. A true simple harmonic wave has no beginning or end. Real waves always have a beginning and an end. Real waves are not always sinusoidal in shape.

Fourier showed that any complicated wave form could be constructed by adding together an infinite number of simple harmonic waves with the appropriate amplitudes, wavelengths, frequencies, and phase relations. Plotting the squares of the amplitudes of the simple harmonic waves against the frequency corresponding to that amplitude produces the Fourier power

spectrum of the complicated wave. Each contributing simple harmonic wave is said to be a Fourier Component of the complicated wave form. Certain complicated wave forms are commonly used:

$$S = A(t) \sin \omega t$$

where $\omega = 2\pi\nu$

is called a modulated wave form because the amplitude is not a constant.

It is easy to show that the Fourier power spectrum of

$$S = \cos \omega_1 t \sin \omega_2 t$$

is as shown in figure 2-2.

The spectrum of $S = (1 + \cos \omega_1 t) \sin \omega_2 t$ is shown in figure 2-3.

The trigonometric identity,

$$\sin a \cos b + \cos a \sin b = \sin (a + b)$$
$$\sin a \cos b - \cos a \sin b = \sin (a - b),$$

is all that is needed for this derivation.

Common wave forms are constructed by modulating a simple harmonic wave with a complicated modulation wave form; that is,

$$A(t) = \sum_{n=0}^{\infty} a_n \cos n\omega_1 t + \sum_{n=1}^{\infty} b_n \sin n\omega_1 t$$

Hence, it is an extension of the foregoing to show that the power spectrum will consist of the "carrier" frequency, ω_2, and the sum and difference frequencies called "side bands" caused by A(t). In most cases, the complicated wave forms have Fourier components corresponding to infinite frequencies. But the magnitude of the high frequency components becomes very small so that only a finite upper limit is sufficient to use in practical applications. A number of common wave forms are analyzed in Reference Data for Radio Engineers, Federal Telephone and Radio Corp., 67 Broad St., New York, N. Y., 1943 and later.

Not only is the analysis of propagating wave forms made easier by using Fourier components but also the analysis of image-forming systems is better understood using the Fourier component concept. A densitometer trace over an image transparency can be analyzed into Fourier components.

Sharp edges in density would produce "high frequency" Fourier components. Since a densitometer trace has no reference to time, the quantity which takes the place of frequency in a densitometer trace is referred to as a spatial frequency.

A wave form which is transmitted through a medium or an electrical circuit may undergo alterations in shape. If the change in shape is due to attenuation of some of the Fourier components of the wave, the wave is said to be distorted by "amplitude distortion." If the change in shape of the wave is due to the shifting in phase of some Fourier components relative to the phase of other components, then the wave is said to be distorted by "phase distortion." Amplitude distortion changes the Fourier power spectrum; phase distortion does not. A medium, which has a velocity of propagation that is different for different wave frequencies, will produce phase distortion in a propagating wave. Such a medium is called a dispersive medium.

Any complicated wave form which is periodic in space or time may be constructed by superposition of simple harmonic waves having frequencies which are integral multiples of some lowest frequency called the fundamental frequency.

The next higher frequency is called the second harmonic; the next higher, the third harmonic, etc. In the case of a wave train which is not periodic, the component waves must have all frequencies in the spectrum so that a simple integer relation between frequencies is not sufficient. The expression of a finite length, nonperiodic wave train in Fourier components must be constructed by integration rather than by simple summing of trigonometric functions.

For periodic wave forms

$$S(t) = \sum_{n=0}^{\infty} a_n \cos n\omega t + \sum_{n=1}^{\infty} b_n \sin n\omega t.$$

For nonperiodic wave trains

$$S(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\omega t} f(\omega)\, d\omega,$$

where $e^{-i\omega t} = \cos \omega t - i \sin \omega t$,

$f(\omega)$ is the Fourier-transform or spectral function,

$i^2 = -1$

$e = 2.7183 \ldots .$

The spectral function, $f(\omega)$, is usually a complex function which corresponds to the two sets of Fourier coefficients A_n and B_n in the periodic case. The Fourier power spectrum is found from $f(\omega)$ by multiplying $f(\omega)$ by its complex conjugate, $f^*(\omega)$, to be defined later.

An important quality of the Fourier transform may be seen from the following specific example by choosing small or large values for a .

Suppose $S = e^{-at^2}$

then $f(\omega) = \sqrt{\frac{\pi}{a}} e^{-\omega^2/4a}$

One can see from this that a sharp wave form (large value for a) requires a broad power spectrum while a long, smooth wave form (small value for a) requires a narrow power spectrum. This property applies in general for other wave shapes.

2.1.5. VECTOR REPRESENTATION OF WAVE AMPLITUDE. Superposition of simple harmonic waves having the same frequency can be more easily accomplished using complex variable addition rather than by the use of the trigonometric identities. Let us make a two dimensional graph with the horizontal axis for real numbers and the vertical axis for numbers multiplied by $i = \sqrt{-1}$. Any complex number written as the sum of a real and an imaginary part — such as $a + ib$ can be plotted on this graph as a point. A two dimensional vector can be defined for each point by placing the tail at the origin and the head at the complex number point on figure 2-4. It will be seen that the addition of two such vectors can be accomplished by adding the two real parts together and the two imaginary parts together to provide a new complex number representing the resultant vector.

$$(a_1 + ib_1) + (a_2 + ib_2) = (a_1 + a_2) + i(b_1 + b_2)$$

The square of the length of any vector is easily found by multiplying the complex number representing that vector by its complex conjugate. The complex conjugate of any complex number is that same complex number but with the sign of the symbol, i, reversed. Hence the square of the length of a vector represented by $a + ib$ is

$$(a + ib)(a - ib) = a^2 - iab + iab - i^2b^2$$

or

$$(a + ib)(a - ib) = a^2 - i^2b^2.$$

But $i^2 = -1$ so $a^2 - i^2b^2 = a^2 + b^2$. The angle ϕ which any vector makes with the real number axis is the phase angle of the simple harmonically varying displacement of the wave. Thus, $\phi = \left(\frac{2\pi x}{\lambda} - 2\pi\nu t\right)$ for a simple harmonic plane wave is the angle which the vector representing that wave motion must make with the real axis. The actual wave displacement, S, at any point in time or space will be taken as the real part of the complex number representing that vector. Now let us suppose we have a simple harmonic wave in the X direction with amplitude, A, so that

$$S = A \cos \left(\frac{2\pi x}{\lambda} - 2\pi\nu t\right).$$

The use of a cosine instead of a sine is not critical since the initial phase of a harmonic motion

depends upon the arbitrary choice of starting time and position. The phase angle of the displacement at x = 0 is

$$\phi = -2\pi\nu t$$

As time progresses, the phase angle increases negatively at a steady rate, $-2\pi\nu$. The vector representing the wave motion must rotate clockwise from a position on the real axis. The real component of the vector is simply A cos ϕ. If one chooses to superpose a second simple harmonic wave having the same frequency but with a different phase relative to the first, it is clear that each vector must rotate at the same rate. Thus, the vector sum must also rotate at the same rate. Therefore, the superposition of two simple harmonic waves having the same frequency and direction but not the same phase and amplitude results in another simple harmonic wave having the same frequency and whose amplitude is the vector sum of the two contributors. The equal rotation of the vectors results from an equal change in phase angle. Considering time fixed and looking at the wave displacement as a function of position, one will see the wave motion represented by the vectors rotating counter clockwise for observer motion in the +x direction.

The superposition at a fixed field point of two waves propagating in different directions can also be represented in the same fashion if they have the same frequency. However, the superposition of waves at a fixed field point but having different frequencies require that the vectors representing the wave motion at that point rotate with different speeds. In that case, the resultant vector amplitude changes with time so that simple harmonic wave motion can not be a result.

2.1.6. POWER OF POLYCHROMATIC WAVE MOTION. Suppose that at a field point the wave motion from a large number of sources with different frequencies are superposed. Let the relative phases and amplitudes of the various contributors be randomly chosen. The resultant vector is then

$$A + iB = (a_1 + a_2 + \ldots a_n) + i(b_1 + b_2 + \ldots + b_n)$$

which will be a function of time in some unknown way. The resultant wave motion will not be simple harmonic. Let us calculate the square of the resultant.

$$A^2 + B^2 = \left(a_1^2 + a_2^2 + \ldots + a_n^2 + a_1a_2 + a_1a_3 + \ldots + a_2a_3 + \ldots\right) + \left(b_1^2 + b_2^2 + \ldots + b_n^2 + b_1b_2 + b_1b_3 + \ldots + b_2b_3 + \ldots\right).$$

Now if there are as many positive values of a and b as negative values and as many large values as small values, then the cross product terms will have as many large values as small values

and as many negative values as positive values so that the sums of the cross terms will tend to be close to zero. However, note that the square terms must all be positive so that the sums of the squares can not ever add to zero unless each is zero separately. Taking the time average of both sides of this relation we find that

$$\overline{A^2 + B^2} = \overline{\left(a_1^2 + a_2^2 + \ldots a_n^2\right) + \left(b_1^2 + b_2^2 + \ldots + b_n^2\right)}.$$

or

$$\overline{A^2 + B^2} = \left(\overline{a_1^2 + b_1^2}\right) + \left(\overline{a_2^2 + b_2^2}\right) + \ldots + \left(\overline{a_n^2 + b_n^2}\right).$$

Hence the average power per unit area passing that field point is equal to the sum of the average power per unit area of each component wave motion. This is the general condition for sunlight, thermal radiation, most artificial illuminants, random sounds and seismic disturbances. Such wave motion is said to be noncoherent.

2.1.7. WAVE GEOMETRY. Consider the wave formula,

$$S = A \sin\left(\frac{2\pi x}{\lambda} - 2\pi \nu t\right),$$

but allow this formula to describe wave motion in a three dimensional space. Evidently, the value of the wave displacement, S, is constant along the y and z directions for some value of x position and time. Thus, the isophase points at a given time all lie in planes parallel to the y, z plane. Such a wave is called a simple harmonic plane wave propagating in the x direction.

The wave motion carries an average power per unit of area in the x direction which is proportional to A^2. Since the wave front has an infinite area and a finite power per unit area, the total average power carried by a simple harmonic plane wave would be infinite. Such a plane wave can not exist in reality but is commonly used for analytical purposes and as an approximation to waves which carry finite average power.

Consider a sound source made by a harmonically pulsating rubber sphere surrounded by air. The pressure waves in the air will radiate away from the surface of the sphere with spherical isophase surfaces or wave fronts. If the power carried by the wave is not lost to the air in transit, the wave displacement, S, (change in air pressure in this case) would be described by

$$S = \frac{A}{r} \sin\left(\frac{2\pi r}{\lambda} - 2\pi \nu t\right)$$

where r is the range from the center of the sphere.

If the size of the sphere is considered to be as small as you please, then the source of such waves is said to be a monochromatic point source.

It is possible to show that the waves of a monochromatic point source can be constructed by the superposition of an infinite number of monochromatic plane waves propagating in all directions from the point source with the appropriate phases and amplitudes.

2.1.8. HUYGENS' PRINCIPLE. Huygen discovered that any propagating wave front in an isotropic homogeneous medium could be generated from any preceding wave front by considering each element of area of the preceding wave front as a small point source of waves generating waves in phase with that preceding wave front. This principle is very valuable in estimating the wave propagation behavior around obstacles or through apertures. Take first a small aperture in an otherwise opaque plate. If a plane simple harmonic wave is incident upon that aperture as shown in figure 2-5, the small part of the wave front which may pass through the aperture may be considered the source of hemispherical waves. The contribution of the other portions of the wave front are missing so that instead of a plane wave appearing on the other side of the plate, only the Huygen's wave contribution of a small section of that wave remains.

One can determine the effect of a small opaque obstacle on an incident plane wave by arguing the converse. When a plane wave is incident upon a small opaque obstacle, the obstacle removes the Huygen's wave contribution to the following wave motion. Thus, instead of a simple plane wave appearing on the other side of the obstacle, one should have a plane wave minus the Huygen's wave contribution that was taken out by the obstacle. This negative Huygen's wave is the same kind of a wave but with opposite phase and corresponds to a redirection of a small part of the power carried by the plane wave. It is clear that a plane wave incident upon a cloud of a large number of tiny opaque obstacles will have a large portion of the incident wave motion redirected in this manner.

One can not apply Huygen's principle to analyze wave motion in all cases since it does not rest entirely upon first principles and ignores polarization effects of wave motion. It is useful in grasping the essential features of the interference and diffraction of wave motion incident upon obstacles. Interference and diffraction are treated extensively in Jenkins and White, "Fundamentals of Optics" chapters 13 through 18. The student should read these chapters carefully. Although the emphasis of these chapters is on visible light, the student should bear in mind that the properties of wave motion are being described and can be applied to microwaves, acoustic waves, and seismic waves. Throughout these chapters, the term, intensity, is used with the symbol,I. Such a use of the terms have been customary in the past and implies only the square of the wave displacement and corresponds to irradiance, E, to be defined later.

When one deals with highly coherent monochromatic radiation two equal values of irradiance of two waves superposed do not add directly at a given point but the resultant irradiance may be any amount between zero and four times the irradiance of one. Nevertheless, the concept of irradiance of the resultant superposed wave is fully consistent with the concept of irradiance for noncoherent radiation as long as the addition of wave displacement is done correctly.

The passage of a wave motion through a boundary between two media having different wave propagation velocities also changes the geometry of the wave front. It is possible to describe these changes in geometry by extending Huygen's principle to include generators with phase delays and certain angular dependent properties. However, another and simpler way of describing the geometry of the wave front is by specifying the directions of propagation of the waves at a large number of points along the wave front. Such a description using these "rays" or directions is called geometrical optics or ray optics. The term optics in this context has come to apply to any kind of wave motion. Geometrical optics will be treated in the section on optics, mirrors, and lenses, and can also be found in Jenkins and White, "Fundamentals of Optics" Part I.

3. THE QUANTITATIVE TREATMENT OF RADIATED POWER

The information which we are to use for remote sensing is to be derived from the quantitative measure of the properties of some force field at a location remote from the materials which are modifying that force field. In the case of radiated force fields, we are to ascertain the identity and location of materials of interest by quantitative measures of the properties of the wave motion arriving at our location. In order to measure or "detect" the nature of this radiation, we must extract some energy from this wave motion. It is helpful to establish a common nomenclature and symbolism for the further discussion of these measures.

There are many nomenclatures and symbol systems for this purpose in existence — too many in fact. The one which will be adopted here is the one adopted by the International Commission on Illumination which is closest to being universally recognized (see table 3-1). This nomenclature may also be extended to apply to acoustic and seismic wave motions so that a common language will help to sweep away apparent distinctions between technologies which only exist because of language.

Energy, Q.

The total energy supplied to wave motion or supplied by wave motion to a receiver is the basic quantity to be measured. It is a measure of the capacity to do physical work or of the expenditure of physical work by a force moving through some displacement — e. g., a pressure expanding through a volume, an electric field moving an electric charge, or a magnetic force moving a magnet. The unit of energy is the joule.

Power, P, or Φ.

Power is the time rate at which energy is being supplied, transmitted, received, or expended. If the power being transmitted is constant in time, then the energy being transmitted increases with time so that $Q = Pt$. Clearly one can transmit the same energy by transmitting a large power over a short time interval or a small power over a long time interval. The energy Q is the quantity being transmitted, the power P is a measure of the time rate of flow. In remote sensing we are certainly concerned with the energy supplied by wave motion; however, we are not interested in waiting too long to accumulate the needed quantity. We are interested in some minimum flow rate in order to be timely. Hence, the principle concern in remote sensing is the power required for timely measurement. The unit of power is the watt so that

1 watt = 1 joule/sec.

It is common for many persons to use the term power and energy as if they meant the same

TABLE 3-1. USA STANDARD SYMBOLS FOR ILLUMINATING ENGINEERING

Quantity*	Symbol*	Defining Equation **	Commonly Used Units	Symbol
Radiant energy	Q, (Q_e)		erg joule calorie kilowatt-hour	 J cal kWh
Radiant density	w, (w_e)	$w = dQ/dV$	joule per cubic meter erg per cubic centimeter	J/m^3 erg/cm^3
Radiant flux	Φ, (Φ_e) P, (P_e)	$\Phi = dQ/dt$	erg per second watt	erg/s W
Radiant flux density at a surface				
Radiant exitance (Radiant emittance)	M, (M_e)	$M = d\Phi/dA$	watt per square centimeter	W/cm^2
Irradiance	E, (E_e)	$E = d\Phi/dA$	watt per square meter, etc.	W/m^2
Radiant intensity	I, (I_e)	$I = d\Phi/d\omega$ (ω = solid angle through which flux from point source is radiated)	watt per steradian	W/sr
Radiance	L, (L_e)	$L = d^2\Phi/d\omega\,(dA \cos\theta) = dI/(dA \cos\theta)$ (θ = angle between line of sight and normal to surface considered)	watt per steradian and square centimeter watt per steradian and square meter	$W \cdot sr^{-1}\,cm^{-2}$ $W \cdot sr^{-1}\,m^{-2}$
Emissivity	ϵ	$\epsilon = M/M_{blackbody}$ (M and $M_{blackbody}$ are respectively the radiant exitance of the measured specimen and that of a blackbody at the same temperature as the specimen)	one (numeric)	–
Absorptance	α, (α_v, α_e)	$\alpha = \Phi_a/\Phi_i$ ***	one (numeric)	–
Reflectance	ρ, (ρ_v, ρ_e)	$\rho = \Phi_r/\Phi_i$ ***	one (numeric)	–
Transmittance	τ, (τ_v, τ_e)	$\tau = \Phi_t/\Phi_i$ ***	one (numeric)	–

NOTE: The symbols for photometric quantities are the same as those for the corresponding radiometric quantities (see above). When it is necessary to differentiate them the subscripts v and e respectively should be used, e.g., Q_v and Q_e.

* Quantities may be restricted to a narrow wavelength band by adding the word spectral and indicating the wavelength. The corresponding symbols are changed by adding a subscript λ, e.g. Q_λ, for a spectral concentration or a λ in parentheses, e.g. $K(\lambda)$, for a function of wavelength.

** The equations in this column are given merely for identification.

*** Φ_i = incident flux
Φ_a = absorbed flux
Φ_r = reflected flux
Φ_t = transmitted flux

TABLE 3-1. USA STANDARD SYMBOLS FOR ILLUMINATING ENGINEERING (Continued)

Quantity*	Symbol*	Defining Equation**	Commonly Used Units	Symbol
Luminous energy (quantity of light)	Q, (Q_v)	$Q_v = \int_{360}^{760} K(\lambda) Q_{e\lambda} d\lambda$	lumen-hour lumen-second (talbot)	lm·h lm·s
Luminous density	w, (w_v)	$w = dQ/dV$	lumen-second per cubic meter	$lm \cdot s \cdot m^{-3}$
Luminous flux	Φ, (Φ_v)	$\Phi = dQ/dt$	lumen	lm
Luminous flux density at a surface				
Luminous exitance (Luminous emittance)	M, (M_v)	$M = d\Phi/dA$	lumen per square foot	lm/ft^2
Illumination (illuminance)	E, (E_v)	$E = d\Phi/dA$	footcandle (lumen per square foot) lux (lm/m^2) phot (lm/cm^2)	fc lx ph
Luminous intensity (candlepower)	I, (I_v)	$I = d\Phi/d\omega$ (ω = solid angle through which flux from point source is radiated)	candela (lumen per steradian)	cd
Luminance (photometric brightness)	L, (L_v)	$L = d^2\Phi/d\omega\,(dA \cos\theta) = dI/(dA \cos\theta)$ (θ = angle between line of sight and normal to surface considered)	candela per unit area stilb (cd/cm^2) nit (cd/m^2) footlambert ($cd/\pi ft^2$) lambert ($cd/\pi cm^2$) apostilb ($cd/\pi m^2$)	cd/in^2, etc. sb nt fL L asb
Luminous efficacy	K	$K = \Phi_v/\Phi_e$	lumen per watt	lm/W
Luminous efficiency	V	$V = K/K_{maximum}$ ($K_{maximum}$ = maximum value of $K(\lambda)$ function)	one (numeric)	–

* Quantities may be restricted to a narrow wavelength band by adding the word spectral and indicating the wavelength. The corresponding symbols are changed by adding a subscript λ, e.g. Q_λ, for a spectral concentration or a λ in parentheses, e.g. $K(\lambda)$, for a function of wavelength.

** The equations in this column are given merely for identification.

thing; however, confusion of these two terms leads to obscure communication between people and probably is the cause of incorrect reasoning in analyzing the functioning of remote sensing devices.

Irradiance, E.

The power in wave motion flows in the direction of wave propagation and hence is distributed in space. Since any object in the way of this flow is finite in size, the power which can be intercepted will depend upon the spatial distribution of the power. The power flow which is intercepted per unit area of receiving area from any direction in the hemisphere on the receiving side of the surface is called irradiance. The unit of irradiance is watts/meter^2. If the sources of wave motion are non-coherent, then the irradiance on a surface due to one source, E_1, adds to the irradiance on the surface due to another source, E_2, so that the total irradiance, E is equal to $E_1 + E_2$.

The precise direction or directions of flow are not necessarily specified as long as the flow is directed within a hemisphere of directions toward the surface.

Exitance, M (also called radiant emittance).

A surface may be the distributed source of non-coherent radiation either by reflecting incoming power or by generating power by some internal mechanism. The term, exitance, is used to label the power per unit area of source surface leaving the surface in any of a hemisphere of directions away from the surface. The units are watts/meter^2.

Solid Angle, Ω.

A solid angle is an extension to solid geometry of the arc angle in plane geometry.

An angle in plane geometry can be measured by the length of the circular arc (with angle apex as center) subtended by the two sides of the plane angle divided by the radius of the circular arc. The ratio of arc length to arc radius is the angle size in radians. A full 360^o plane angle is equal to 2π radians since $2\pi R$ is the arc length of a full circle with arc radius R.

The solid angle of a cone is measured by the area of a sphere (with center at the cone apex) intersected by the cone divided by the square of the radius of the sphere. A full sphere of solid angle must have 4π steradians since the surface of a sphere is $4\pi R^2$ when the radius is R. Clearly a cone which is wide enough to be a plane surface has a solid angle of 2π steradians since a "plane" cone intersects a hemisphere of spherical surface.

In calculating the solid angle subtended by a finite plane area such as a camera aperture at a long distance R from a point normal to the direction from point to aperture, one assumes that the area of the aperture is insignificantly different from the area of spherical surface

which would be intersected by the cone so that the solid angle subtended by the aperture, Ω is given well enough by:

$$\Omega = \frac{\text{Area of aperture}}{R^2}$$

If the plane aperture area is not normal to the line between the point and the aperture center, then the solid angle subtended is found by using the projected area of the aperture normal to that line. Although, the solid angle is a dimensionless quantity, a unit symbol is used to distinguish the angular measure from other possible measures and also to assist in unit analysis during computation. The unit symbol of steradian is sr.

Radiance, L.

In most cases, we will be interested in calculating the power intercepted by a receiving aperture due to the exitance, M, from some distant surface. In that case, it will be necessary to know in more detail the directional distribution of power leaving each element of the surface. The quantity, radiance, is the amount of power per unit projected source area per unit solid angle subtended by the receiving aperture in the direction of the receiving aperture. Radiance and brightness are similar concepts. Brightness specifically applies to the visual effect of radiance. A good piece of diffuse white paper does not change in brightness as the angle of view is changed. Such a paper is said to have a constant radiance with aspect angle if the power per unit solid angle from such a paper varies proportionally to the projected area of the paper in the direction of view. The ideal diffuse surface having a constant radiance with aspect angle is called a Lambertian surface.

Intensity, I.

If a source of power is so small as viewed from a distance that the spatial extent of the source is immaterial to a power transfer calculation, an intensity, I, of the source may be defined as the power per unit solid angle leaving that source. A distant star is an example of a source for which an intensity may be established. The units are watts/steradian. The intensity need not be the same in all directions, however. As a matter of fact, the variation of intensity as a function of viewing angle may supply important data by which the nature of the source can be ascertained.

Absorptance, α.

The absorptance of a surface is defined as the ratio of the power absorbed by the surface to the power incident upon that surface.

$$\alpha \equiv \frac{\text{P absorbed}}{\text{P incident}}$$

Reflectance, ρ.

The reflectance of a surface is defined as the ratio of the power reflected by the surface to the power incident upon that surface.

$$\rho \equiv \frac{\text{P reflected}}{\text{P incident}}$$

Transmittance, τ.

The transmittance of a surface is defined as the ratio of the power transmitted through the surface to the power incident upon that surface.

$$\tau \equiv \frac{\text{P transmitted}}{\text{P incident}}$$

By conservation of energy for a passive surface the energy flowing toward a surface must be conserved. Hence, the time rate of flow of energy (power) must also be accounted for. The amount of energy at any moment should be either stored in the surface by absorption, sent back by reflection or sent on forward by transmission. No other possibility is admitted. Hence, one can see that

$$\text{P incident} = \text{P absorbed} + \text{P reflected} + \text{P transmitted}$$

Dividing both sides of this relation by P incident we find that

$$1 = \frac{\text{P absorbed}}{\text{P incident}} + \frac{\text{P reflected}}{\text{P incident}} + \frac{\text{P transmitted}}{\text{P incident}}$$

or $1 = \alpha + \rho + \tau$

In the general definition of absorptance, reflectance, and transmittance no specification as to the direction of incident and reflected power is made. In many situations, the directional properties of these quantities are important. Modifiers used with these terms imply directional restrictions or other special conditions of measurement of the incident, reflected, and transmitted power. A discussion of directional reflectance is given in Appendix I. Similar modifiers can apply to the absorptance and transmittance.

3.1. SPECTRAL QUANTITIES

Each of the foregoing quantities, Q, P, E, M, L, and I may be made specific to the energy or power transmitted by a group of simple harmonic components of a noncoherent wave. Since any finite noncoherent wave must be described by an infinite number of simple harmonic waves with infinitesimal amplitudes, the power carried by a finite but narrow spectral band of these components tends to increase as the spectral band is widened to include more components. The power per unit wavelength interval is finite and is given the symbol

P_λ where the wavelength, λ, appears as a subscript. The units of P_λ are watts/(micrometer of wavelength) or watts per the unit wavelength which is appropriate for the spectral region. The energy delivered in a short interval of time to a receiver per unit wavelength interval of component waves, Q_λ, is then P_λ times the short time interval. Similarly the contribution per unit of small wavelength interval of simple harmonic components can be symbolized for the other power quantities, E_λ watts/m^2(μm) L_λ watts/(projected m^2) sr (μm), I_λ watts/sr(μm).

The spectral quantities, such as P_λ, are assumed to be a function of the wavelength of interest.

From the definition of α, ρ, and τ one can see that the spectral power may be substituted for the total power so that a spectral absorptance, reflectance and transmittance may be defined as a function of wavelength. However, these spectral quantities are symbolized as α (λ), ρ (λ), τ (λ) to show spectral dependence. The quantities are still dimensionless so that a subscript is not used.

3.2. ILLUMINATION QUANTITIES

Historically, the quantitative measurement of noncoherent visible light has been of major importance because of the importance of human vision. The human eye is not equally responsive to power of all wave lengths in the visible region, however. An irradiance of one watt per square meter on the human retina for a spectral band in the red part of the visible spectrum appears as a dim light compared to 1 watt per square meter of green light. The illumination engineers wished to obtain a power like unit so that one unit of this "visible" power would be perceived as bright no matter what wavelength in the visible spectrum one may choose. The quantity of "visible" power or luminous power is called the lumen so that at a wavelength of 0.555 μm one watt of power would correspond to 680 lumens. The correspondence between watts of radiant power and the visible quantity, lumen, changes as a function of wavelength depending upon the relative responsiveness of the eye. The response of the normal human eye relative to its maximum response at 0.555 μm as a function of wavelength is called the luminous efficiency, or the visibility curve or the relative luminosity factor and is determined by extensive measurement of the daylight vision of a large number of people. From the visibility curve (figure 3-1) one can determine that if 1 watt of 0.555 μm radiant power provides 680 lumens then 10 watts of 0.650 μm radiant power are required to produce the same value of 680 lumens and hence the same degree of visual response because the visual response has decreased to only one tenth of its peak value for such a wavelength. For any arbitrary spectral distribution of power the luminous power P_v, in lumens is found by using the luminosity curve, $K(\lambda) = 680 \frac{\text{lumens}}{\text{watt}} V(\lambda)$ and the spectral radiant power, $P_{e,\lambda}$

Thus,

$$P_v \text{ lumens} = 680 \frac{\text{lumens}}{\text{watt}} \int_0^\infty V(\lambda) P_{e,\lambda} \frac{\text{watt}}{\mu\text{m}} d\lambda(\mu\text{m}),$$

where $V(\lambda)$ is the visibility curve value at wavelength, λ. The same set of symbols and power flow concepts are applied to the luminous power, P_v, as are applied to the actual power flow. If context does not make clear when the use of luminous power and radiant power is intended a subscript is to be added to the symbols.

Thus, P_v is luminous power (lumens) and

P_e is radiant power (watts).

All other illumination quantities may be determined in the same manner as the luminous power is determined. The relation, $P_{v,\lambda} = 680 \frac{\text{lumen}}{\text{watt}} V(\lambda) P_{e,\lambda}$, between spectral power and luminous spectral power can be used to determine α_v, ρ_v, and τ_v.

For example,

$$\rho_v = \frac{680 \int_0^\infty V(\lambda) P_{e,\lambda}(\text{reflected})\, d\lambda}{680 \int_0^\infty V(\lambda) P_{e,\lambda}(\text{incident})\, d\lambda} = \frac{P_v(\text{reflected})}{P_v(\text{incident})}$$

There are a large number of luminous units in the literature. Learning all of them would not be worth while until they are encountered. A few of these are most popular. The unit of I_v which may be expressed in lumens per steradian is also called the candle and more recently the candela. The unit of L_v which may be expressed in lumens per steradian per projected square meter may many times be expressed by a different sized unit called the lambert; where

$$1 \text{ lambert} = \frac{10^4}{\pi} \text{ lumen sr}^{-1}\text{m}^{-2}.$$

The unit of E_v which may be expressed in lumens/m^2 is often expressed as a (metercandle). Since these specialized units make unit analysis difficult and add nothing more, they are not recommended for computational purposes.

3.3. PHOTOPIC AND SCOTOPIC VISIBILITY CURVES

The human eye response to different wavelengths of light changes significantly from daylight adapted conditions to dark adapted conditions. Daylight adapted vision is called photopic vision. The definition of the lumen from the photopic visibility curve has been discussed. If one is to discuss night vision, the lumen must be defined using the scotopic visibility curve. In most problems of remote sensing, night vision is not important. However, in the literature

on vision one must be on guard as to whether or not the lumen refers to the scotopic or photopic visibility curve. In scotopic vision the maximum response of the eye occurs at 0.510 μm and 1 watt at that wavelength produces 1,746 lumens in scotopic vision.

4. GEOMETRICAL OPTICS

One of the most important properties of wave motion in relation to remote sensing is the direction of propagation of incident radiated power. Geometrical optics treats these directions as rays. The interaction of these rays of propagation with macroscopic obstacles (ignoring the diffraction or interference contribution predicted by Huygen's principle) makes possible a simple description of the functions of mirrors and lenses. One of the functions of mirrors and lenses is to sort out the incident power in accordance with direction of propagation of the incoming superposed components. The power radiated from some field point can then be identified at the receiver by the direction from which that contribution came. Quantitative measurements can then be applied to ascertain the nature of the material at that field point and its spatial distribution. The following discussion reviews the essential features of geometrical optics.

4.1. INDEX OF REFRACTION

Electromagnetic waves propagating through empty space travel at a velocity, c, (about 3×10^8 meters/sec.). Electromagnetic waves can also penetrate matter particularly if the matter does not conduct electric current easily. The velocity of propagation of electromagnetic waves in matter is not the same as the velocity in free space. The index of refraction of poorly conducting matter is defined as the ratio of the velocity of propagation in free space to the velocity v in the matter.

$$n = \frac{c}{v}.$$

The index of refraction depends upon the kind of matter and is an attribute which could be measured remotely to help identify materials. The emphasis of the following discussion is on the functions of mirrors and lenses used in electromagnetic sensors; however, the same relationships hold with matter at the remote end of the remote sensing process.

Further, the behavior of mirrors and lenses are not unique to electromagnetic waves but apply to acoustic and seismic waves equally well. (In such a case, the relative index of refraction of one medium to another must be defined using the acoustic or seismic velocities.)

4.2. SNELL'S LAW

It is an experimental fact that when waves traveling in a material (1) with index of refraction n_1 is incident on the plane boundary separating this material (2) with index of refraction n_2, the incident wave is divided into two parts. One part, the reflected wave, is reflected back into the material (1). The second part, which travels through the material (2) is called the refracted wave. If as in figure 4-1 we consider the rays associated with the incident, reflected,

and refracted waves, and the angles which they make with the normal to the surface separating the materials, we can write down certain laws. Call θ_1 the angle between the incident ray and the normal, and θ_1' the angle between the reflected ray and the normal, then the law of reflection states

$$\boxed{\theta_1 = \theta_1'.} \tag{1}$$

Now if we call θ_2 the angle between the refracted ray and the normal, the law of refraction (also called Snell's law) is

$$\boxed{n_1 \sin \theta_1 = n_2 \sin \theta_2} \tag{2}$$

There is a third relationship between the three rays. It is, THE INCIDENT, REFLECTED, AND REFRACTED RAYS AND THE NORMAL TO THE SURFACE ALL LIE IN THE SAME PLANE. The plane containing the incident ray and the normal is called the INCIDENT PLANE. Thus the three rays and the normal lie in the incident plane. If the boundary surface is curved, we can still apply the above laws by using the angles between the rays and the normals to the surface at the points where the incident rays meet the surface. This is illustrated in figure 4-2.

When there are two surfaces separating three different types of materials, the laws of reflection and refraction are applied to each surface. Thus for the case illustrated in figure 4-3, we have for the surface separating ① and ②

$$n_1 \sin \theta_1 = n_2 \sin \theta_2$$

and for the surface separating ② and ③

$$n_2 \sin \theta_2' = n_3 \sin \theta_3.$$

The extension to more than two surfaces is obvious.

It is a well known and easily demonstrated fact that refraction directionally separates white light into its component colors. This phenomenon illustrates that, in general, the index of refraction is a function of wavelength. We can see this from the law of refraction as shown in figure 4-4. White light in air (n = 1) is incident on the surface of a material whose index is n' at the angle θ. Therefore, for all the wavelengths, the incident angle is θ. But after refraction, the different wavelengths make different angles θ' with the normal. For this case we write

$$\sin \theta = n' \sin \theta',$$

where θ is the same for all wavelengths, but not θ'. From this equation, it is easy to see that if θ' depends on wavelength (λ) then so must n' in order that the equation be true. The dependence of index of refraction and therefore the velocity of propagation on wavelength is called

DISPERSION. In order to spread radiation into its component wavelengths, many instruments make use of a "prism" or block of material with a triangular cross-section. When radiation passes through the first face of the prism (see figure 4-5) dispersion directionally separates the incident radiation into its spectral component waves, further separation takes place upon passing through the second face.

Let us examine the law of refraction, Equation (2), as the angle of incidence is changed. This is illustrated in figure 4-6 where a ray from a material of index n_1 is incident upon one of lesser index n_2. The incident and reflected angles are θ_1 and the refracted angle is θ_2. As θ_1 is increased from figure 4-6(a) to 4-6(b), we see by Equation (2) that θ_2 also increases, and since n_2 is less than n_1 then θ_2 must be greater than θ_1. Increasing θ_1 we arrive at the case where $\theta_2 = 90^\circ$ (but $\theta_1 < 90^\circ$). This indicates that the refracted ray travels along the surface. For this case we call the incident angle "critical angle" and denote it by θ_c. Its value is obtained from Equation (2) by substituting $\theta_2 = 90^\circ$ giving

$$\sin \theta_c = \frac{n_2}{n_1}. \tag{3}$$

When θ_1 is increased further, so that $\theta_1 > \theta_c$, we get the requirement that $\sin \theta_2$ must be greater than one in order to satisfy Equation (2). This is impossible. The result is that there is no refracted ray when $\theta_1 > \theta_c$, only the incident and reflected rays. Such a phenomenon is called TOTAL REFLECTION. Remember that total reflection can only occur when the radiation is traveling from a material of higher index of refraction into one of lower index of refraction.

It is easily demonstrated experimentally that when an unpolarized beam of light of a single wavelength is incident on a quartz or calcite crystal, there is in addition to the reflected beam, two refracted beams instead of one. The phenomenon is called DOUBLE REFRACTION and is illustrated in figure 4-7 for calcite. Consider a ray traveling in air which is incident at the angle θ on the calcite. When the angle of refraction for the two rays is measured, it is found that the law of refraction

$$1 \times \sin \theta = n' \sin \theta'$$

holds for one ray and not for the other. The ray which satisfies the above equation is called the ordinary or O ray, the other is called the extraordinary or E ray.* Furthermore, the incident, reflected, and ordinary rays all lie in the same plane containing the normal while

* The use of the symbol, E, as an abbreviation for EXTRAORDINARY is customary in the literature and should not be confused with the symbol for irradiance in this context.

the extraordinary ray does not lie in this plane. If the two rays are checked for polarization, it will be found that the E and O rays are each linearly polarized but in perpendicular directions.

4.3. FORMATION OF IMAGES

The rays from a point source are along the radii of a sphere centered at the source (see figure 4-8). A pencil of radiation is defined as the radiation within a narrow cone and bounded by rays from the same source. A convergent pencil is one for which the bounding rays travel towards a common point of intersection. A divergent pencil is one for which the bounding rays proceed from a common point. Consider the reflection of a pencil of rays from a point source, P, located in front of a plane mirror*. This is illustrated in figure 4-9 where for simplicity only the bounding rays of the pencil are drawn. If the surface is not totally reflecting, that is if there are refracted rays, we shall neglect the refracted rays for the present. The rays proceeding from the point, P, will, after reflection, appear to be coming from a point, P', behind the surface. The point P' from which the rays appear to come is called the VIRTUAL IMAGE of the object P because the rays do not actually come from there, they only appear to do so. If the rays actually did pass through P', it would be called the REAL IMAGE of P. A plane reflecting surface does not produce a real image. Now by application of relation (1) it can be shown that P' appears to be the same distance behind the surface (or mirror) as P is in front. Let us measure distances always from the reflecting surface and call S the distance to P and S' the distance to P', both measured normal to the surface. Since distance can, in general, be measured in one of two directions for any point, a convention must be adopted to decide whether a distance is called positive or negative:

> IF THE DISTANCE TO THE OBJECT IS MEASURED IN THE DIRECTION OPPOSITE TO THE INCIDENT DIRECTION, CALL IT POSITIVE (OTHERWISE IT IS NEGATIVE). IF THE DISTANCE TO THE IMAGE IS MEASURED IN THE SAME DIRECTION AS THE REFLECTION CALL IT POSITIVE (OTHERWISE IT IS NEGATIVE).

Hence for reflection from a plane surface,

$$\boxed{s = -s'}\,. \tag{4}$$

The negative sign appears in front of S' because it is measured in a direction opposite to the reflected rays (see figure 4-10).

Next let us consider what happens to a pencil of rays after refraction at a plane surface. This is illustrated in figure 4-10. The refracted rays appear to come from the virtual image

---

* The symbol P, here, represents a position and does not refer to the power of the source.

P'. However, if we look at another pencil of rays which has a different angle of incidence, it will appear to come from a different image point P''. Thus a refracting plane surface does not produce a point image for a point object. Instead the image is blurred. If we confine our radiation to a narrow cone, then to a good approximation a point source is imaged to a point. Furthermore, if we confine ourselves to a pencil of rays which makes only a small angle of incidence with the normal, then we may write

$$\boxed{n's = -ns'}\ , \tag{5}$$

where s is the distance to the object, s' the distance to the image, n the index of the material in which the object is located, and n' the index of the material from which observations are made. We also use a sign convention for refraction:

> IF THE DISTANCE TO THE OBJECT IS MEASURED OPPOSITE TO THE DIRECTION OF INCIDENCE IT IS POSITIVE. IF THE DISTANCE TO THE IMAGE IS MEASURED IN THE SAME DIRECTION AS THE DIRECTION OF REFRACTION IT IS POSITIVE.

We see from figure 4-10 why there is a negative sign in from of s' in equation (5).

One of the most important imaging devices is the lens. A lens is constructed of a material which is transparent to the radiation to be used. First we shall discuss thin lenses and later thick lenses. A thin lens is defined as one whose thickness can be neglected in calculations. In figure 4-11 are illustrated some of the more common types of lenses. Those lenses which are thicker at the center and thinner at the edges are called converging or positive lens. Those which are thinner at the center and thicker at the edges are called diverging or negative lens.

For a lens there are a number of important terms to be defined. The AXIS is a line through the center of the lens and perpendicular to each surface at the points of intersection. A ray along the axis of the lens passes through the lens undeviated. Every lens has a PRIMARY FOCAL POINT and a SECONDARY FOCAL POINT. These are illustrated for both positive and negative lenses in figure 4-12. The primary focal point F is a point on the axis located for a positive lens at a position for which a diverging pencil from that point is refracted into a parallel beam. For a negative lens it is the position for which a pencil converging towards that point is refracted into a parallel beam. For a positive lens the secondary focal point F' is that point on the axis at which rays parallel to the axis intersect (or are focussed) after passing through the lens. With a negative lens, the rays parallel to the axis are refracted into a pencil which appears to come from F'. The distance from the lens to F is called f, and from the lens to F' is called f'. For a thin lens, f = f', if air is on both sides of the lens. The distance f is known as the focal length. For a converging or positive lens f is taken as positive, for a diverging or negative lens f is negative.

The plane through a focal point and perpendicular to the axis is called a FOCAL PLANE. Parallel rays making some angle with the axis are brought to focus at some point on the focal plane as illustrated in figure 4-13.

If we place a point object in from of a lens, the lens will form a point image of this object. The image will be real or virtual depending on the type of lens and the position of the object. Figure 4-14 illustrates a point object P placed in front of a positive lens so as to form a real image P'. For any thin lens we have the equation

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \tag{6}$$

where s is the distance from the lens to the object and s' the distance from the lens to the image. The sign convention is as before: the distance to the object is positive if measured in a direction opposite to the incident rays, and s' is positive if measured in the direction of the refracted rays.

When the object is not a point but is extended then the image is also extended. Illustrated in figure 4-15, we have an object with end points P and Q which has an image with end points P' and Q'. For extended objects, a quantity of interest is the lateral magnification m. It is defined as the ratio of the height of the image y' to the height of the object m

$$m = \frac{y'}{y} \tag{7}$$

The sign convention for the heights are:

IF THE HEIGHTS ARE MEASURED UPWARDS FROM THE AXIS, THEY ARE POSITIVE, OTHERWISE THEY ARE NEGATIVE.

For a thin lens one can show that another expression for the lateral magnification is

$$m = \frac{-s'}{s} \tag{8}$$

From a consideration of the sign conventions the following statements can be made:

If s' is positive the image is real

If s' is negative the image is virtual

If m is positive the object and image or either both upright, or both upside down

If m is negative, one is upright and the other is upside down.

Many optical systems consist of more than one lens. If this be the case, then the treatment is simple. We treat each lens of the system separately, but the object for each lens after the first is the image formed by the preceding lens. An example of a system with three thin lens is shown in figure 4-16.

Depicted are lenses L_1, L_2, and L_3 together with their appropriate focal points. Also shown are two sample rays. The object O for the first lens, a distance s_1 produces an image I_1 a distance s'_1 from L_1. Notice that I_1 is formed behind L_2 and the rays do not actually cross there, it is the virtual object for L_2. Also the distance of I_1, $-s_2$, from the lens L_2 is negative because it is measured from the lens in a direction opposite to the incident rays. Next, the image I_2 of I_1 is formed by L_2. This is a real image. This image becomes the object for L_3 which forms the real image I_3. Calling the distances between L_1 and L_2 as "a" and between L_2 and L_3 as "b" the equations are:

$$\frac{1}{s_1} + \frac{1}{s'_1} = \frac{1}{f_1}$$

$$\frac{1}{s_2} + \frac{1}{s'_2} = \frac{1}{f_2} \quad \text{where } s_2 = -(s'_1 - a)$$

$$\frac{1}{s_3} + \frac{1}{s'_3} = \frac{1}{f_3} \quad \text{where } s_3 = b - s'_2.$$

Another imaging device is the spherical mirror. It is a portion of a silvered sphere, which uses reflection to form the image. Figure 4-17 illustrates the two types of spherical mirrors, a concave or converging mirror and a convex or diverging mirror.

As with lenses, there are important terms to be defined for spherical mirrors. (These are shown in figure 4-18.) The AXIS is a line along that diameter of the sphere which bisects the mirror. The FOCAL POINT, F, is that point on the axis at which incident rays parallel to the axis are brought to focus with a concave mirror; while for a convex mirror, it is that point from which the rays parallel to the axis appear to come after reflection. The distance f along the axis from the mirror to the focal point is positive for a concave lens and negative for a convex lens. The relationship between the focal length and the radius of curvature γ, is $f = r/2$ for a concave mirror and $f = -r/2$ for a convex mirror.

That plane through F which is perpendicular to the axis, as shown in figure 4-19 is called the FOCAL PLANE. Parallel rays making arbitrary angles with the axis are focussed at some point on the focal plane.

The properties of a spherical mirror are much like that of a lens. A point source shown in figure 4-20 placed at a distance s from the mirror surface will produce a point image at a distance s from the mirrors. The distances satisfy the equation

$$\boxed{\frac{1}{s} + \frac{1}{s'} = \frac{1}{f}}\ , \qquad (9)$$

with the convention that the distance s to the object is positive when measured opposite to the direction of the incident rays (otherwise it is negative) and the distance s' to the image is positive if measured in the direction of the reflected rays (otherwise it is negative). When s' is positive the image is real, when negative it is virtual. For extended objects of height y the magnification m is

$$\boxed{m = \frac{y'}{y}} \tag{10}$$

where y' is the height of the image. The sign convention is the heights y and y' are positive when measured upwards from axis and negative when measured downwards. It may be shown that

$$\boxed{m = -\frac{s'}{s}} \tag{11}$$

as with lenses. Figure 4-21 illustrates the various lengths and heights as associated with an object PQ and its image P'Q' as formed by a spherical mirror.

When the thickness of a lens cannot be neglected compared with the focal length, it is called a thick lens. It can be shown for a thick lens that the relationships between distances are the same as for a thin lens if these distances are properly measured. Before proceeding, we must again define important terms. To illustrate these terms refer to figure 4-22. First, there is the AXIS which is defined exactly as for the thin lens. Also, the primary focal point F and the secondary focal point F' are defined as before, and illustrated in figure 4-12. Figure 4-22 (a) and (b) illustrates the passage of two rays through a thick lens as they proceed from and towards the focal points.

Included in the figures are the vertices A_1 and A_2 which are the intersections of the axis with the surfaces of the lens. Extend from the lens surfaces two lines along the directions of incidence and refraction for each ray of a point source at F and F' until these lines intersect. Those planes through the intersections and perpendicular to the axis are called PRIMARY and SECONDARY PRINCIPAL PLANES, depending upon whether the source is at the primary or secondary focal points. The point where the primary principal point intersects the axis, H, is called the primary principal point; the point where the secondary principal point intersects the axis, H', is the secondary principal point. Now Equation (6), namely, is still valid if the object distance s is measured from the primary principal plane parallel to the axis, the image distance s' is measured from the secondary principal plane parallel to the axis, and the focal distance f is measured from the principal focal plane. In addition, f = f'. The sign conventions for s, s', y, and y' are the same as for the thin lens. Generally for thick lens, the principal points, planes, and focal points need not be symmetrically located with respect to the vertices. Illustrated in figure 4-23 are the positions of the principal planes and points for some

representative converging lenses of equal focal length. Notice that the labeling of the primary and secondary planes depends upon which face of the lens the radiation is first incident.

4.4. SHORT REVIEW OF THE THEORY OF STOPS

Every optical system has an aperture that effectively limits the power traversing the system. This aperture may be a lens periphery, a diaphragm, or some other obstacle and is referred to as the aperture stop. Another stop is called the first stop; it controls the field of view. Additional stops are sometimes used to reduce aberration effects and glare.

4.4.1. THE APERTURE STOP — ENTRANCE AND EXIT PUPILS. We define the aperture stop in the following way. Of all the possible stops, that one which limits the divergence of the rays from a point P on the optical axis (i.e., which determines which rays from P will traverse the system) is the aperture stop. It should be noted that the aperture stop with respect to some position of P in the object space is not necessarily the aperture stop with respect to another location of P.

The entrance pupil is defined as the image of the aperture stop as imaged through the optical elements to the left of the stop. (We are using the standard convention in which the light is incident on the optical elements from the left.) Similarly, the exit pupil is defined as the image of the aperture stop as imaged through the optical elements to the right of the stop. When the aperture stop is placed at the first principal focal point F of a system, the exit pupil is at infinity. Such a system is called telecentric on the image side. A system telecentric on the object side has the aperture stop at the second focal point of the system, so that the entrance pupil is at infinity.

4.4.2. THE PRINCIPAL RAY OR CHIEF RAY. That ray coming from the object (not necessarily on the optical axis) which is directed toward the center of the entrance pupil, or which passes through the center of the aperture stop is called the principal ray, or chief ray. The principal ray acts as an axis of symmetry for the bundle of rays it represents and so defines the direction of the pencil as it traverses the system.

4.4.3. THE FIELD STOP — ENTRANCE AND EXIT WINDOWS. In addition to the aperture stop other stops or diaphragms that are present put a limit of the size of the object whose image can be formed. The stop which limits the field of view is called the field stop. We define the field stop and its associated entrance and exit windows in the following manner. Of all the images formed by the optical elements to the left of any stop, that one whose edge subtends the smallest angle at the center of the entrance pupil limits the field to the greatest extent and is called the entrance window. The physical stop of which the entrance window is

the image is called the field stop. Similarly that image of the field stop formed by the optical components to its right, is known as the exit window. The exit window is that stop image which subtends the smallest angle at the exit pupil. The angle that the diameter of the entrance window subtends at the center of the entrance pupil is known as the angular field of view in the object space. The angle subtended at the center of the exit pupil by the diameter of the exit window is known as the angular field of view in the image space.

4.4.4. RELATIVE APERTURE AND f NUMBER. For a lens which forms the image of an extended object, the amount of power intercepted from a small radiating area of the object and transmitted to the image plane is directly proportional to the area of the entrance pupil. The irradiance at the image of that small area is inversely proportional to the image area. $\left(E = \frac{P}{A}\right)$. The area of the entrance pupil is proportional to the square of the pupil radius and the image area is proportional to the square of the focal length; hence, the ratio of the square of these two dimensions is a measure of the irradiance at the image. The ratio of the entrance pupil diameter to the focal length of the lens is known as the relative aperture. The reciprocal of the relative aperture is called the f number.

4.4.5. DEPTH OF FOCUS AND DEPTH OF FIELD. The depth of focus for an optical system is defined as the linear distance by which the image may be shifted on the optical axis from a given reference plane before introducing a predetermined acceptable amount of blurring or defocusing. In a similar fashion, the depth of field is the amount by which the object may be shifted along the optic axis before introducing the same amount of blurring or defocusing. Generally, the measure of the acceptable blur is given as the linear diameter of the acceptable blur spot.

4.5. ABERRATIONS IN GEOMETRICAL OPTICAL SYSTEMS

To this point, it was assumed that a lens or mirror could sort out the incoming power with perfect one to one correspondence with the direction to the field point contributing that power. Any real lens or mirror system will fail to do this to some extent. When such failure occurs due to factors other than the expected diffraction effects, the optical system is said to have aberrations. Aberrations have an important place in the study of remote sensing tools mainly because a failure to establish the direction of incoming power precisely may lead to a failure to deduce remotely located material properties by further power analysis utilizing spectral component power distributions and polarization. Frequently, optical systems which have insignificant aberrations are expensive. In applications where aberrations may be harmless, some considerable savings in equipment investment could be realized by not purchasing high precision optics that are unnecessary.

The student is well advised to read Chapter 9 of Jenkins and White "Fundamentals of Optics" in order to acquaint himself with the various types of lens and mirror aberrations. It will tend to provide reasons why photographic cameras may contain from two to twelve different lens components in order to perform well enough for remote sensing purposes.

Frequently, one may see advertisements for optical apparatus that contain remarkable specifications of magnification, f/no. lenses for high speed photography, or high magnification factors for telescopes and microscopes. In general, these specifications will be met but the other essential features which are necessary to provide image fidelity are missing. In many cases, this editor had found that these wonderful specifications are useless because of the presence of extensive aberrations of all kinds which were not mentioned plus many defects which are not normally considered as aberrations such as, scattered light from poorly polished lens surfaces, multiple reflections from lens surfaces, mechanical instability of the optical structure, misalignment of the optical parts, and low functioning reliability. This is not to say that optical components and systems must necessarily be expensive to be useful. Modern mass production techniques have reduced the cost of high quality optical instruments with very little harmful aberration and there is no reason why further reductions could not take place in the future. Lens and mirror designs are now greatly aided by modern digital computers. Large numbers of designs can be tried to reduce the most harmful aberrations for a given application to acceptable minimums. In the days before high speed computers, lens design was truly an art. Many very good designs were discovered by trial, were patented, and now carry the inventors name or a trade name. One should always bear in mind, however, that "good" lens design for one application may not likely be aberration free for another application. The results of computer lens design indicates that for a fixed number of lens components, the reduction of one aberration usually results in the increase of another. The problems of aberration reduction is greatly simplified if the only place where the power in the image plane is to be used is on the optic axis. For a distant point source, the paraboloidal mirror is aberration free at that one position in the image plane. Figure 4-24 shows the resolution of a variety of optical systems near the optic axis. Please note that resolution as used here does not specify directly the quality of image formation since other factors are involved. In figure 4-25 is shown the properties of a perfect paraboloidal mirror for $1/2^{o}$, 2^{o}, and 4^{o} off the optic axis. Figure 4-26 shows the limiting blur due to diffraction effects for a perfect lens or mirror for visible wavelength radiation. There is little value in reducing aberrations much below the point where diffraction effects degrade image formation.

Additional References for Aberrations: Handbook of Military Infrared Technology, pp. 381ff; Manual of Photogrammetry, pp. 77ff.

5. HUMAN VISION

The human eye and the properties of sight play a vital role in remote sensing. As the main information channel to the brain from the external world, the eye accomplishes the following functions:

1. The unaided eye is the most commonly used sensor to make aerial or space observations of the earth directly.

2. Visual aids providing magnification to assist in direct observation are common and low cost.

3. Vision is used predominantly in observing and interpreting the output of other sensors which do not provide interpretation.

The properties of this visual channel, consisting of an optical system, a transducing system, and a transmission and data processing system, set certain limitations to its use as a search and interpretation mechanism. These properties are discussed below.

5.1. VISUAL ACUITY

Visual acuity is a measure of the ability of the eye to resolve closely spaced objects. It is usually expressed as the reciprocal of the visual angle of resolvable objects as measured in minutes of arc. Visual acuity depends on a number of factors including the types of objects observed and the available contrast and light levels. In general terms, however, the normal eye can resolve two lines separated by about a minute of arc. The eye is also able to detect a single line on a uniform background whose width subtends substantially smaller angles. Under conditions of good contrast, a single dark line subtending only 0.5 second of arc has been detected. The visual acuity is critically dependent upon the direction of the object relative to the foveal direction (or the "direction of normal visual attention"). The visual acuity for objects rapidly decreases with increasing angle off the foveal direction. Because human vision is almost an automatic process, this reduction in off-foveal acuity is not always appreciated. In any process involving search, however, the effect of this angular varying acuity will be an important factor in predicting realistic search success.

5.2. SENSITIVITY

The eye is sensitive to radiant energy lying roughly between 0.4 and 0.7 micron, with the peak response occurring at about 0.555 micron under normal lighting conditions. With a 2 mm pupil diameter, the eye forms an f/8 imaging system. The largest pupil size is about 8 mm, so that the lowest f number is f/2.

At normal light levels, the brightness of a surface as observed by the eye approximates a logarithmic function of the radiance of the light. The least perceptible brightness contrast is about 2%.

5.3. RESPONSE TO LOW LIGHT LEVELS

When an individual passes from light to dark surroundings, a period of dark adaptation of the eye occurs covering a period of 30 to 40 minutes, during which time the eyes become accustomed to the dark. At the end of this period, the sensitivity of the eye has increased by a factor of many thousands. When the individual reenters a lighted area, a process of light adaptation occurs, the time required being much shorter than for dark adaptation.

At very low levels of illumination, sensitivity to color does not exist. Various objects in the visual field are distinguished only by differences in brightness. The peak response of the eye occurs at about 0.555 micron under daylight conditions, but shifts to 0.510 micron under weak illumination. As a result of this shift, a red surface which appears to be equal in brightness to a blue surface at ordinary light levels, will appear much darker than the blue surface at low levels.

5.4. FLICKER

When light reaches the eye as a series of intermittent flashes, these are distinguished as individual flashes if the repetition rate is below a certain value called the critical flicker frequency. This frequency increases with the intensity of the light and the area of the field, ranging from 10 to 40 Hz.

5.5. RESPONSE TO COLOR

Along with the response to light level and image shape, the eye responds in a limited way to the spectral distribution in the received light to give the sensation of color. The spectral sensitivity of vision is limited principally by the absorption of the photochemical materials of the transducer elements. The saline solutions within the eye limit the transmission beyond this range to some degree also. Ultraviolet burns the outer eye structures. A certain amount of near infrared radiation might reach the retina without being perceived. Saline solutions begin to absorb infrared radiation just beyond the visible red spectral range. The tristimulus theory of color appears to be sufficient to predict most experimental results although there is considerable question as to mechanism. Three visibility functions, $\alpha(\lambda)$, $\beta(\lambda)$, $\gamma(\lambda)$, are assumed that mutually overlap but have a maximum at different wave lengths. The peaks were taken to be 440 mμ, 550 mμ, and 590 mμ by the 1920 Committee on Colorimetry, OSA. Thus any light sensation can be represented by three numbers each of which represents the degree of excitation of the respective hypothetical color receptor. The level of illumination

is represented by the sum of the excitations. If the three excitation levels are represented by A, B, and C and the level of light by A + B + C, then the color is represented by

$$\frac{A}{A+B+C}, \frac{B}{A+B+C}, \frac{C}{A+B+C}.$$

These three quantities are not independent since

$$\frac{A}{A+B+C} + \frac{B}{A+B+C} = 1 - \frac{C}{A+B+C}$$

so that one may express the chromaticity by using any two ratios as coordinates in a cartesian coordinate system or all three as the perpendicular distances from the sides of a triangle. The triangular plot is called the color triangle.

The tristimulus theory predicts that an infinite number of spectral distributions can provide the same color sensation.

$$A = \int_0^\infty \alpha(\lambda)\rho(\lambda)E_\lambda(\lambda)\,d\lambda$$

$$B = \int_0^\infty \beta(\lambda)\rho(\lambda)E_\lambda(\lambda)\,d\lambda$$

$$C = \int_0^\infty \gamma(\lambda)\rho(\lambda)E_\lambda(\lambda)\,d\lambda$$

A knowledge of the numbers A, B, and C and the functions $\alpha(\lambda)$*, $\beta(\lambda)$, $\gamma(\lambda)$ and $E_\lambda(\lambda)$ will not allow for a unique solution to find $\rho(\lambda)$.

However, these formulae suggest that the functional form of $\rho(\lambda)$ may be more accurately inferred if an object is viewed under a series of different spectral irradiance distributions.

Color matching is most successful with large adjacent areas. Color perception decreases with small areas, decreases at low light levels, and is altered by surrounding colors in the visual field.

5.6. SEARCH PROPERTIES OF THE EYE

The search properties of the eye are best analyzed by the use of the detection lobe theory. The essence of detection lobe theory is the calculation of the probability of detecting a target item by the employment of an imaginary volume in space representing the field of view of an

* This temporary use of $\alpha(\lambda)$ should not be confused with the spectral absorption of a surface. Here it represents a response property of the eye.

observer-device combination. This volume in space is so defined that a target within that volume is detected with certainty and a target outside that volume remains undetected with certainty. The detection lobe is moved about over a scene in some search pattern. A detection lobe is not only determined by the characteristics of the device and eye but also by the properties of the intervening medium and target-background characteristics.

For the daylight unaided eye the relation between the threshold contrast for detection, angular subtense, a, of the target, and angle, θ, off the foveal axis is given* as

$$C_t = 1.75\ \theta^{1/2} + \frac{19\theta}{a^2}$$

where C_t is the threshold apparent contrast in percent.

θ is in degrees, $0.8 \leqq \theta \leqq 90$.

a is in minutes of arc and is proportional to (L/r) where L is a target dimension and r is the range to the target.

C_t is considered constant for $0 \leqq 0 \leqq 0.8$.

Also, the apparent target-background contrast, C, is related to the inherent contrast, C_0, by the relation

$$C = C_0\ e^{-3.44r/V}$$

where r is the range to the target, V is the meteorological visibility.

By requiring the range r to be that value for which $C = C_t$ then

$$C_0 e^{-3.44r/V} = 1.75\,\theta^{1/2} + \frac{19\theta}{a^2}$$

which defines θ as a function of r. For visual angles from the fovea smaller than θ, detections are assured. Outside the angle θ, detections are considered improbable. Thus, $\theta = \theta$ (r) when considered a figure of rotation about the visual axis defines the detection lobe for the unaided eye in daylight. The intersection of the detection lobe with the terrain defines an area wherein target detections are assured provided that line of sight exists between the observer and the target.

*OEG Report 56, Search and Screening, Chapter IV, 1946.

If one were to view the terrain through a binocular, or other visual aid, having negligible optical loss, then the detection lobe is changed in two ways. First, the apparent angle, a, is increased by the angular magnification, M; and secondly, the detection lobe angle, θ, in object space appears to the eye to be at angle $M\theta$. Thus, the basic relation becomes

$$C_0 e^{-3.44r/V} = 1.75(M\theta)^{\frac{1}{2}} + \frac{19(M\theta)}{(Ma)^2}.$$

The unaided eye may be considered as a special case of a binocular where M = 1 and the field of view, (FOV), is 180 deg.

The detection lobe theory applies best to small objects in an otherwise expansive scene. Long thin objects such as roads and railroads require modified treatment. It is assumed also that strong color contrast is not present.

When applying this theory to binoculars, it must be remembered that the apparent field of view is less than 180° because of the field stop in the binocular. Hence, the detection lobe will be limited accordingly.

The same theory will apply as well to the search problem in finding particular items in photographic records made by a camera, infrared scanner, or radar. Obviously, the meteorological visibility term may not apply in these cases.

Whatever search pattern is assumed, one must be constrained to a glimpse rate of not more than 3 per sec. because of the finite time required for perception.

5.7. PERCEPTION

The transmission and data processing systems of the visual perception process are not well understood. An important property of these systems is that a rather long time is required to complete the perception process. The eye moves in a series of quick motions and fixations. Perception is not generally possible during the motion but occurs when the eye is held steady during a fixation upon an image. These glimpses last about 1/3 sec. Although the transducing action of image formation is faster than this, the perception time continues after the transducing action for a relatively long time.

The process of visual search and image interpretation will be limited in spatial and spectral resolution, processing time, and light level.

5.8. VISUAL AIDS AND STEREO-VISION

Visual aids are a highly specialized but important application of optics. The human eye is to form the final link in the optical chain. The optical system in front of the eye is designed to match the requirements of the eye in such a way that the eye forms the limiting part of the optical system. The pupil of the eye should become the aperture stop of the optical system.

Visual aids such as telescopes and binoculars are used for assisting in the direct observation of remote objects. Visual aids such as magnifiers, microscopes, and stereoscopes are used for the examination and the visual measurement of data obtained by other sensors.

5.8.1. BRIGHTNESS LIMIT.

An important property of visual aids is that no visual aid can improve the brightness of a scene for the eye. The perception of brightness depends upon the irradiance of the retinal area. If an area a × b shown in figure 5-1 is illuminated so that the radiance, L_r, is established, then the total luminous power received by the eye from that area is

$$P = a \times b \times L_r \times \frac{A}{R^2}$$

where A is the area of the entrance pupil of the eye and is constant for a particular light level.

Assuming no optical losses in the eye, all of that power is imaged in the area

$$\left(a \cdot \frac{f}{R}\right) \times b\left(\frac{f}{R}\right) = \frac{baf^2}{R^2}.$$

Thus, the irradiance of the retinal area is

$$\frac{PR^2}{baf^2} = \frac{abL_r AR^2}{R^2 abf^2} = \frac{L_r A}{f^2}.$$

The irradiance thus depends upon the illumination and the fixed properties of the eye. Provided that radiation is not lost on its way to the eye, the irradiance in the retinal area is constant, independent of range or apparent size of the area. It can be seen that the quantity, $\frac{A}{f^2}$, is the solid angle of the cone of light made available to the retina. Thus, the only way to increase the irradiance of the retina is to increase L_r if the image is to remain in focus or to increase the area of the entrance pupil A. The size of the pupil is automatically adjusted by the ambient light level. The most that a visual aid can be expected to do is to change the apparent size of the image on the retina without loss of brightness.

When the exit pupil of the visual aid coincides with the entrance pupil of the eye in both position and size, then all of the light passing through the visual aid aperture stop can pass into the eye. If the exit pupil is larger than the entrance pupil of the eye, then some of the light passing through the visual aid is not used. If the exit pupil is smaller than the entrance pupil of the eye, then the cone of rays entering the eye is less than would be available with the unaided eye so the brightness will be reduced below the unaided eye level.

This leads to different specifications for daylight and night binoculars. A prism binocular is simply a pair of astronomical telescopes using prism reflections to provide an erect image. See figure 5-2. It is not difficult to show that the angular magnification and the objective lens diameter is determined by the size of the exit pupil of the eye.

$$D = (ExP) \cdot AM$$

where D is the diameter of the objective lens, (ExP) is the diameter of the exit pupil, AM is the angular magnification apparent to the eye without loss of brightness.

If the dark adapted eye has an entrance pupil of 7 mm and the angular magnification is 7 then the diameter of the objective must be at least 49 mm if no loss of brightness is desired. A common night binocular is the seven power, fifty millimeter diameter objective type. A common daylight binocular is the 8 power, 35 mm objective which does not lose brightness for entrance pupils for the eye up to 35/8 mm in diameter.

5.8.2. MAGNIFIERS. The simplest magnifier for examining data such as photographic images is the single lens. The single lens is held approximately one focal length from the object being examined. The angular magnification, AM, is

$$AM = \frac{25 \text{ cm}}{f \text{ cm}}$$

The distance of 25 cm is the closest comfortable reading distance for the unaided eye. If a lens of a given shape is simply reduced in size proportionally in every dimension, the focal length is also reduced. Hence, a very small lens can be used for obtaining a large AM. However, a very small lens requires that the eye must be very close to the lens in order to see much of the object area. Ones nose will bump the object. The working distance is too small. Another serious problem arises from chromatic aberration. The focal length varies slightly for different wave lengths with corresponding different AM. The color fringes around details of the image interfere with the interpretation when AM values are greater than 4 for a simple lens. Using two lenses spaced apart a distance, d, so that

$$d = \frac{f_1 + f_2}{2}$$

tends to correct chromatic aberration. This magnifier is called a doublet. The working distance is zero when $f_1 = f_2$. The first lens must touch the object. There is still room for the nose due to the lens spacing. A millimeter scale reticle may be etched in the first lens surface so that small distance measurements may be made on photographs without parallax error. Some doublets compromise good chromatic correction to give more working distance.

A better magnifier for higher powers and greater working distances is known as the Hastings Triplet. It is made of three lenses cemented together. One lens is a different glass to accomplish chromatic correction.

5.8.3. STEREOSCOPES AND STEREOSCOPIC VISION. Stereoscopic vision is a mental process which combines two separate retinal images and makes them into one image using the parallax differences to judge range. The ability of the brain to do this is remarkable. When two vertical sticks are placed so that one eye sees only the front stick with the front stick hiding the back stick; the other eye will be able to see both sticks. The brain will be able to infer the proper position and number of sticks. The ability to resolve two sticks as separate and countable requires an angular separation of about 1 arc minute. However, a normal observer can detect parallax angles of 30 arc seconds, many can detect 10 arc seconds and some individuals have detected as little as 5 arc seconds.

The convergence of the eyes functions so as to bring the two retinal images to the appropriate positions on the retina to achieve fusion by the brain. This convergence also aids to some degree in "feeling" the range to a near object. For near objects the lens of the eye also must change focus to accommodate for the range.

When one examines a pair of stereo photographs at close range, the normal response of the eye is to converge and accommodate for the near range so that the stereo effect is not achieved. Without visual aids, one must not permit convergence but must allow short range accommodation which is contrary to normal eye response. Holding a piece of cardboard or paper between the two photographs and perpendicular to the plane of the photographs will help the eyes to accomplish this.

The use of two simple magnifiers allows the relaxed eye accommodation with non-convergence so that fusion of the images can occur more easily. This is the principle function of a stereoscope using simple magnifiers. However, data may be presented for stereo viewing which is so large that the correct distance between pupils cannot be arranged physically. The common 9" × 9" aerial camera film is such a case. Additional optics are required to bring the images of the two pictures to the correct separation. Both mirrors, in the form of periscopes, and prisms can be used to do this.

Normally, stereoscopes are applied to images which replicate the visual data of an actual environment. Abstracted data can also be presented to the brain using stereoscopic vision.

It is not difficult to draw stereo pictures and three dimensional graphs of data. Figure 5-3 illustrates such data plots. One plot is made first without regard to parallax. The second plot is made by horizontally displacing the points of the first plot by the distance between pupils minus the desired parallax distance corresponding to the z coordinate of the data point.

6. CLASSICAL PROPERTIES OF ELECTROMAGNETIC WAVES

The presence of electric or magnetic fields can only be known by the interaction of these fields with matter.

Static fields can be explored by test charges and magnets. The test results show evidence that these are vector fields.

Dynamic fields are explored by electric probes and magnetic loops and show that the following relations always hold:

$$\vec{\nabla} \times \vec{\mathcal{E}} = -\frac{\partial \vec{B}}{\partial t}, \quad \vec{\nabla} \cdot \vec{D} = \rho$$

$$\vec{\nabla} \times \vec{H} = \frac{\partial \vec{D}}{\partial t} + \vec{J}, \quad \vec{\nabla} \cdot \vec{B} = 0$$

$\mu \vec{H} = \vec{B}$, $\epsilon \vec{\mathcal{E}} = \vec{D}$ for isotropic media.

The Poynting vector, $\vec{S} = \vec{\mathcal{E}} \times \vec{H}$ watts/m^2,
measures the power density transported by the wave in the direction of $\vec{S}$.

From these equations one predicts the possibility of electromagnetic waves.

$$\nabla^2 \vec{\mathcal{E}} - \mu\epsilon \frac{\partial^2 \vec{\mathcal{E}}}{\partial t^2} = 0, \quad \nabla^2 \vec{H} - \mu\epsilon \frac{\partial^2 \vec{H}}{\partial t^2} = 0$$

where $v = \frac{1}{\sqrt{\mu\epsilon}}$ is the velocity of propagation.

In free space $\vec{\mathcal{E}}$ and $\vec{H}$ are at right angles to each other and both are at right angles to the direction of $\vec{S}$. The velocity in free space is 3×10^8 m/sec.

The direction of polarization is taken as the direction of the $\vec{\mathcal{E}}$ field.

All equations are linear differential equations so that the superposition principle holds. It is customary to express complicated wave forms as the superposition of many different plane waves having each its own characteristic frequency, polarization and direction of propagation.

The generation of radiation requires an accelerated rate of change of either a magnetic dipole moment or an electric dipole moment.

The power radiated is then

$$P = \frac{\omega^4 \mu \sqrt{\mu\epsilon}}{12\pi} |p_e|^2 \text{ watts,}$$

or

$$P = 173 \times 10^{-15} \nu^4 |p_e|^2 \text{ watts for free space,}$$

where p_e is the magnitude of a harmonically changing dipole moment.

p_e = size of the charge times the maximum displacement of the charge.
(Units are coulomb meters)

From this formula one can calculate the power radiated from a charged ball made to oscillate in space.

The corresponding formula for a magnetic dipole is

$$P = 160\pi^6 \left(\frac{R}{\lambda}\right)^4 J_0^2 \text{ watts,}$$

where R is the radius of a wire loop

λ is the wave length of the radiation

J_0 is the magnitude of a sinusoidal current in the loop.

The formula requires that $R << \lambda$ in order to apply.

Just as the emission of electromagnetic power requires an accelerated rate of change of dipole moment, so also the absorption of power cannot be accomplished without an accelerated rate of change of dipole moment but in the opposite way. Absorption is the reverse of emission. An emitter which can emit only polarized radiation can absorb only polarized radiation. An emitter which can emit only monochromatic radiation can absorb only monochromatic radiation having the same frequency.

The spectrum of useful electromagnetic radiation is very great (see figure 6-1). The technology for generating and receiving these various wavelengths is quite varied but the radiation differs only in scale and not in kind.

When electromagnetic waves are reflected by matter, the pattern of reflection depends upon the structure of the reflecting surface.

When electromagnetic waves fall upon a flat conducting plate, the electric field causes current to flow in the surface to the plate in the direction of the electric field and in time with it. In turn, the alternating current forms a radiating source of electromagnetic waves away from the plate. The total electric field at the conducting plate is then the sum of the incident and the reradiated electric fields. If the plate is a perfect conductor, the total

electric field strength at the surface will be zero at all times so that the reradiated wave must have the opposite phase to the incident wave at the surface and be of the same magnitude. When electromagnetic waves fall upon a dielectric plate, the electric field causes the material of the plate to polarize so that a dipole moment is induced in the material which changes in time but not always in phase with the electric field. Reradiation again occurs because of this changing polarization or varying dipole moment.

The total radiated field is the sum of the incident field and the reradiated field as before. In this case, however, the total field will be composed of an incident field, a reflected field and a transmitted field moving with a different velocity of propagation in the dielectric. The phase difference between the incident field and the reradiated field is caused by the inertia of heavier ionic parts of the solid dielectric and the elastic restoring forces which hold them in place. These heavier parts have natural resonant modes of vibration that are characteristic of the material so that the reflecting and transmitting properties can be used to help identify these materials.

6.1. ATOMIC AND MOLECULAR SPECTRA

The absorption and emission of electromagnetic radiation by atoms and molecules follows the same basic principles as was described. Since dielectrics and conductors are generally made of atomic and molecular aggregates, one should expect that the same basic processes should occur. The state of aggregation alters the characteristics of the emission or absorption.

The classical model of the atom as a planetary system with the electron spinning in orbit around the nucleus leads to a time varying dipole moment which would produce radiation until the electron falls into the nucleus. Observed evidence contradicts that prediction. In order to make a theory which does predict observed evidence, the quantum mechanical theory was invented. This new theory predicts that the dipole moment is not varying in time when the electron is in a particular set of possible states of motion about the nucleus. If some other kind of motion is induced, the dipole moment does change with time and radiation or absorption occurs until a nonradiating state of motion is reached. Further, the frequency which is radiated depends upon the difference in total energy of the initial and final states of motion,

$$\Delta Q = h\nu$$

where h is Planck's constant

$$h = 6.6256 \times 10^{-34} \text{ Joule seconds.}$$

Of course, this frequency is the central frequency of a complex wave form. The wave form is by no means an endless simple harmonic wave. The transmitted wave form is a finite wave train which may be produced in as short a time as 10^{-8} seconds.

Since the change in the total energy of motion of the electron determines the frequency of absorption or emission, an energy level diagram is used to describe the total energy of the nonradiating states of motion. The energy levels for nonradiating states of motion are represented by the horizontal lines show in figure 6-2. These energy levels are sorted into groups. The required change in dipole moment occurs when a change in energy is made from one level in one group to a level in an adjacent group. Every kind of atom has such an energy level diagram appropriate to it.

If we now take two atoms of the same kind and push them together, the energy level diagram for the combination changes so that the radiative properties of a molecule are not the same as that of the isolated atoms. The molecular spectra of compounds of dissimilar atoms do not resemble the spectra of the isolated components. In addition to the absorption and emission due to changes in electronic motion in a molecule, the component ions of the molecule can also form a changing dipole moment by spinning and by vibrating against each other. These motions contribute to the total energy of the molecule and must be accounted for in the energy level diagram. The energy difference between one nonradiating spinning motion and the next of higher energy is usually very small. The energy difference for vibrational changes is larger than the spinning (or rotational) and the energy difference for electronic motion changes is usually the largest of all. The appearance of the energy level diagram for molecules is banded and the spectra of molecules which are taken on a spectrometer also consequently appear in bands so that molecules are said to produce band spectra. These band spectra are a consequence of the molecular species and therefore provide a means of identification.

Frequently, in molecular spectra, certain chemical radicals produce vibrations which are characteristic of the radical — such as O-H at 2.8 μ; N-H at 3.03, 6.12, 6.46 μ; C-H at 3.4, 6.81 μ; C-O with bands at 9, and 13.47 to 14.20 μ. These features are widely used in infrared spectroscopy for the identification of species and in the control of chemical reactions.

A further state of aggregation introduces still more changes in the absorption and emission spectra. As atoms or molecules are combined to form crystalline solids, the spectra become still more broadly banded. A crystal of a particular compound may be considered to be a giant molecule with many identical small molecules as components. Thus, solids tend to lose much of the spectral character of the constituents of which it is made. Nevertheless, significant differences still remain between spectra of solids because the nature of the constituents and the crystalline forms continue to govern the motions which cause a change in dipole moment. Liquids have spectra which show the consequences of a different kind of aggregation. Thus, liquid water and solid water (ice) should not have the same spectral properties although one should expect some degree of similarity. Once a crystal of some compound has aggregated a million or so molecules, the crystal behaves essentially the same spectrally as any larger

amount of the material for short wavelengths. A new atom or molecule added to the outside of the crystal is so far away from the majority of those already there that its effect is not great enough to be noticed. For long wavelengths, the gross electromagnetic properties of the aggregate and the geometrical form determine the spectral properties.

7. A REVIEW OF BLACKBODY RADIATION LAWS

There is a natural division or grouping of noncoherent radiant power sources which depends upon the nature of the wavelength distribution of the emitted energy. One type of source emits radiation over a very broad and continuous band of wavelengths. A plot of its spectral power, P_λ, versus wavelength is a smooth curve which usually passes through only one maximum. This type is called a continuous spectrum source, or more briefly, a continuous source. Another type of source is one which radiates strongly in some relatively narrow spectral intervals, but, in other wavelength intervals, the source does not radiate at all. A plot of P_λ versus wavelength reveals a series of emission bands or lines. The curve is discontinuous and the source is called a discontinuous spectrum source, a line source, or a band source. In this section we will deal only with the continuous spectrum sources.

7.1. THERMAL RADIATION

All objects are continuously emitting and absorbing radiation. The emitted radiation results from the acceleration of electrical charges within the material. The interactions of these charges in solid bodies are very complex. Thus, it is very difficult to apply electromagnetic theory to the electrons in a solid to explain such things as the observed variations in the spectral emissivity of real solids. Instead, thermodynamics has been utilized to predict and explain the radiative behavior of such bodies. This mode of attack has led to the very useful concept of the ideal radiator. Certain radiating surfaces exist whose radiation characteristics are completely specified if their temperature is known. These surfaces, which are continuous spectrum sources, are known as ideal thermal radiators, or "blackbodies." The following section shows why a so-called blackbody is an ideal radiator.

7.2. KIRCHHOFF'S LAW

Consider a material object S located within the walls of an evacuated cavity which is maintained at a uniform temperature, T. No matter what material S is made of and whatever its initial temperature, it will finally come to and remain at temperature T. As long as S and the walls of the cavity differ in temperature, there will be a net flow of radiant energy toward the cooler of the two. Eventually the stream of radiation will be the same in all directions and it will be the same in all enclosures which are at the same temperature no matter what materials are used to make them.

At thermal equilibrium, then, the power per unit area radiated by S must equal the power per unit area absorbed by it.

$$M = \alpha E,$$

where M ≡ the radiant emittance or exitance, or power per unit area emitted by S,

α ≡ the absorptance of S,

E ≡ the power per unit area striking S.

Since α cannot be greater than unity, the emitted power is a maximum for a perfect absorber, a so-called blackbody. For such ideal radiators $\alpha = \epsilon = 1$. Where ϵ is the ratio of M emitted by a surface to M emitted by a black surface at the same temperature. For other less efficient absorbers, the emissivity is correspondingly lower so that in all cases

$$\epsilon = \alpha \tag{1}$$

It turns out that this is also true for each spectral component of the radiation, that is, $\epsilon(\lambda) = \alpha(\lambda)$. This can be proven using the second law of thermodynamics. Equation (1) is a statement of Kirchhoff's law. The relationship

$$P_i = P_\alpha + P_\rho + P_\tau, \tag{2}$$

where P_i is the incident radiant power, P_α, the absorbed power, P_ρ, the reflected power, and P_τ, the transmitted power, is another statement of the conservation of energy. By dividing both sides by P_i, this relationship may be rewritten as

$$\alpha + \rho + \tau = 1, \tag{3}$$

where α, ρ, and τ have the usual meaning.

For an opaque body $\tau = 0$, so that equation (3) becomes

$$\alpha = 1 - \rho, \tag{4}$$

indicating that surfaces of high reflectance are poor emitters!

7.3. THE STEPHAN-BOLTZMANN LAW

In 1879, T. Stephan found experimentally that the radiant emittance of an object was proportional to its absolute temperature raised to the fourth power. Five years later, L. Boltzmann deduced the law theoretically by utilizing thermodynamic consideration. In terms of radiant emittance, we can write the Stephan-Boltzmann Law as

$$M_{bb} = \sigma T^4, \tag{5}$$

where M_{bb} is the power per unit area radiated into a hemisphere by a perfect radiator, or blackbody; σ is known as the Stephan-Boltzmann constant and T is the absolute temperature. The value of σ has been found to be 5.6687×10^{-8} watts m^{-2} $(\text{deg K})^{-4}$.

Equation (5) provides a simple means for computing the radiant emittance or radiance of any blackbody source for which the temperature can be determined. This is not the limit of its usefulness. It turns out that many common surfaces are "grey" or nearly so. In other

words, their emissivity (not their radiant emittance) is independent of wavelength. Thus, a relationship very useful in computing the radiant emittance of a real body having an emissivity ϵ is

$$M = \epsilon \sigma T^4. \tag{6}$$

This relationhip requires the use of true temperature. True temperature can be determined in some cases with a thermometer or thermocouple directly.

7.4. PLANCK'S RADIATION LAW

There are several conceptual models which can form the basis for a deductive justification of Planck's Radiation Law. It is interesting to note that conceptually these models are quite diverse. They vary from the pure classical through the hybrid classical-quantum mechanical to quantum statistical. We will here outline only one way to view the physical processes described by the law.

If the photon is considered a free particle then the number of discrete energy levels available for photons in the frequency interval ν to $\nu + d\nu$ when they are enclosed in a cavity at thermal equilibrium can be shown to be

$$dz = \frac{8\pi V \nu^2}{c^3} d\nu ,$$

where V is volume of the cavity, ν is frequency of the radiation, and c is the velocity of light in vacuum. The energy density per unit frequency interval can be obtained by assigning a mean energy $\overline{Q}_\nu$ to each corresponding energy level. Thus,

$$w_\nu = \frac{8\pi \nu^2}{c^3} \overline{Q}_\nu$$

calculation of $\overline{Q}_\nu$ implies that a statistical distribution function for the probability that the energy states at ν are filled is known. This distribution function is known and was deduced independently by Bose and Einstein for particles which do not obey the quantum conditions implied by the Pauli exclusion principle. The form of their distribution is

$$g(Q) = \frac{C}{e^{-\frac{Q_0}{kT}} e^{\frac{Q}{kT}} - 1}$$

Here C is a normalizing constant, Q_0 a reference zero point energy, and Q is the energy of the particle. k is Boltzmann's constant and T is the absolute temperature of the environment of the particle. For the applications which we consider here $Q = h\nu$ is 10 to 100 times greater

than kT and g(Q) may be approximated by the classical Maxwell-Boltzmann distribution, i.e.,

$$g(Q) = Ce^{-\frac{h\nu}{kT}}$$

with this approximation we can calculate the average energy associated with the states between ν and $\nu + d\nu$. Thus,

$$\overline{Q}_\nu = \frac{\sum_{n=1}^{N} nh\nu e^{-\frac{nh\nu}{kT}}}{\sum_{n=1}^{N} e^{-\frac{nh\nu}{kT}}}$$

$$\overline{Q}_\nu = \frac{-\frac{d}{dkT}\sum_{n=1}^{N} e^{-\frac{nh\nu}{kT}}}{\sum_{n=1}^{N} e^{-\frac{nh\nu}{kT}}}$$

$$\overline{Q}_\nu = \frac{h\nu}{e^{\frac{h\nu}{kT}} - 1}$$

The radiation energy density thus becomes

$$w_\nu d_\nu = \frac{8\pi h\nu^3}{c^3} \frac{d\nu}{e^{\frac{h\nu}{kT}} - 1},$$

which is the basic form of Planck's Law. Generally one measures not the energy density, but the radiant flux density or radiant emittance into a hemisphere (2π steradians) emanating from a flat blackbody radiator. This more useful form can be obtained by multiplying w_ν by the factor $\frac{c}{4}$. Doing this we get the radiant emittance for unpolarized radiation

$$M_\nu(\nu,T)\,d\nu = \frac{c}{4} w_\nu d\nu = \frac{2\pi h\nu^3 d\nu}{c^2 e^{\frac{h\nu}{kT}} - 1}.$$

Utilizing $\lambda\nu = c$, M can be written in terms of wavelength

$$M_\lambda(\lambda)\,d\lambda = \frac{2\pi c^2 h}{\lambda^5}\,\frac{d\lambda}{e^{\frac{hc}{\lambda kT}} - 1} \tag{7}$$

This is the most useful form of Planck's Law for practical use. Equation (7) has been experimentally verified to such an extent that its validity is unquestioned and it is assumed to completely describe all observable phenomena of blackbody radiation.

7.5. THE RAYLEIGH-JEANS' AND WIEN'S RADIATION LAWS

For the case of $ch/\lambda kT << 1$, $e^{ch/\lambda kT}$ can be expanded in series and to first order Planck's Law becomes

$$M_\lambda(\lambda T)\,d\lambda = \frac{2\pi c}{\lambda^4} kT d\lambda \qquad \lambda >> \frac{ch}{kT}$$

which is the Rayleigh-Jean's Law.

When $ch/\lambda kT >> 1$, $e^{h/\lambda kT} >> 1$ and Planck's formula reduces to

$$M_\lambda(\lambda T)\,d\lambda = \frac{2\pi \nu^2 h}{\lambda^5} e^{-\frac{ch}{\lambda kT}} d\lambda \qquad \lambda << \frac{ch}{kT}$$

which is Wien's radiation Law.

Theoretical Determination of the Stephan-Boltzmann Constant

If Planck's Law is integrated over all frequencies, the Stephan-Boltzmann equation should result. Thus,

$$M(T) = \frac{2\pi h}{c^2} \int_0^\infty \frac{\nu^3 e^{-\frac{h\nu}{kT}}}{\left(1 - e^{-\frac{h\nu}{kT}}\right)} d\nu$$

$$= \frac{2\pi h}{c^2} \int_0^\infty e^{-\frac{h\nu}{kT}} + e^{\frac{2h\nu}{kT}} + , \ldots \nu^3\, d\nu$$

$$= \frac{2\pi^5 h}{15c^2}\left(\frac{kT}{h}\right)^4 = \sigma T^4$$

$$\sigma = \frac{2\pi^5 k^4}{15c^2 h^3} = 5.672 \times 10^{-5} \frac{\text{erg}}{\text{cm}^2\ \text{deg}^4\ \text{sec}}.$$

σ is found to agree very well with the experimentally determined value thus confirming the validity of Planck's Radiation Law.

7.6. WIEN'S DISPLACEMENT LAW

The Planck radiation formula shows that the spectrum of the radiation shifts toward shorter wavelengths as the temperature of the radiator increases. This is a common observation with practical sources. A metal, when heated, radiates first at very long wavelengths in the infrared. As it becomes hotter it begins to radiate more at shorter and shorter wavelengths, first in the red end of the visible spectrum and finally more and more toward the blue. Thus, the metal appears successively deep red, red, orange, yellow, and finally (when sufficient blue and violet light are being radiated), white. The Wien displacement formula relates the wavelength of maximum emission to the temperature of the radiator. This relationship can be obtained by setting the derivative of $M_\lambda(\lambda T)$ with respect to λ equal to zero and solving for $\lambda = \lambda_m$. After simplification this gives

$$\left(1 - \frac{hc}{5\lambda_m kT}\right)\exp(hc/\lambda_m kT) = 1.$$

Replacing $hc/\lambda_m kT$ by x, we have

$$\left(1 - \frac{x}{5}\right)e^x = 1.$$

Using an iterative method or a table of natural logarithms we find that x = 4.965. Thus

$$\frac{hc}{k\lambda_m T} = 4.965.$$

Introducing numerical values for h and k as $h = 6.6252 \times 10^{-34}$ joule sec, $k = 1.3805 \times 10^{-23}$ joule/deg K we find that

$$\lambda_m T = 2893\ (\mu\text{m})\ \text{degrees}.$$

This is the Wien displacement formula.

7.7. GREYBODIES

Non blackbody radiators, i.e., real materials may be characterized by a spectral emissivity which is also in general a function of temperature. However, this temperature dependence is appreciable only for extremely high temperature. For our purposes here, the spectral emissivity may be operationally defined as

$$\epsilon(\lambda) = \lim_{\Delta\lambda \to 0} \frac{\int_{\lambda-\Delta\lambda}^{\lambda+\Delta\lambda} M_{m,\lambda}\, d\lambda}{\int_{\lambda-\Delta\lambda}^{\lambda+\Delta\lambda} M_{bb,\lambda}\, d\lambda}$$

where $M_{m,\lambda}$ is the spectral radiant emittance of the real radiator and $M_{bb,\lambda}$ is the spectral radiant emittance of a blackbody radiator at the same temperature.

A Greybody is defined as a radiator which has the same spectral emissivity at all wavelengths. A selective radiator on the other hand has different emissivities over different wavelength intervals.

If θ is the angle with the normal of the radiating surface, then a further refinement is to introduce a dependence of the emissivity on θ. The blackbody radiation from a unit area into the solid angle $d\Omega$ at angle θ is $wc/4\pi \cos\theta\, d\Omega$. Hence, for a non-blackbody radiator one has $\epsilon(\lambda,\theta)(w_\lambda c/4\pi)\cos\theta\, d\Omega$. A material which radiates the same angular distribution as a blackbody radiator, i.e., so $\epsilon(\lambda,\theta)$ is independent of θ, is said to obey Lambert's Law. Since a $\cos\theta$ factor appears in converting the actual area of the radiator to the projected area normal to the line of sight of the observer, a Lambertian radiator has the same radiance in any direction.

7.8. BLACKBODY TABLES

For very careful work it is necessary to use tables of blackbody functions with their additional precision and attendant difficulties. Since each set of tables is a thick book in itself, only references to these works are given here.

S. A. Golden, Spectral and Integrated Blackbody Radiation Functions, Research Report 60-23, Rocketdyne Division, North American Aviation, Inc., Canoga Park, Calif. (1960).

For discussing radiometric quantities in extant literature, the older set of symbols are used where $W \equiv M$, $N \equiv L$, $u \equiv w$.

In Golden, table I provides $W_\lambda/W_{\lambda max}$ and $W_{0-\lambda}/W$ and their first derivative as functions of $c_2/\lambda T$.

The intervals are as follows:

$c_2/\lambda T$	$\Delta c_2/\lambda T$
0-2	0.01
2-5	0.02
5-10	0.05
10-25	0.10
25-50	0.20

Table II provides $W_{\lambda max}$ and W as functions of T from 0°K to 10,000°K in 10°K intervals. Values of the radiation constants are

$$c_1 = 2\pi hc^2 = 3.7413 \times 10^{-5} \text{ erg cm}^2 \text{ sec}^{-1}$$

$$c^2 = hc/k = 1.4388 \text{ cm } ^oK$$

In M. Pivovonsky and M. Nagel, Tables of Blackbody Radiation Functions, Maxmillan; N. Y. (1961) table I is a tabulation of (1) N_λ vs. λ and T; (2) the ration $N_\lambda(\lambda T_1)/N(0.56\ \mu, T_1)$; (3) $N_{0-\lambda}/N$. These are tabulated to five significant figures for $\lambda = 0.2\mu$ to 0.590μ in 0.005μ intervals and between 0.590μ and 1.2μ in 0.01μ intervals from 800 to 40,000°K in intervals varied to meet the needs of the range.

Table II continues Table I for $\lambda = 1.1\mu$ to 1100μ at temperatures from 20 to 13,000°K with four-figure accuracy.

Table III includes: (1) $N_\lambda/N_{\lambda max}$; (2) a restatement of the wavelength ratios of Table I and II; and (3) a function for computing derivatives of the Planck function. These are plotted for λT from 0.01 to 0.99μ°K. Procedures for evaluation at higher values of λT are given.

Table IV has: N, $N_{\lambda max}$, λ_{max} for temperatures from 1000° to 2500°K at 2°K intervals, from 2500 to 5500°K in 5°K intervals and from 5500° to 10,000°K in 10°K intervals.

Table V repeats Table IV but for wavenumbers, and Table VI repeats Table IV for reciprocal temperatures.

Table VII gives luminance from 800° to 1796°K in 4°K intervals and a table of luminance and chromaticity coordinates.

Table VIII is a temperature correction table — for revised physical constants.

In M. Czerny and A. Walther, Tables of the Fractional Function of the Planck Radiation Law, Springer-Verlag, Berlin (1961), $W_{0-\lambda}/W$ is plotted vs. T/c_2. The tables are independent of c_2. It also includes the first and second derivatives of this function and

$$\frac{W_{0-\lambda}}{W} + \frac{1}{4}\frac{T}{c_2}\frac{d}{d\lambda}\left(\frac{W_{0-\lambda}}{W}\right)$$

and its first derivative.

Other tables in print are as follows:

Parry Moon, J. Opt. Soc. Am. 38, 291 (1948).

A. N. Lowen and G. Blanch, J. Opt. Soc. Am. 30, 70 (1940).

E. Jahnke and F. Emde, Tables of Functions, Dover, New York (1945).

A. G. DeBell, Rocketdyne Research Report 59-32, Rocketdyne Division of North American Aviation, Inc., Canoga Park, Calif. (1959).

7.9. BLACKBODY CURVES

Figures 7-1, 7-2, 7-3, and 7-4 provide information about the spectral distribution of blackbody radiation, and are self explanatory. Figure 7-5 is a plot of the fractional radiant emittance below a given point as a function of the variable λT.

An especially useful curve is a log-log plot of the Planck equation. The shape of the curve is identical for all T and need only be shifted along the line representing the Wien displacement Law. Such curves are shown in figures 7-3 and 7-4. The straight lines are the "sliding lines." Every blackbody curve for any temperature can be obtained by moving a curve of the same shape along this line. Thus a "do-it-yourself" slide rule can be constructed by putting an overlay on this figure, tracing the curve and the line, and placing an index marker at 6000°F, the temperature of the top curve. Then by keeping the lines overlapped and setting the index marker at the desired temperature, the template becomes the blackbody curve for that temperature.

Summary of Equations and Constants

$$M \equiv W = \sigma T^4$$

$$M_\lambda \equiv W_\lambda = 2\pi c^2 h \lambda^{-5} (e^{hc/kT\lambda} - 1)^{-1}$$

$$M_\nu \equiv W_\nu = 2\pi c^{-2} h \nu^3 (e^{h\nu/kT} - 1)^{-1}$$

$$L_\lambda \equiv N_\lambda = W_\lambda/\pi \text{ (only Lambertian surfaces, e. g., blackbody)}$$

$$w_\lambda \equiv u_\lambda = 4W_\lambda/c$$

$$w_\lambda \equiv u_\lambda = 2W_\lambda/c$$

$$\lambda_{max} T = 2897.9 (\mu {}^\circ K)$$

$$h = 6.6252 \times 10^{-34} \text{ w sec}^2$$

$$\pi = 3.1416$$

$$c = 2.99793 \times 10^8 \text{ m sec}^{-1}$$

$$\sigma = \frac{2\pi^5 k^4}{15 c^2 h^3} = 5.6686 \times 10^{-12} \text{ watt cm}^{-2} ({}^\circ K) \text{ (Stephan-Boltzmann constant)}$$

$$k = 1.38047 \times 10^{-16} \text{ erg}({}^\circ K)^{-1} \text{ (Boltzmann's constant)}$$

$$e = 2.71828 \text{ (base of Napierian logarithms)}$$

$$h/k = 4.079 \times 10^{-11} \text{ sec } {}^\circ K$$

Quantity	c_1	c_1
Energy density	$8\pi ch$	4.99 joule μ^4 m^{-3}
Emittance	$2\pi c^2 h$	3.7413×10^8 w μ^4 m^{-2}
Radiance	$2c^2 h$	1.19×10^8 w μ^4 m^{-2} sr^{-1}

$$c_2 = ch/k = 1.4388 \times 10^4 \ \mu {}^\circ K$$

7.10. THE GENERAL ELECTRIC BLACKBODY SLIDE RULE

This rule, designated GEN-15C, is available from the General Electric Company, 1 River Road, Schenectady, New York; it costs about one dollar. Calculations which can be made on the rule are as follows (see figure 7-6):

1. Conversions of temperatures among Celsius, Kelvin, Fahrenheit, and Rankine by setting the temperature on one scale and reading it on another — scales ABKL.
2. Multiplication by the use of standard C and D log scales.

3. Total blackbody radiant emittance by setting the temperature of the blackbody source on a temperature scale and reading on the E scale (w cm^{-2}). An emissivity scale associated with the E scale permits direct calculation for graybodies; read the value on the E scale under the appropriate emissivity.

4. Incremental blackbody radiant emittance $W_{\Delta\lambda}$ at maximum. The power density for a 1μ bandpass can be read directly from the $W_{\lambda max}$ or F scale.

5. The ratio of W_λ at any wavelength λ to that at λ_{max}, $W_\lambda/W_{\lambda max}$. The temperature is set on a temperature scale; then $W_\lambda/W_{\lambda max}$ is read from the G scale opposite the desired λ on the H scale. Thus, one can find $W_{\lambda max}$ for a given λ on the $W_{\lambda max}$ scale and then calculate the value of W_λ at any wavelength on the $W_\lambda/W_{\lambda max}$ scale.

6. The blackbody radiation in any spectral interval. Set the temperature scale at the appropriate temperature. Then on the $W_{0-\lambda}/W_{0-\infty}$ or J scale read the percentage radiation that lies below a particular wavelength λ_1 (on the I scale). Do the same for λ_2, and subtract.

7. Conversion of range in nautical miles to range in centimeters with the aid of a straight edge, and vacuum calculation of irradiance. These can be made with the QRST scales.

8. Conversion from w $in.^{-2}$ to Btu ft^{-2} hr^{-1}.

9. Number of photons sec^{-1} cm^{-2} from a blackbody at index temperature.

Useful constants and other combinations of these calculations are also available.

List of References

1. A. Sommerfeld, Thermodynamics and Statistical Mechanics, Academic Press, New York, 1956.
2. Handbook of Physics, E. Condon and H. Odishaw, Eds., McGraw Hill Book Company, 1958.
3. P. Kruse, et al., Elements of Infrared Technology, John Wiley and Sons, Inc., New York, 1963.
4. M. Holter, et al., Fundamentals of Infrared Technology, MacMilliam Company, New York 1962.
5. W. Smith, Modern Optical Engineering, McGraw Hill, New York, 1966.
6. Handbook of Military Infrared Technology, William L. Wolfe, Ed., Office of Naval Research, Department of the Navy, Washington, D. C., 1965.

8. BASIC PRINCIPLES OF RADIATION DETECTORS

In order to determine the amount of received energy in a given time interval, some form of material must absorb the radiant energy at the receiver location so that the condition of that material is altered in a way which can be monitored. The material which is to absorb this energy and be monitored is called the detector. (Types and characteristics of detectors are listed in Table 8-1.) The term detector has a number of meanings. In this context it refers to a transducing action changing the radiant form of energy into a stored localized form of energy. In a radio circuit, a "detector" performs the function of a demodulator isolating the modulation signal from the carrier signal produced by a remote transmitter. (The antenna wire performs the function of absorbing the radiant energy to change the radiant energy into a localized form for monitoring.) In human perception, to detect implies a recognition or discovery process. As used here, a detector does not discover, recognize, or demodulate. A detector transduces energy for monitoring purposes.

An obvious way to transduce radiant energy into local energy for monitoring is to change absorbed radiant energy into heat so that the material displays a rise in temperature. Some temperature-dependent property of the material can be monitored by local measurement systems to reveal the occurrence of absorbed radiation. This kind of detector is called a thermal detector. An ordinary liquid thermometer with a blackened bulb is one such transducer. Since there is nothing mysterious about such detectors save for the elaborateness of design no further discussion of these are needed. Another class of detectors, called quantum detectors, produce a monitorable change in property after absorption of radiant energy but a measureable temperature rise is not required. Such detectors as photo-emissive tubes, photographic film, silicon solar cells, television vidicon camera tubes, and many high speed infrared detectors are quantum detectors. In order to understand the operation of these detectors, it will be necessary to examine the interaction of electromagnetic radiation with atomic matter in more detail.

8.1. PARTICULATE NATURE OF RADIATION

Electromagnetic radiation has been described as a smooth wave motion. This served well for describing reflection from mirrors and transmission through lenses. However, if we want to know in more detail about the interaction of atoms with EM radiation the smooth wave motion must be modified to allow for the particle or bullet like effects of electromagnetic radiation.

Einstein in an experiment performed in 1905 showed that EM radiation had the properties of a flow of small indivisable quantities of energy called photons. In his experiment, he found that if EM radiation of sufficiently high frequency (or short wavelength) falls on the surface of

TABLE 8-1. INFRARED DETECTOR CHARACTERISTICS

Detector	Operating Temperature °K	Mode of Operation*	Impedance Megohms	Time Constant μ seconds	Type
PbS	300	PC	1-10	150-500	Polycrystalline film
	195	PC	5-20	~900	"
	77	PC	5-50	500-3000	"
PbSe	77	PC	2-100	10-30	"
PbTe	77	PC	50-1000	1-10	"
InSb	77	PV	.001-.01**	<1	Single crystal
	77	PC	.001-.01	<1	"
	195	PC	10-100	<1	"
	300	PEM	4.5-90	<1	"
Ge:Au (P-type)	77	PC	0.1-5	<0.1	"
Ge:Au (N-type)	77	PC	1-40	~50	"
Ge:Hg	35	PC		<1	"
Ge:Cu	15	PC	10^{-5}-10^{-3}	<1	"
Ge:Cd	25	PC		<1	"
Ge:Zn (1)	4	PC	~2	<1	"
Ge-Si:Au (1)	21	PC	170	<1	"
	50	PC	0.5	<1	"
Ge-Si:Zn (2)	21	PC	70	<1	"
	50	PC	30	<1	"
Te	77	PC	5×10^{-4}-2×10^{-3}	~60	"
InAs	300	PV	25×10^{-6}**	~2	"
CdS	300	PC	~1000	12,000-300,000	"
CdSe	300	PC	~100	6000-60,000	"
Se-SeO	300	PC	~10,000	2000	Polycrystalline film
Thermistor	300	temperature dependence of resistance	3	1500	Thermistor flake
Golay cell	300	gas expansion	-	20,000	Pneumatic cell

*PC - Photoconductive
PV - Photovoltaic
PEM - Photoelectromagnetic

**Dynamic Impedance

some metals, electrons are ejected from the surface.

Two significant observations were made during this experiment.

1. When the irradiance of the incident EM radiation is increased with the same spectral distribution, more electrons are given off but the velocity range with which they leave is unchanged.

2. Electromagnetic radiation causes the ejection of an electron from the metal surface only when the frequency of the EM radiation exceeds some threshold frequency ν_c.

Einstein's explanation for this using elementary radiation particles is as follows: A photon strikes the metal and is completely absorbed. It disappears and gives up a definite amount of energy $h\nu$ to one of the free electrons in the metal. Thus the electron's kinetic energy is increased by the amount equal to the energy of a photon. Before the electron can escape from the surface it must have at least enough energy to overcome the binding energy of the metal for the electron, $h\nu_c$. In addition, some excess energy over the critical value may be required because the electrons at some depth below the surface which absorb a photon are not able to escape without suffering collisions with atoms in the metal and losing some of their kinetic energy. The energy lost in this way varies from one electron to another, thus they emerge with a certain range of velocities. Increasing the irradiance simply increases the rate at which this happens but does not change the velocity range.

The important consequence of this experiment is that it provides evidence that energy of an EM wave is delivered only in localized regions — not spread evenly over the wave front.

One may conclude that the smallest wave train having a central frequency, ν, generated by a source must have sufficient amplitude and duration so that the total energy, Q, carried by that wave train is equal to $h\nu$; otherwise, a wave train cannot be radiated. Moreover, if that wave train is intercepted by matter, then that smallest wave train delivers all of that energy, Q, in one small spot but at some unpredictable position along the wave front. The probability that the energy will be delivered within some small area of wave front is proportional to the square of the elemental wave amplitude at that position. Now if enormous numbers of such elemental wave trains are superposed to make a sensible stream of radiation, the number of elemental wave trains which do deliver their energy to a given macroscopic area will be, with high precision, proportional to the square of the resultant wave amplitude. However, if the number of wave trains delivering energy per second is reduced radically, either by reducing the receiving area to atomic sizes or by reducing the generation rate, one will find that the fluctuation of the number of deliveries from the average will become very noticeable and measureable. It is apparent, then, why the interaction of electromagnetic radiation with atomic receiving areas requires a description of radiation which utilizes this probabilistic

energy delivery property of elemental wave trains where $Q = h\nu$ while the treatment of gross power for large areas and high power levels are adequately described by the mathematically smooth wave motion concept.

Because the probabilistic delivery of elemental quantities of energy is more of interest in discussing atomic absorption and the interference and diffraction properties of the wave motion are not of direct interest for this purpose, a particle-like description of radiation is commonly used. The probabilistic energy delivery is thought of as due to an elemental "bullet" of radiation shot out from the source. Each elemental bullet is called a photon. Therefore, one can picture EM radiation as a stream of photons which interact with solids on a microscopic level. [Note how photon rates from blackbodies can be computed using the radiation slide rule.]

8.2. WAVE NATURE OF PARTICLES

Now let us take a look at the basic theory of Atomic Physics from the classical quantum-mechanical point of view.

DeBroglie, a French Scientist, pondered the wave and particle nature of EM radiation and probably asked himself whether nature was symmetrical in this respect—i.e., do particles have wavelike properties? In 1924 he postulated that material particles as well as light quanta have associated with them an intrinsic frequency ν and wavelength, λ, where the wavelength and frequency can be expressed as

$$\lambda = \frac{h}{p} = \frac{h}{mv} \tag{1}$$

$$\nu = Q/h \tag{2}$$

Experiments by Davisson and Germer showed that their relationships held for the electron which was thought to be only a particle with some electric charge. The dual characteristics of an electron, i.e., that it is both a wave and a particle, leads to an interesting result. Consider the circular orbits of an electron about the nucleus of a hydrogen atom. There should be an integral number of electron wavelengths about the orbit. Thus,

$$2\pi r = n\lambda \tag{3}$$

where $n = 1, 2, 3, 4, \ldots$

λ = electron wavelength

r = orbit radius

From DeBroglie's relationship ($\lambda = h/mev$)

$$2\pi r = n\left(\frac{h}{mev}\right) \tag{4}$$

Square both sides of this expression

$$r^2 = n^2\left(\frac{h^2}{m_e^2 v^2}\right) \tag{5}$$

Since the centripital force F_c must equal the force of electrical attraction F_e

$$F_e = \frac{q_c^2}{r^2} = \frac{m_e v^2}{r} = F_c \tag{6}$$

Therefore, $v^2 = q_e^2/m_e r$ and if we substitute this into equation (5) we have

$$r = \left(\frac{h^2}{4\pi^2 q_e^2 m_e}\right) n^2 \tag{7}$$

This gives us an expression for the radius of the electron orbit in terms of some constants and the quantum number, n, which is the integral number of wavelengths in the orbits. Putting in the numbers we have ($h = 6.63 \times 10^{-27}$ erg sec, $q_e = 4.8 \times 10^{-10}$ stat coulombs, and $m_e = 9.11 \times 10^{-28}$ gm).

$$r = (0.53 \times 10^{-8})\, n^2 \text{ centimeters}$$

Therefore, the smallest possible radius occurs when n = 1, thus,

$$r_0 = 0.53 \times 10^{-8} \text{ cm}$$

This agrees with experimental results for the hydrogen atom.

8.3. APPLICATION TO ATOMS

Now let us look at the total energy of a hydrogen atom. It is the sum of the kinetic energy and the electrostatic potential energy (which is negative). Thus

$$Q_t = (1/2\, m_e v^2) - (q_e^2/r) \tag{8}$$

We know from equation (6) that

$$m_e v^2 = q_e^2/r$$

Therefore,

$$Q_t = 1/2\left(\frac{q_e^2}{r}\right) - \frac{q_e^2}{r}$$

$$Q_t = -1/2\left(\frac{q_e^2}{r}\right) \tag{9}$$

Note that this expression indicates that when the radius of a hydrogen atom is increased, the energy of the atom is increased. But we have already shown that the radius can only exist at discrete positions. Therefore, the energy of the atom must be discrete as well. Using equation (7) and substituting it into equation (9) we have

$$Q_t = -\frac{1}{2}\left(\frac{q_e^2}{0.53 \times 10^{-8}}\right) \cdot \frac{1}{n^2}$$

$$Q_t = -\frac{11.5 \times 10^{-20}}{0.53 \times 10^{-8}} \cdot \frac{1}{n^2}$$

$$Q_t = -21.7 \times 10^{-12} \cdot \left(\frac{1}{n^2}\right) \text{ Joules} \tag{10}$$

This is an interesting expression. From it we can tell what energy we must add to the atom to cause the electron to jump from radius r_1 to radius r_2 (e.g., n = 1 and n = 2 respectively). This energy we must add, ΔQ, becomes

$$\Delta Q = Q_2 - Q_1 = -21.7 \times 10^{-12}\left[\frac{1}{(2)^2} - \frac{1}{(1)^2}\right]$$

$$\Delta Q = 21.7 \times 10^{-12}\left[1 - \frac{1}{4}\right]$$

$$\Delta Q = \left(\frac{3}{4}\right)21.7 \times 10^{-12} = 16.3 \times 10^{-12} \text{ Joules}$$

Problem for the student: Calculate how much energy, ΔQ, must be absorbed by the hydrogen atom to drive the orbiting electron to an infinite orbit. Under this condition the atom is said to be ionized.

Recall that the energy of a photon is given by

$$Q_{ph} = h\nu = \frac{hc}{\lambda} \tag{11}$$

where h is Planck's constant and ν is the frequency, c is the velocity of light, and λ is the wavelength. Niels Bohr further postulated that an atom can go from energy level Q_1 to an energy level Q_2 by absorbing a photon or it can go the opposite way — from energy level Q_2 to energy level Q_1 by emitting a photon. See figure 8-1.

8.4. ADDITIONAL QUANTUM NUMBERS

Calculations of the photon wavelengths generated under these conditions agree remarkably well with what is observed experimentally. There was one problem, however; when experimental results were studied carefully they showed that the photon wavelengths predicted by the theory often turned out to be a group of wavelengths close together and clustered about the value predicted by equation (10). The theory which was developed to explain this difference is beyond the scope of this course, but I shall provide you with a modified picture of the atom as we described it above to account for this change.

We have shown how an atom's energy, Q, is quantized, (with the quantum number n) but in addition to this, there are other things about an atom which are quantized. They are

(1) s, Electron spin, s = +1/2 or - 1/2

(2) ℓ, Orbital quantum number, ℓ = 0, 1, 2. . . (n-1)

(3) M, Magnetic quantum number, M = 0, ±1, ±2, . . . ±ℓ*

The electron can spin in one of two directions only (see figure 8-2). The orbital quantum number, ℓ, means there are various elliptically shaped orbits allowed for each quantum number n (see figure 8-3).

The magnetic quantum number, M, is illustrated in figure 8-4.

8.5. QUANTUM NUMBERS FOR MORE THAN ONE ELECTRON

Wolfgang Pauli stated his famous principle which is NO TWO ELECTRONS CAN HAVE THE SAME VALUES FOR THE FOUR QUANTUM NUMBERS n, ℓ, s and M when together in the same atomic structure.

Thus, for the energy level n = 1, all of the combinations of four quantum numbers are

n = 1	ℓ = 0, M = 0, s = +1/2
	ℓ = 0, M = 0, s = -1/2

for n = 2 we have

n = 2	ℓ = 0, M = 0, s = +1/2
	ℓ = 0, M = 0, s = -1/2
	ℓ = 1, M = 0, s = +1/2
	ℓ = 1, M = 0, s = -1/2
	ℓ = 1, M = +1, s = +1/2
	ℓ = 1, M = +1, s = -1/2
	ℓ = 1, M = -1, s = +1/2
	ℓ = 1, M = -1, s = -1/2

* The symbol, M, in this context should not be confused with the symbol for exitance or radiant emittance.

This means for n = 1 there are two allowed electronic states and, therefore, there can be two electrons at n = 1. Similarly for n = 2, there are 8 allowed electron states. This kind of analysis can be continued on up to 238 electrons which exist in the Uranium 238 atom.

8.6. APPLICATIONS TO SOLID STATE DETECTORS

We now have a reasonable understanding, for our purposes, of atomic theory. Our attention should now be devoted to the basic principles of solid state theory. Ultimately we hope to understand how a detector operates, and this involves understanding solid state physics and the interaction of solids with radiation.

All modern infrared photo detectors are made of crystalline solids. This means that the atoms which compose the solid are arranged in regularly spaced intervals as shown in figure 8-5. What happens to the energy structure of the individual atoms when they are brought closer together to form the solid? Remember from my preceding lectures that the energy structure of an atom is a discrete series of allowed energy levels as shown in figure 8-6. As the individual atoms are brought closer together the electrons from one atom are attracted not only by their own nucleus but also by the nucleus of other atoms in the solid. For illustration purposes let us form a one dimensional solid — in other words, a closely packed chain of atoms all alike. If we examine what happens to the energy levels as the atoms in the chain are brought closer together, we would find that the discrete energy level as shown in figure 8-6 would broaden into bands as the value of 'a' gets smaller (as shown in figure 8-7). The characteristics of a solid depend upon the way in which these bands broaden, and the number of electrons which are available to fill them. As a simple example, let us look at a linear chain of Lithium atoms. By looking at the Periodic Table we can see that Lithium atoms each have three positive charges in the nucleus (three protons) and three electrons revolving about the nucleus. This means that if we have N number of atoms in a chain, there will be 3N total electrons in the chain. Remember that previously in the lecture we presented the Pauli Exclusion Principle which states that no two electrons in an atomic system can have identical sets of the four quantum numbers n, s, ℓ, M. This exclusion principle leads us to the conclusion two electrons can exist at n = 1 and 8 electrons can exist at n = 2. Thus, since we have N atoms in the chain, there can only be 2N electrons at the n = 1 level and 8N electrons at the n = 2 level. This means that the Lithium chain of atoms, which has 3N total electrons, will have 2N electrons in the n = 1 level and the n = 2 level which has capability of containing 8N electrons will contain only 1N electrons. Figure 8-8 illustrates the situation for Lithium.

Our linear chain solid will have its atoms separated by an amount a_0 as indicated in figure 8-8. If we look at the energy level diagram along the chain — which we will call the x-direction, we will find a situation as shown in figure 8-9. If we apply electrical contacts on each end of our chain and place a battery in the circuit as shown in figure 8-10, we will

find that there are 1N electrons free to move along the chain. Therefore, Lithium is a good conductor of electricity. It is possible to find materials which have a property such that the valence band is filled and the conduction band is completely empty as shown in figure 8-11. The highest filled state is by definition the valence band while the next unfilled state is the conduction band. Since the conduction band is empty, there are no electrons free to move along the chain, the material is called an insulator. However, if we could cause some of the valence band electrons to jump up in energy into the conduction band this would change the electrical conductivity — the material would be a slightly better electrical conductor. We could cause an electron to jump across the energy gap, ΔQ, by shining a photon on the chain of atoms. The photon must have at least energy ΔQ to give the electron enough energy to jump the gap. The increase in conduction with increase in photon flux is called the photoconductive effect. Any photon of higher energy than ΔQ will cause this effect but photons of energy less than ΔQ will not. We can observe this change in conductivity of the material when the proper photons illuminate the chain, by monitoring the electrical voltage across the resistor which is in the circuit.

Thus, we have the makings of a solid state quantum detector. If we choose the proper materials such as InSb to make our crystalline solid, infrared photons can cause electrons to jump into the conduction band. The response of the detector for equal photon flux at each wavelength would be as shown in figure 8-12. Since the energy of a photon is

$$Q_{photon} = \frac{hc}{\lambda}$$

the energy is smaller for longer wavelength photons. The value λ_c in figure 8-13 is the longest wavelength to which the detector is sensitive. The response of an infrared detector to equal radiation power at each wavelength is shown in figure 8-13. If we plot the detector response to equal radiation power in a logarithmic scale the curve would appear as in figure 8-14. This is the general shape of infrared detector spectral responses as reported in most text books and manufacturer's advertisements.

The same solid state theory we have used to explain the operation of detectors is useful in understanding the operation of some kinds of spectral filters. Figure 8-11 shows the energy band picture of an insulator. If we take a crystalline solid with long wavelength limit λ_{c_1} we should find that photons with wavelengths longer than λ_{c_1} will not be absorbed by the material but will instead be transmitted. Therefore, if we select our materials correctly for the detector and the filter (see figure 8-15) the filter should absorb radiation of wavelength shorter than λ_{c_1}. The detector itself sets the limit to the longest wavelength photon, namely λ_{c_2}.

8.7. DETECTOR RESPONSIVITY

The responsivity R of a detector used to detect noncoherent optical and infrared radiation is defined as the change in signal output, ΔS, of the detector divided by the change in incident power, ΔP, on the detector.

$$R = \frac{\Delta S}{\Delta P} \text{ signal units per watt}$$

The signal output is usually proportional to the incident power in some narrow spectral band. Radio antennas and acoustic or seismic detectors usually produce output signals proportional to the input wave displacement so this definition of responsivity is not used for them. Where the signal output is in the form of a voltage change or current change the units of R are volts/watt or amperes/watt respectively. For detectors that are primarily used for visible radiation, the responsivity, R_v, is normally given in amperes/lumen. Detectors do not produce the full required output signal immediately upon changing the incident power level. If the incident power level is abruptly raised, the output signal will immediately begin to rise but will require a finite time to reach its final higher value. When the power is sharply reduced, the detector signal begins to decrease immediately but again requires a finite time to reach the final lower value. When abrupt incident power changes are made the time required for a detector to reach an output value within about 36% of its final value is called the detector response time. It is evident that fluctuating incident power level cannot be accurately measured by the fluctuating output signal if the fluctuations in power are too rapid. The responsivity can be measured using a harmonically varying power level where the frequency of variation is f. It will be found that the harmonically varying signal output will be much reduced for frequencies above some critical frequency, f_c, where

$$f_c = \frac{1}{2\pi \times (\text{response time})}$$

The responsivity is dependent on the modulation frequency, f, of the fluctuating incident power.

A detector is to be employed in a radiometer so that as the radiometer intercepts in rapid succession radiant power from various objects in a scene, the output signal should be a measure of that power with good fidelity. Clearly, the speed of making such a series of measurements with fidelity will be limited by the response time of the detector.

The responsivity is also dependent upon the wavelength of the radiant power as was described earlier. When the responsivity to monochromatic power is expressed, the term spectral responsivity is used and symbolized by $R(\lambda)$ to specify that a particular narrow spectral range centered at wavelength λ is used to produce the signal. It is assumed (and usually true enough) that

$$\Delta S_{total} = \Delta S_1 + \Delta S_2 = R(\lambda_1)\Delta P(\lambda_1) + R(\lambda_2)\Delta P(\lambda_2)$$

or that the total signal due to power contributions in different spectral bands is equal to the sum of the signals which would have been generated by these contributions separately. A knowledge of the spectral dependence of $R(\lambda)$ and the incident spectral power is all that is needed to predict the signal output for fluctuation rates below f_c. Thus,

$$\Delta S_{total} = \int_0^\infty R(\lambda) \frac{\text{signal units}}{\text{watt}} \cdot P_\lambda \frac{\text{watt}}{\mu\text{m}} \cdot d\lambda(\mu\text{m})$$

8.8. DETECTOR NOISE

All detectors, without exception, produce some small fluctuation in the output even when the time average of the input power is held constant. That output fluctuation is called noise. There are many causes of noise. Some causes are inherent in the structure of the detector. Noise from such sources is called internal detector noise. Another important source of noise is caused by the fact that it is impossible in principle to create a flow of radiant power which is without fluctuations in time. The noise due to a steady average flow of power is called photon noise and arises from the probabilistic delivery of energy along the wave front of the elemental component waves of radiation. Photon noise increases when the average power level increases. Internal detector noise may not increase with power level on the detector. A detector which operates under conditions such that the main contribution to the noise output is photon noise is said to be operating in the photon noise limited condition. Noise is a random fluctuation of the output about the average output signal level. It can be considered as the result of the superposition of an infinite number of very low amplitude simple harmonic output variations with all frequencies represented but with relative phases chosen at random. If the detector and/or amplifier following the detector does not respond to all frequencies, then only those noise components will be observed for which response is possible. Thus, a detector-amplifier combination responding only to a very narrow band of fluctuation frequencies will provide an output noise fluctuation much less than a detector-amplifier combination responding to a very large band of fluctuation frequencies because fewer low amplitude fluctuations add together to make the noise output. However, unless the signal power fluctuations which are to be detected are also composed of the same narrow frequency band of simple harmonic fluctuation components, the signal will not appear with full fidelity in the output.

Noise fluctuations appear on a TV set tuned to a vacant channel. It looks like a swirling snow storm. The audio system responds to noise with a sound which is like rain in a forest of trees.

The fluctuation of noise output about the average signal level is as much positive as it is negative on the average so that the average fluctuation is zero. The variance of the fluctuation (the mean square) is not zero. A measure of the degree of noise fluctuation is taken as the square root of the variance or the root-mean-square abbreviated as rms. The rms noise fluctuation is proportional to the square root of the fluctuation frequency band pass allowed.

8.9. NOISE EQUIVALENT POWER (NEP)

If a detector has a constant responsivity with power level then the signal output of the detector can be predicted from the incident power level as

$$S = R\,P.$$

If the power is reduced to the point where the signal amplitude is equal to the rms noise output, then that power is called the noise equivalent power. It is that input power which produces a signal amplitude comparable to the noise output fluctuation. The signal to noise ratio is the signal amplitude divided by the rms noise fluctuation. The noise equivalent power is that incident power which will provide a signal to noise ratio of unity. Since $S = RP$ then $S_n = R\,(NEP)$ where S_n is the rms noise fluctuation value and (NEP) is the noise equivalent power. The NEP can be calculated from measured values of S_n and R. The NEP is a good measure of the smallest change in signal power which can be distinguished reliably by the detector when the noise output is primarily due to internal detector noise, the NEP can be considered as one figure of merit for the detector. A low value of NEP indicates a "good" detector capable of making reliable measurements on very small incident power levels. In order to express the figure of merit in such a way that a larger number implies a good detector, the term detectivity was invented where the detectivity, D, is defined as

$$D = \frac{1}{NEP}$$

The NEP or the D value for a detector is in part determined by the geometry of the detector and the band of fluctuation frequencies, Δf, which are utilized in the output signal of the detector. Consider for example two identical detectors each having receiving area, A. By placing them together to permit the same irradiance E to be incident upon them and by connecting them together so that the output signals S of each add together, one will find that the total signal for the pair will be 2 S. The noise fluctuations of these two detectors are not in phase with each other since the sources of noise are independent. Thus, the rms noise fluctuation of the two detectors together is only $\sqrt{2}\,S_n$. (The noise fluctuation in one may be at times subtracting and at times adding to the noise fluctuations of the other). The signal to noise ratio of each separately is S/S_n but together it is

$$\frac{2\,S}{\sqrt{2}\,S_n} = \sqrt{2}\,\frac{S}{S_n}.$$

It is a simple extension in reasoning to show that three identical detectors so connected would produce a signal to noise ratio of $\sqrt{3}\,S/S_n$ and for N detectors. $\sqrt{N}\,S/S_n$. Now a single large area detector may be considered to be many small area detectors connected together inherently. Therefore, the signal to noise ratio of a detector when exposed to uniform irradiance will be proportional to the square root of the area of the detection because the number of identical small area detectors that would be required to make up any detector would be proportional to the area of the detector.

As previously stated in discussing detector noise, the rms noise fluctuation, S_n, is proportional to the square root of the fluctuation frequency band pass, i.e., $\sqrt{\Delta f}$, in the frequency range of detector operation. Hence, the noise equivalent power will also depend upon the frequency band pass one chooses to use.

A figure of merit, which is independent of the choice of area and the band pass that one may choose, can be made to represent some inherent quality of a detector type. That figure of merit is called D-star and is written D^*. The quantity D^* is a normalized form of detectivity, D, to remove the area and band pass dependence (see figure 8-16). D^* is defined as

$$D^* = \sqrt{A\,\Delta f}\,D$$

or using NEP,

$$D^*\ \text{watt}^{-1}\,\text{m}\,\text{Hz}^{1/2} = \frac{\sqrt{A\,\Delta f}}{NEP}$$

The spectral detectivity $D(\lambda)$, spectral D-star $D^*(\lambda)$, and spectral $NEP(\lambda)$ are defined using the spectral responsivity $R(\lambda)$.

A knowledge of the spectral responsivity and detectivity will permit the calculation of the responsivity for any arbitrary spectral distribution of power. Thus,

$$\frac{S}{P} = \frac{\int_0^\infty R(\lambda)P_\lambda\,d\lambda}{\int_0^\infty P_\lambda\,d\lambda} = R.$$

The signal to noise ratio that can be expected from a detector limited by internal detector noise is given simply by

$$\frac{S}{S_n} = \int_0^\infty D(\lambda)P_\lambda\,d\lambda$$

In the Handbook of Military Infrared Technology, the spectral $D^*(\lambda)$ is given for many kinds of infrared detectors. The nomenclature used in the handbook is not consistent with that used here. On page 486 of the handbook the $D^*(\lambda)$ is plotted for the InSb cooled photoconductive detector. The maximum of the spectral responsivity is designated as $R_{\lambda\,max}$. Notice that this subscript does not imply the spectral density (per μm wavelength). Use of the wavelength symbol should be reserved for spectral density. To be consistent with the nomenclature presented here one could write $R(\lambda\ max)$. One can see from the plot of $D^*(\lambda)$ that $\lambda\ max = 5.5\ \mu m$. The symbol τ is used to express the response time and should not be confused with the symbol for transmittance. Context will distinguish the two uses. The value of D^* depends upon the fluctuation frequency, f. Thus, $D^*(\lambda)$ is measured at one frequency and is denoted as $D^*(\lambda, f)$. On page 486 the detector was measured with monochromatic power interrupted at 1,000 Hz (or cycles per second).

Detectors for which the photon noise predominates do not have a detectivity which is characteristic of the detector since the source of noise is external to the detector. Although a detectivity can be defined under any particular conditions of irradiance, the limiting noise will depend upon the spectral distribution of the irradiance as well as the magnitude. Photo-emissive detectors that are commonly used in the visible light spectral range are almost always photon noise limited in daylight remote sensing applications. For photo-emissive detectors the signal to noise ratio for small signals is given approximately by

$$\frac{S}{S_n} = \left(\frac{\Delta i}{e}\right) \Big/ \sqrt{\frac{i\,\Delta f}{e}}$$

Where Δi is the change in current due to a signal fluctuation, i is the average current, e is the charge of an electron and Δf is the fluctuation frequency band pass of the following amplifier. Figure 8-17 shows the typical circuit connection for photoconducting and photovoltaic detectors.

APPENDIX I

DEFINITIONS OF, AND DISTINCTIONS BETWEEN, BIDIRECTIONAL AND DIRECTIONAL REFLECTANCES

In discussing the radiation reflected from a surface, it is often necessary to introduce two parameters called bidirectional and directional reflectance. It is unfortunate that both these parameters bear the name of reflectance since they describe completely different physical quantities. Only one of the two conforms to what is usually meant by reflectance, i.e., the ratio of reflected to incident powers. This short note attempts to clarify the distinction between the two, and indicates their relationship. A sample calculation is given.

The terms and definitions presented here are essentially those proposed by Nicodemus.* Bidirectional reflectance and directional reflectance are denoted by ρ' and ρ_d, respectively. To define these two quantities, consider an infinitesimal surface element dA exposed to incident radiation. The incident irradiance in the element of solid angle $d\Omega_i$, which is at angles θ_i and ϕ_i to the normal of dA (see figure 1), is dE_i. In terms of the incident radiance L_i in this solid angle:

$$dE_i(\theta_i, \phi_i) = L_i(\theta_i, \phi_i) \cos \theta_i \, d\Omega_i, \tag{1}$$

where $d\Omega_i = \sin \theta_i d\theta_i \, d\phi_i$. Consider the radiation reflected into some element of solid angle $d\Omega_r$ making angles θ_r and ϕ_r with the normal to dA (see figure I-1). In general, the irradiance dE_i in each of the different solid angles $d\Omega_i$ makes some contribution to the reflected radiance L_r in a given solid angle $d\Omega_r$. That portion of the reflected radiance in a given solid angle $d\Omega_r$ at angles θ_r and ϕ_r which results from the reflection of the irradiance in a particular solid angle $d\Omega_i$ at angles θ_i and ϕ_i is denoted by $dL_r(\theta_r, \phi_r)$. The bidirectional reflectance, ρ', is then defined by:

$$dL_r(\theta_r, \phi_r) = \rho'(\theta_i, \phi_i; \theta_r, \phi_r) \, dE_i(\theta_i, \phi_i) \tag{2}$$

or

$$\rho'(\theta_i, \phi_i; \theta_r, \phi_r) = \frac{dL_r(\theta_r, \phi_r)}{L_i(\theta_i, \phi_i) \cos \theta_i \, d\Omega_i}$$

---

* Nicodemus, F. E., Directional Reflectance and Emissivity of an Opaque Surface, Technical Memorandum No. EDL-G266, Electronic Defense Laboratories, Mountain View, Calif., 19 May 1966.

The dimensions of ρ' are:

$$\frac{[\text{watts/(cm}^2\text{ steradian)}]\text{ reflected}}{[\text{watts/cm}^2]\text{ incident}}$$

From equation (2), it is seen that ρ' can take on any positive value.

Although ρ' is called bidirectional reflectance, it does not represent the ratio of reflected to incident power (or even of reflected to incident irradiance). More correctly, it should be designated by some term including the words "distribution function," since it is a measure of how the incident irradiance in a solid angle $d\Omega_i$ is distributed (after reflection) among the various solid angles $d\Omega_r$.

Let us proceed to the definition of the directional reflectance, which will be denoted by ρ_d. If radiation, well collimated within an infinitesimal solid angle $d\Omega_i$ at angles θ_i and ϕ_i to the normal dA, is incident on dA, then the total power incident on dA is

$$dP_i = dE_i(\theta_i, \phi_i)\, dA = L_i(\theta_i, \phi_i) \cos \theta_i\, d\Omega_i\, dA \tag{3}$$

In general, this incident radiation will be reflected into all directions, the total reflected power from dA being called dP_r. The directional reflectance for collimated incident radiation (ρ_{di}) is defined by

$$dP_r = \rho_{di}\, dP_i. \tag{4}$$

The reflected power from dA can be found from the radiance by:

$$dP_r = \left[\int_h L_r(\theta_r, \phi_r) \cos \theta_r\, d\Omega_r\right] dA, \tag{5}$$

where the integration is over a hemisphere. If all the incident radiation is confined to the solid angle $d\Omega_i$ at the angles θ_i and ϕ_i, then from (2)

$$L_r(\theta_r, \phi_r) = \int dL_r \int \rho'(\theta_i, \phi_i; \theta_r, \phi_r)\, L_i(\theta_i, \phi_i) \cos \theta_i\, d\Omega_i,$$

since $L_i(\theta_i, \phi_i)$ is zero for all θ_i and ϕ_i except those equal to the incident direction, we have

$$L_r(\theta_r, \phi_r) = \rho'(\theta_i, \phi_i; \theta_r, \phi_r)\, L_i(\theta_i, \phi_i) \cos \theta_i\, d\Omega_i.$$

$L_i(\theta_i, \phi_i)$ is the value of radiance in the incident direction.

Substitution of this, (5) and (3) into (4) yields

$$\left[\int_h \rho'(\theta_i,\phi_i;\ \theta_r,\ \phi_r)\ (L_i(\theta_i,\ \phi_i)\cos\theta_i\, d\Omega_i)\cos\theta_r\, d\Omega_r\right] dA = \rho_{di} L_i(\theta_i,\ \phi_i)\cos\theta_i\, d\Omega_i\, dA$$

or,

$$\rho_{di}(\theta_i,\ \phi_i) = \int_h \rho'(\theta_i,\ \phi_i;\ \theta_r,\ \phi_r)\cos\theta_r\, d\Omega_r. \tag{6}$$

In analogous fashion, if the incident radiation comes in from all angles and we are interested in the power reflected into a certain solid angle $d\Omega_r$ making angles θ_r and ϕ_r with the normal to dA, we can define a directional reflectance ρ_{dr} for this case, with the result

$$\rho_{dr}(\theta_r,\ \phi_r) = \int_h \rho'(\theta_i,\ \phi_i;\ \theta_r,\ \phi_r)\cos\theta_i\, d\Omega_i. \tag{7}$$

By the Helmholtz reciprocity theorem,

$$\rho_{di}(\theta_i,\ \phi_i) = \rho_{dr}(\theta_r,\ \phi_r) \tag{8}$$

for $\theta_i = \theta_r$ and $\phi_i = \phi_r$. And thus for particular angles θ and ϕ:

$$\rho_{di}(\theta,\ \phi) = \rho_{dr}(\theta,\ \phi) = \rho_d(\theta,\ \phi) \tag{9}$$

Notice that ρ_d is a true reflectance, the ratio of reflected to incident power. As such, it must be less than or equal to unity

$$\rho_d \leqq 1. \tag{10}$$

However, ρ' may have any positive value consistent with the requirement (obtained by substituting (6) into (10)):

$$\int_h \rho'(\theta_i,\ \phi_i;\ \theta_r,\ \phi_r)\cos\theta_r \sin\theta_r\, d\theta_r\, d\phi_r \leqq 1. \tag{11}$$

As an example, consider a perfect specular reflector,

$$\rho_{di} = \rho_{dr} = \rho_d = 1,$$

and

$$\rho' = \frac{1}{2}\,\delta(\sin^2\theta_r - \sin^2\theta_i)\,\delta(\phi_r - \phi_i \pm \pi),$$

where the δ's are Dirac delta functions. These functions are presented in figure I-2. Using (6) and integrating over ϕ_r, we have

$$\int_0^{\pi/2} \rho']_\phi \cos\theta_r \sin\theta_r d\theta_r = \rho_d = 1,$$

where

$$\rho']_\phi = \int_0^{2\pi} \rho' d\phi_r$$

Hence, if we plot $\rho']_\phi \cos\theta_r \sin\theta_r$ vs. θ_r as in figure I-3, the area under the curve must equal ρ_d (or unity).

As another example, assume that all quantities are independent of the azimuth angle ϕ, then

$$\rho_{di}(\theta_i) = \int_h \rho(\theta_i, \theta_r) \cos\theta_r \sin\theta_r d\theta_r d\phi_r = 2\pi \int_0^{\pi/2} \rho'(\theta_i, \theta_r) \cos\theta_r \sin\theta_r d\theta_r.$$

If we plot $\rho' \cos\theta_r \sin\theta_r$ vs. θ_r (see figure I-4), then the area under the curve must equal $\rho_{di}(\theta_i)/2\pi$ for each θ_i. Furthermore, if the reflector is Lambertian ρ' is a constant (and therefore also ρ_{di}). Performing the integration over θ_r:

$$\rho_{di} = \pi\rho' = C \text{ (a constant).}$$

Figures I-5(A) and (B) are graphs of ρ_{di} vs. θ_i, and $\rho' \cos\theta_r \sin\theta_r$ vs. θ_r for this case. Note that the area under the $(\rho' \cos\theta_r \sin\theta_r)$- curve must equal $C/2\pi$.

As an example for use in this memo, laboratory measurements were made on a wooden block painted with a 50% - 50% mixture of black and white 3M paint. A Beckman DK-2 spectrophotometer was used to obtain directional spectral reflectance values (ρ_d) and a goniometer (of our own design) to obtain bidirectional reflectance (ρ'). The wavelength used in both cases was 6328 Å. Freshly smoked MgO was used in the Beckman unit as a standard, and the bidirectional reflectance of MgO was also measured (ρ'_{MgO}).

Two measurements were made with the Beckman unit using radiation normally incident on the block. The sample was rotated 90^o about the normal as axis for the second measurement. The average value of the ratio ρ_d to $\rho_{d,MgO}$ was

$$\frac{\rho_d}{\rho_{d,MgO}} = 10.8 \times 10^{-2}$$

Many measurements of $\rho' \cos\theta_r$, where θ_r is the reflected polar angle, were made with the goniometer for a given ϕ_r and $\phi_r + \pi$ (ϕ_r is the reflected azimuth angle). The incident radiation was always normal to the block, and collimated. The results are presented in Table I. In Table II are presented the average of measurements of $\rho'_{MgO} \cos\theta_r$.

TABLE I

ϕ_r		$\phi_r + \pi$
θ_r	$\langle \rho' \cos \theta_r \rangle_{average}$	$\langle \rho' \cos \theta_r \rangle_{average}$
10°	0.03460	0.03543
20°	0.03163	0.03293
30°	0.02903	0.02973
40°	0.0540	0.02665
50°	0.02173	0.02265
60°	0.01720	0.01833
70°	0.01215	0.01293
80°	0.00632	0.00722

TABLE II

θ_r	$\langle \rho'_{MgO} \cos \theta_r \rangle_{average}$
10°	0.297
20°	0.274
30°	0.248
45°	0.190
60°	0.130
70°	0.0822
80°	0.0397

The tabulated values for the block and MgO were multipled by sin θ_r and plotted against θ_r (as in figure I-4 and I-5) and a smooth curve drawn through the points. The areas were determined with a planimeter. For ϕ_r, the area under the curve (A) divided by the area under the curve for MgO(A_{MgO}) was, approximately,

$$\frac{A}{A_{MgO}} \approx 14.6 \times 10^{-2}$$

Correspondingly, for $\phi_r + \pi$, the ratio of areas is

$$\frac{A}{A_{MgO}} \approx 12.8 \times 10^{-2}.$$

Since, as we have previously shown,

$$\int_0^{\pi/2} \rho' \cos \theta_r \sin \theta_r d\theta_r = A = \rho_d,$$

then

$$\frac{A}{A_{MgO}} = \frac{\rho_d}{\rho_{d,MgO}}$$

The difference between the values obtained from the Beckman and those from integration of the goniometric values could be due to neglect of polarization effects and also errors involved in determining the areas under the curves.

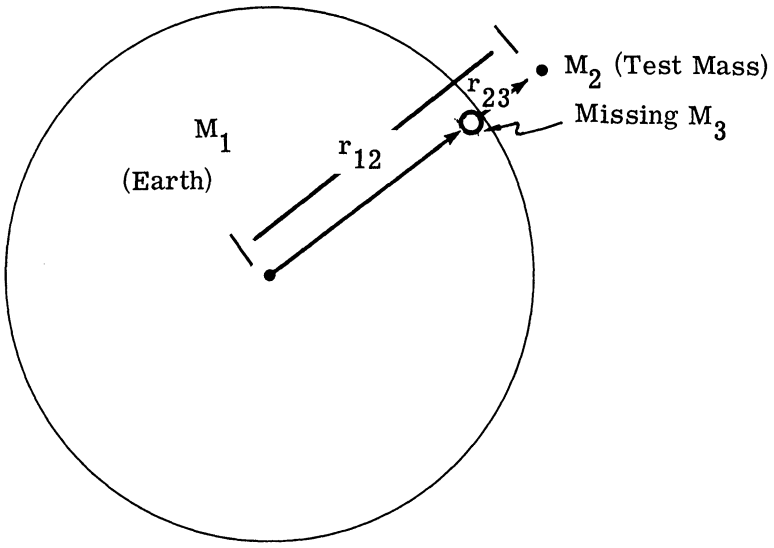


FIGURE 2-1. EFFECT OF A VOID ON THE GRAVITATIONAL FIELD

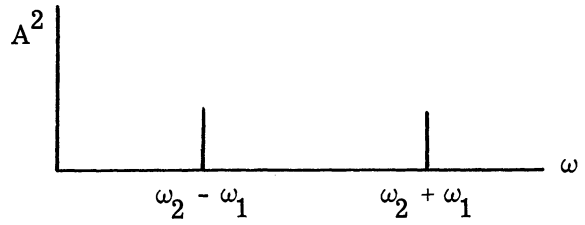


FIGURE 2-2

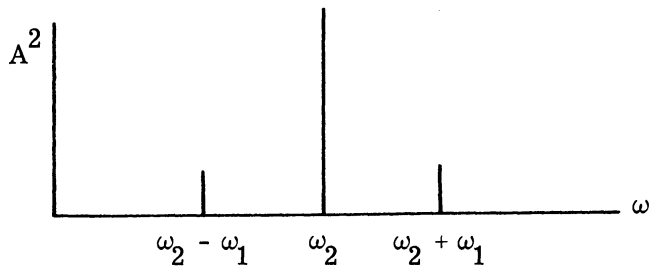


FIGURE 2-3

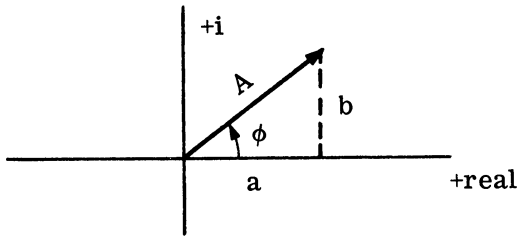


FIGURE 2-4

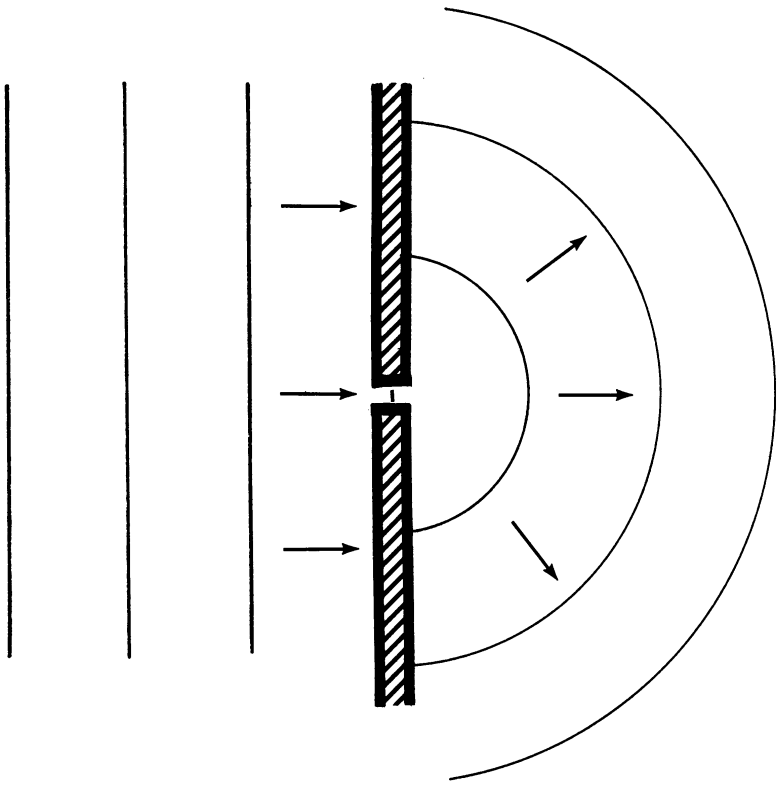


FIGURE 2-5. DIFFRACTION THROUGH A HOLE LEAVES ONLY ONE HUYGEN WAVELET

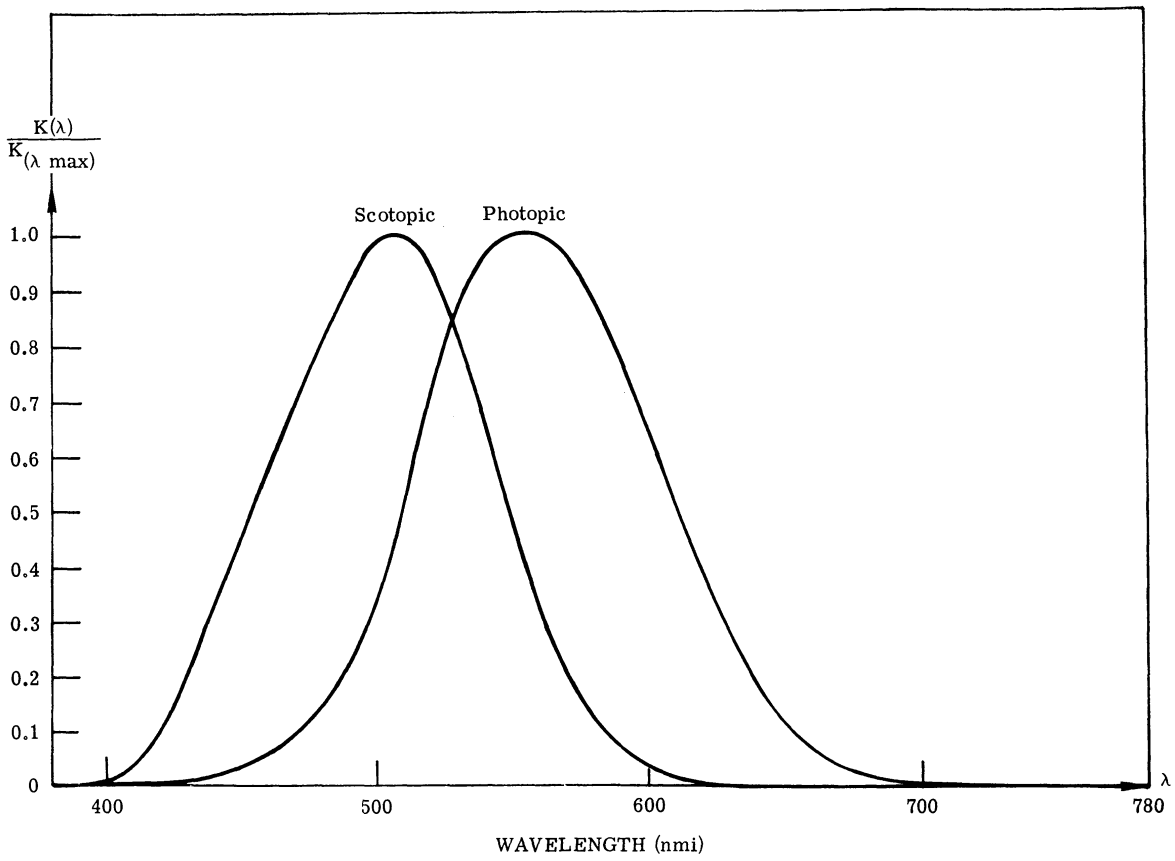


FIGURE 3-1. VISIBILITY CURVE. Normalized luminosity versus wavelength for photopic (daytime) and scotopic (nighttime) vision.

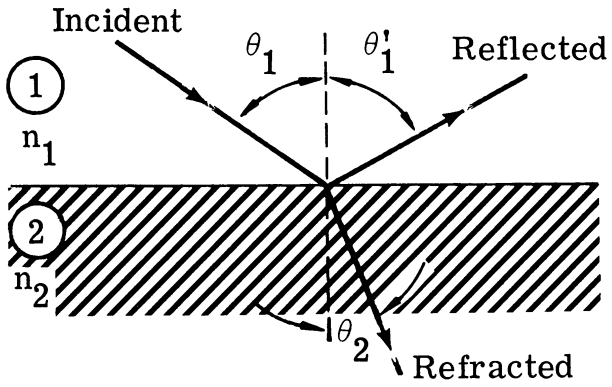


FIGURE 4-1

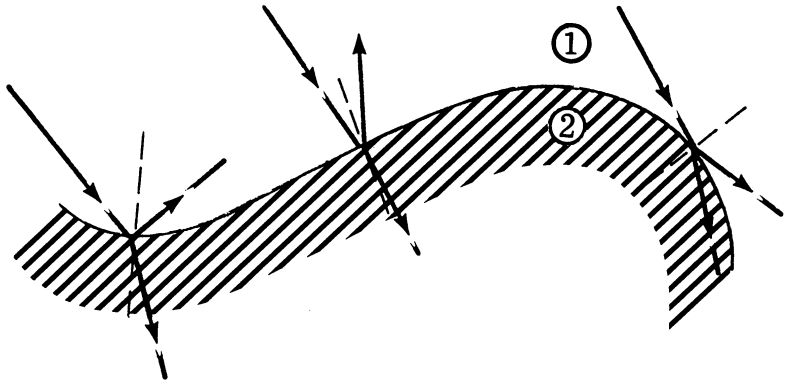


FIGURE 4-2

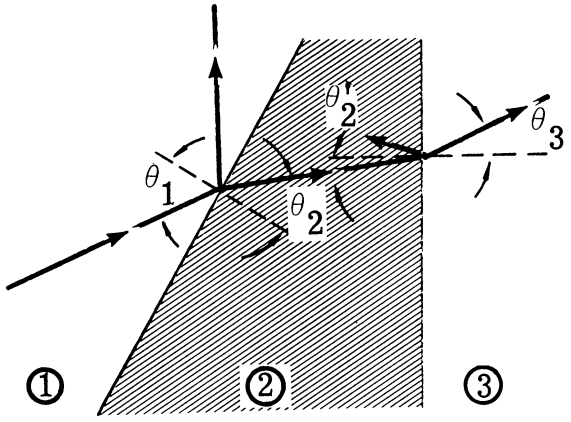


FIGURE 4-3

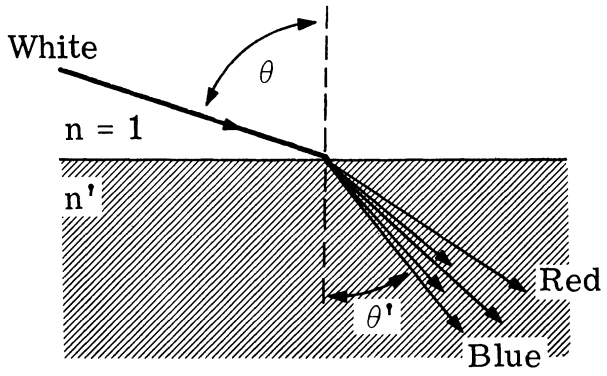


FIGURE 4-4

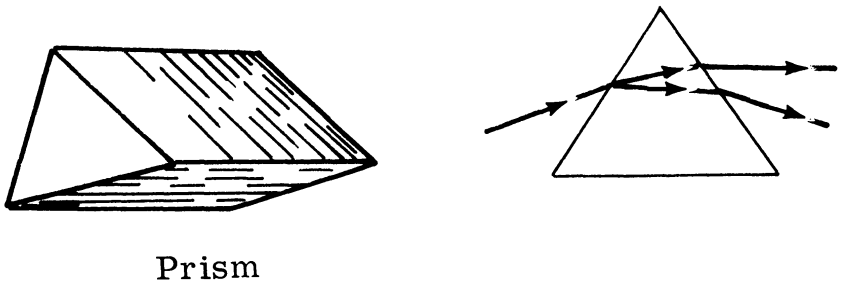


FIGURE 4-5

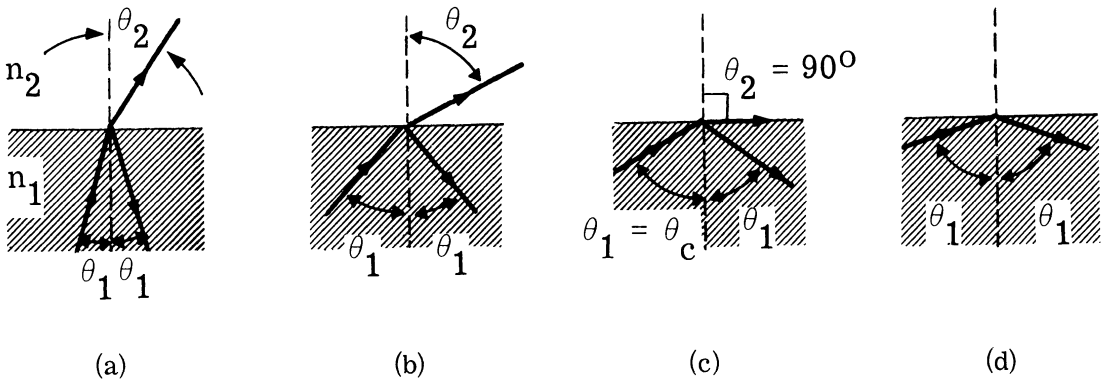


FIGURE 4-6

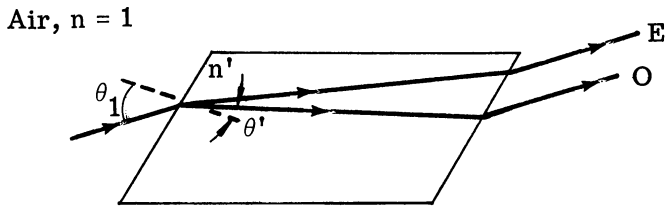


FIGURE 4-7

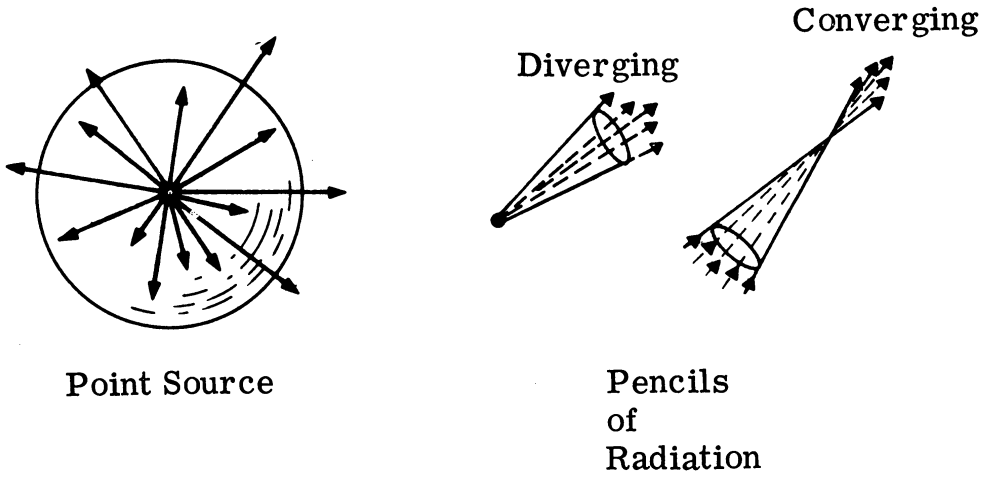


FIGURE 4-8

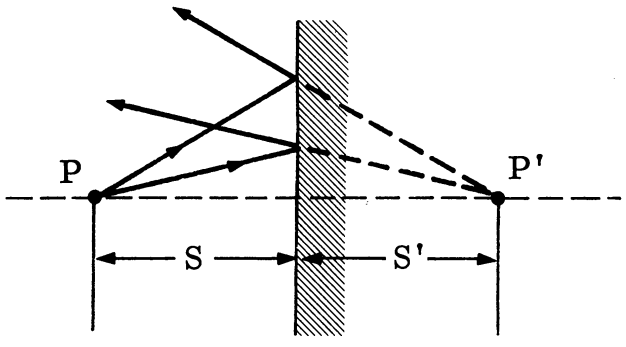
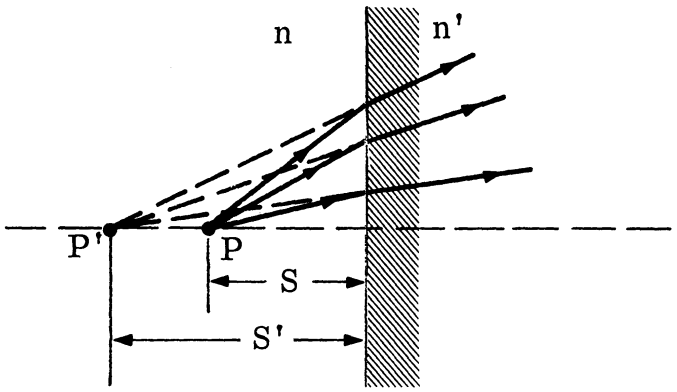


FIGURE 4-9



$n' > n$

FIGURE 4-10

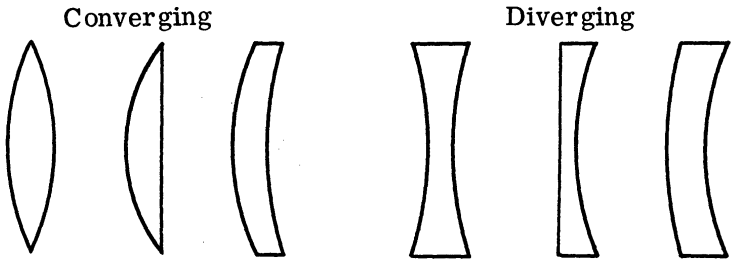


FIGURE 4-11

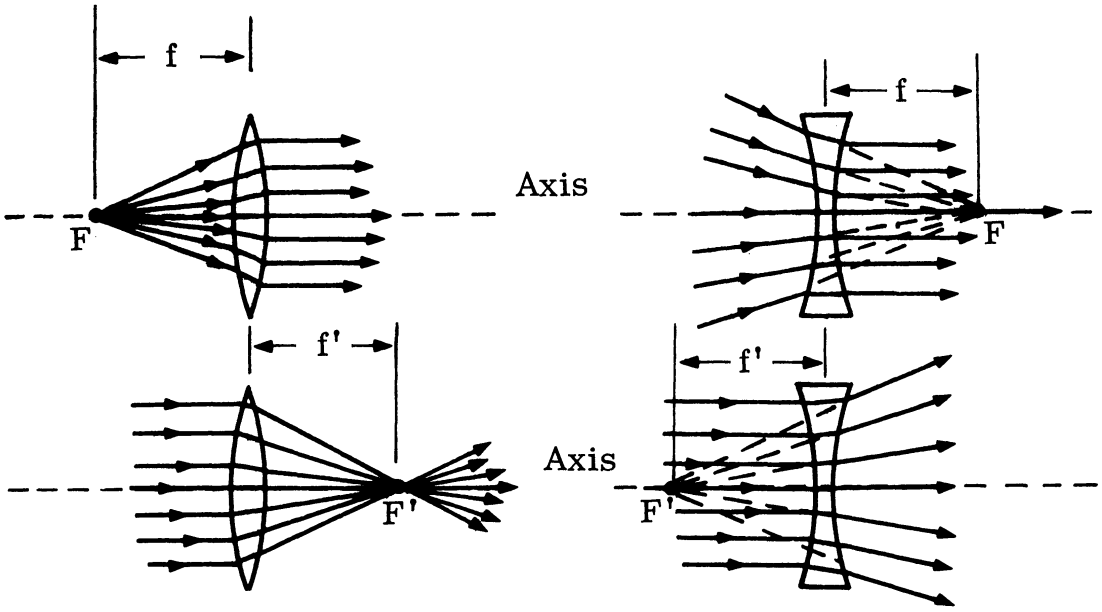


FIGURE 4-12

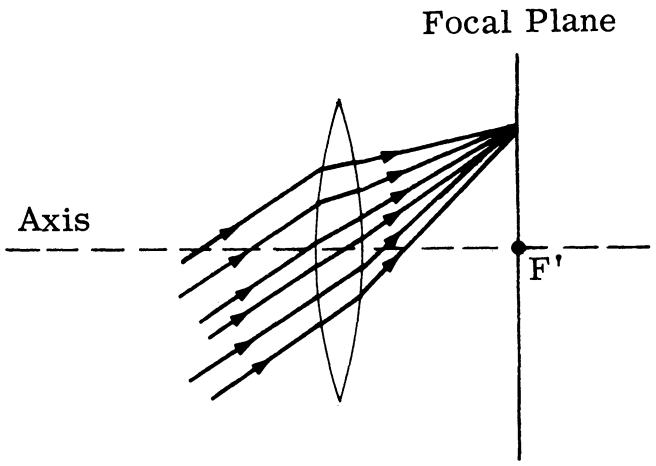


FIGURE 4-13

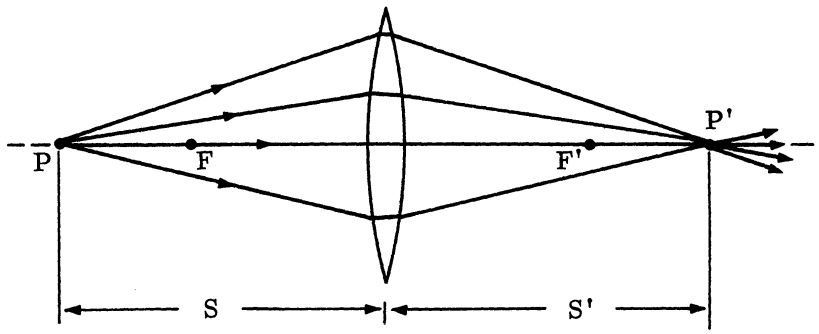


FIGURE 4-14

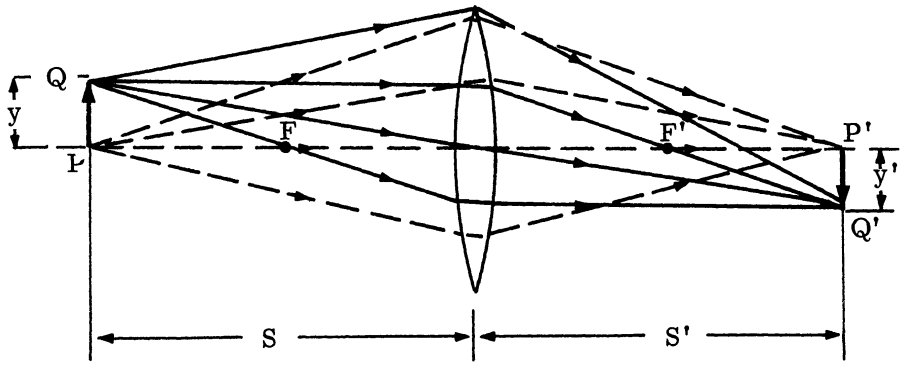


FIGURE 4-15

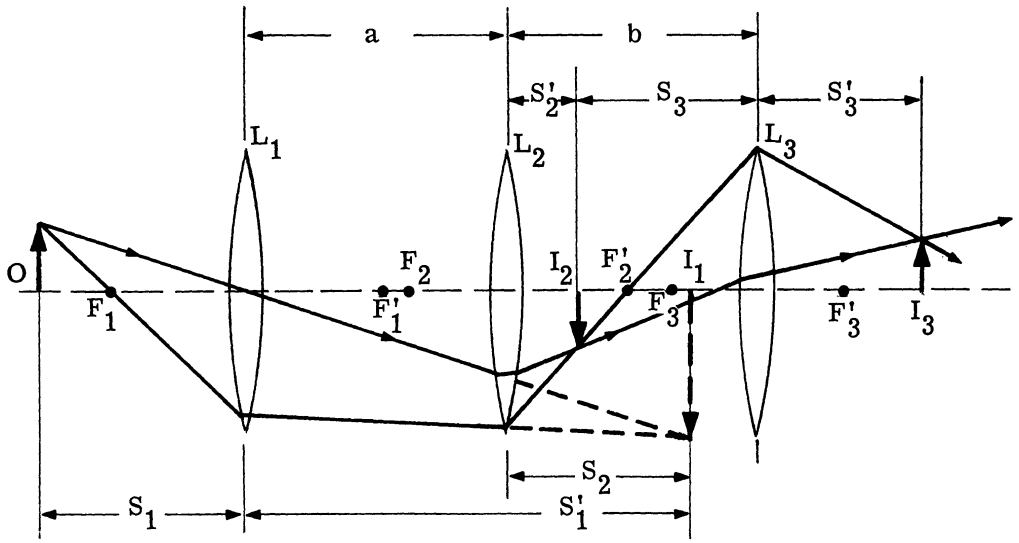


FIGURE 4-16

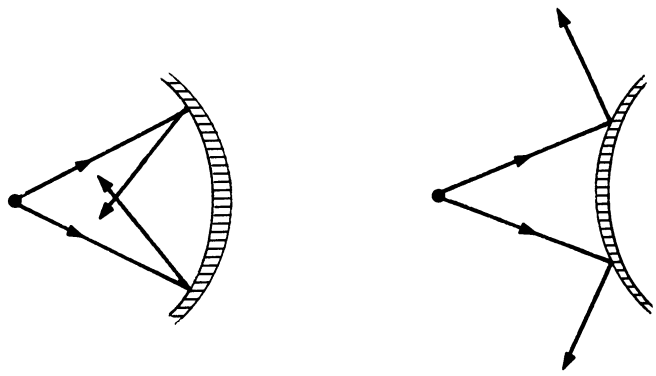


FIGURE 4-17

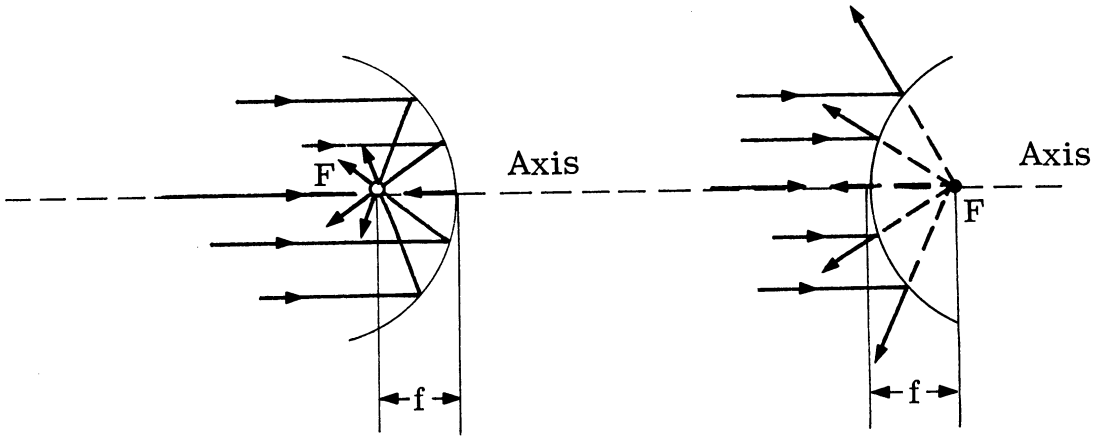


FIGURE 4-18

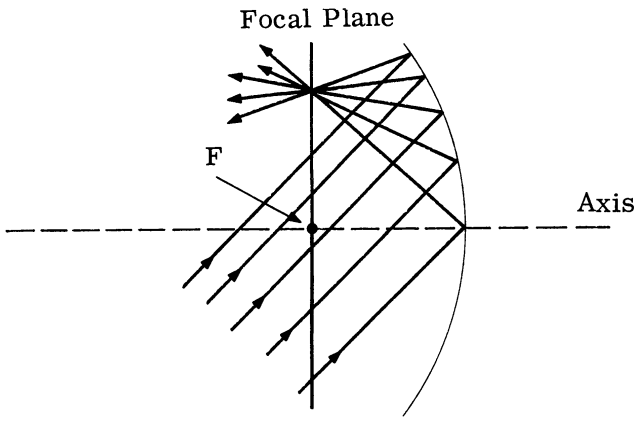


FIGURE 4-19

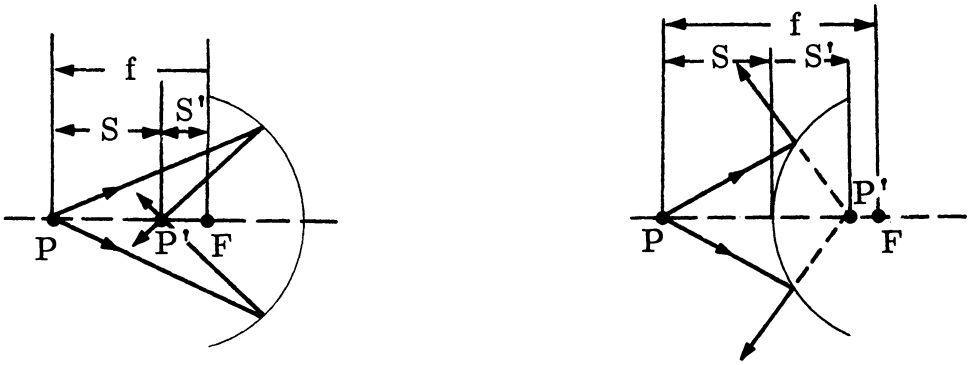


FIGURE 4-20

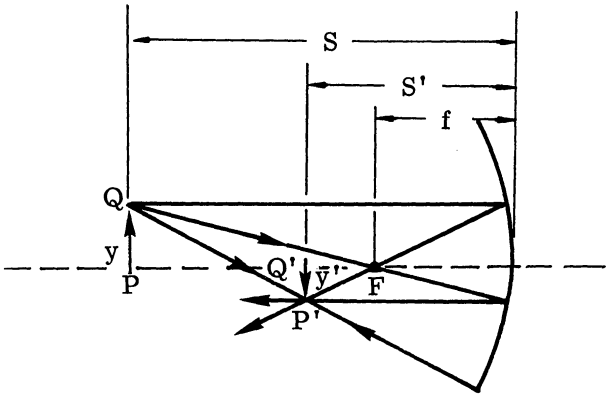


FIGURE 4-21

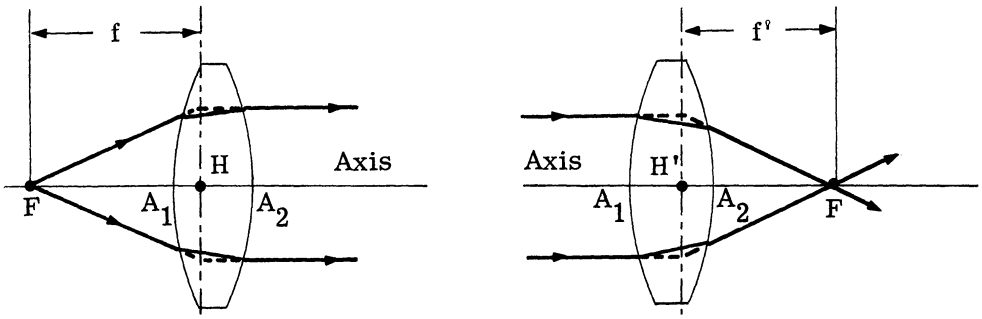


FIGURE 4-22

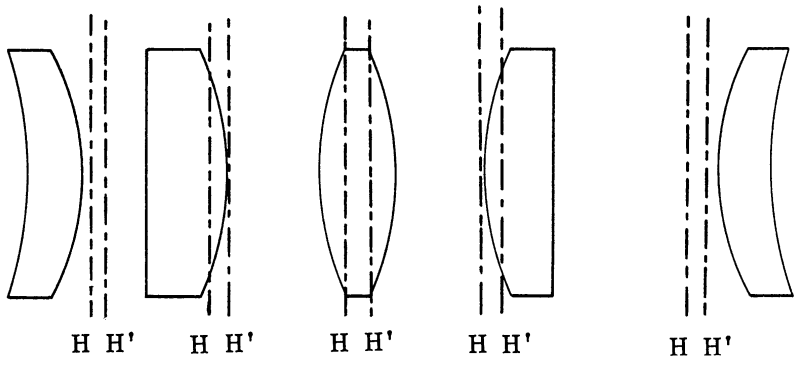


FIGURE 4-23

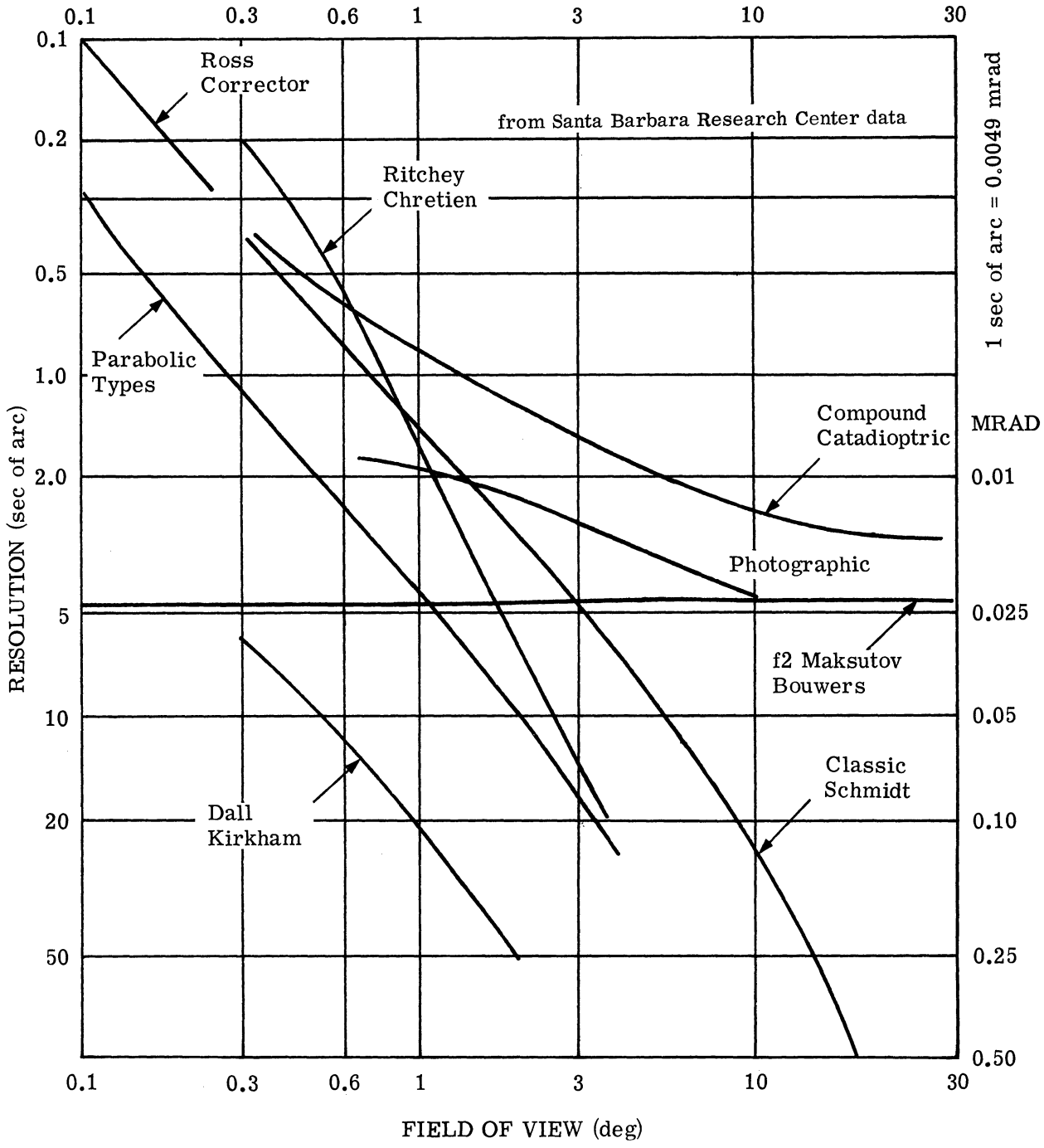


FIGURE 4-24. RESOLUTION OF TYPICAL OPTICAL SYSTEMS

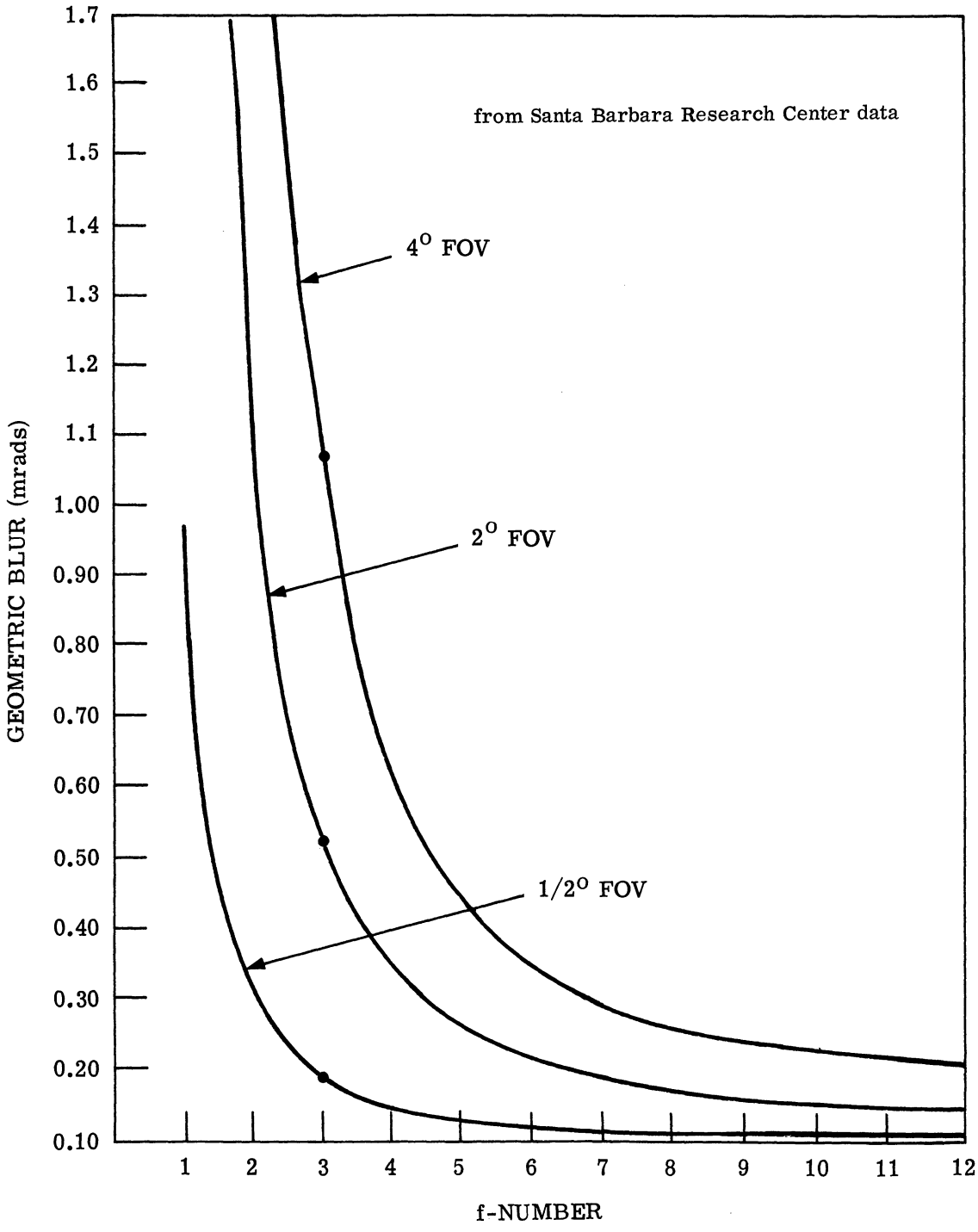


FIGURE 4-25. OFF-AXIS ABERRATIONS OF PARABOLIC REFLECTOR

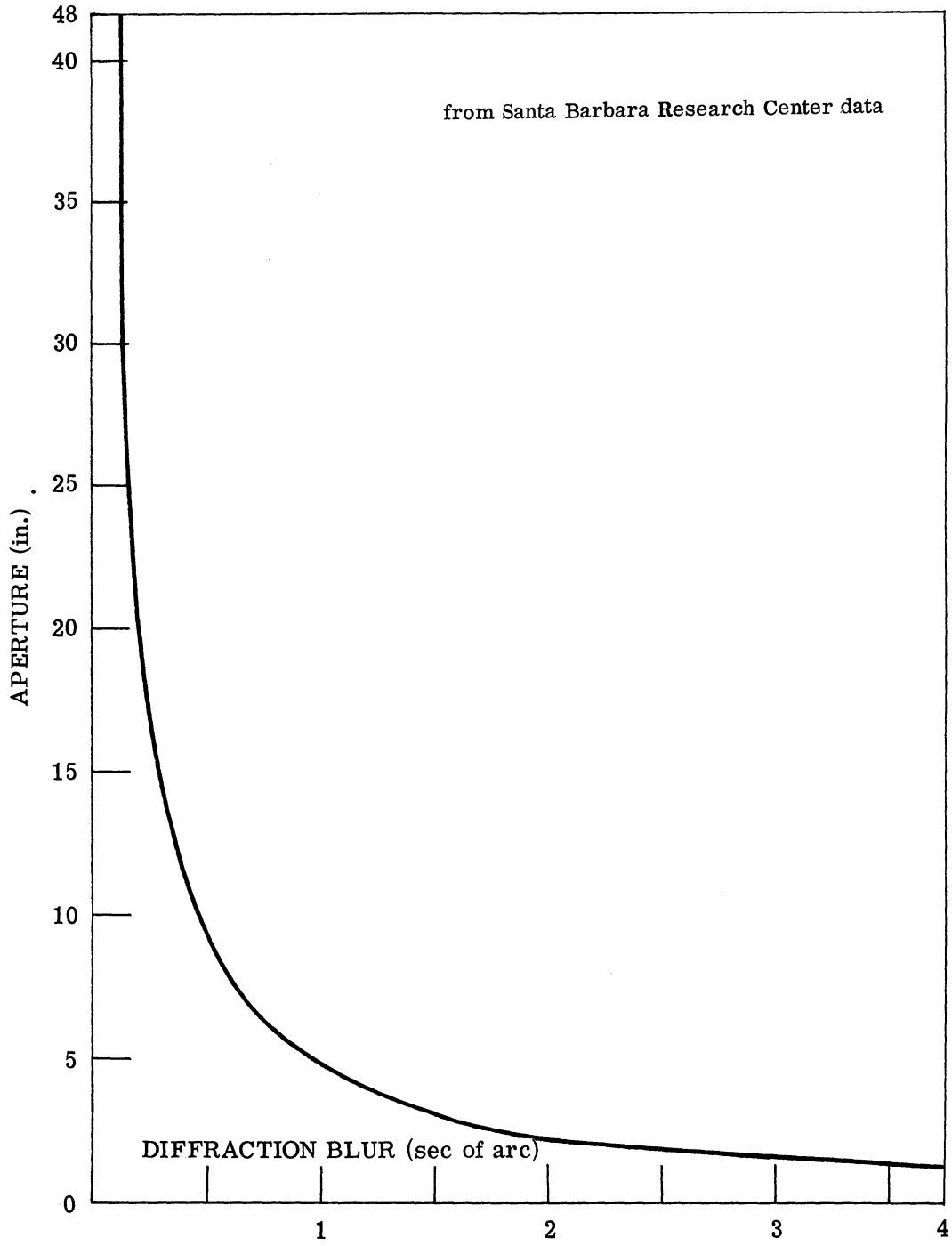


FIGURE 4-26. DIFFRACTION LIMIT FOR TELESCOPES

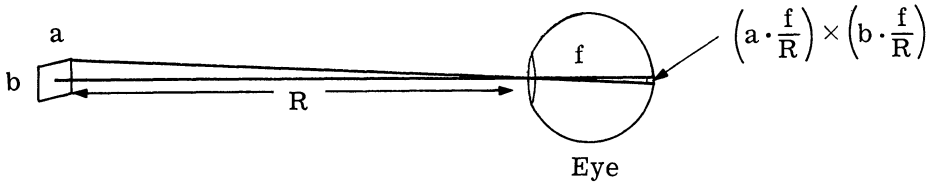


FIGURE 5-1. IMAGE OF A DISTANT ELEMENT OF BRIGHT AREA

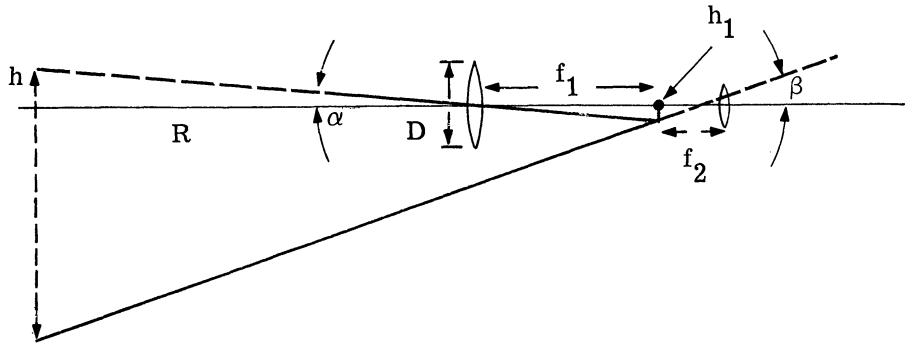


FIGURE 5-2. SIMPLE REFRACTOR TELESCOPE SCHEMATIC

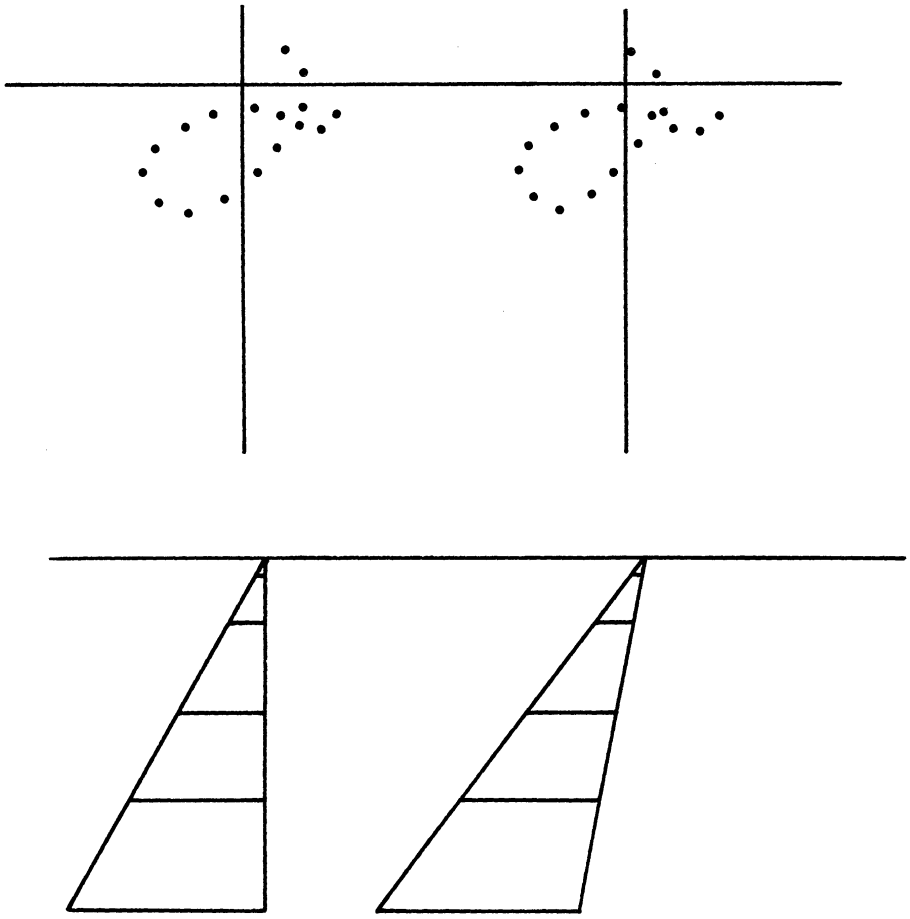


FIGURE 5-3. STEREO PLOTTED DATA

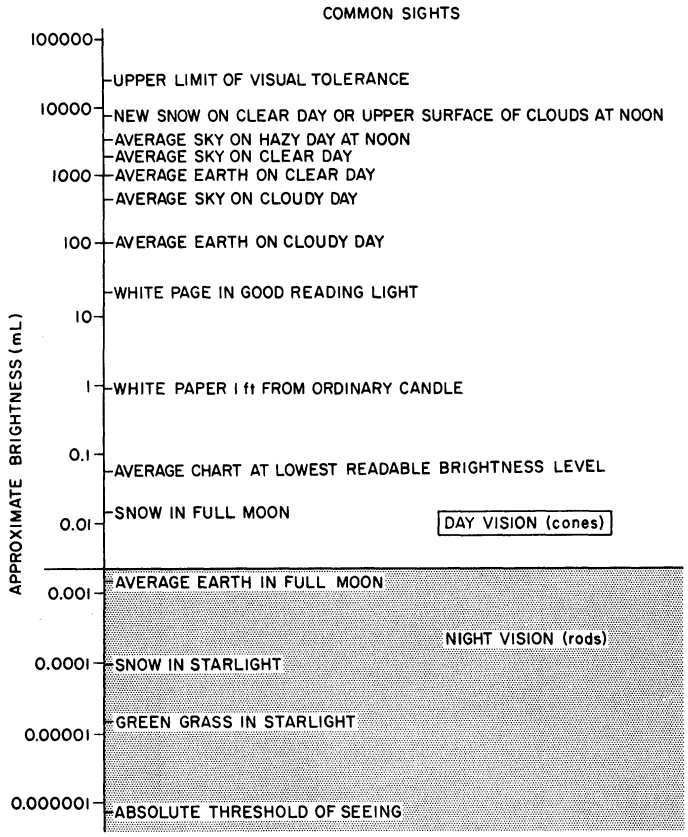


FIGURE 5-4(a). COMMON SIGHTS

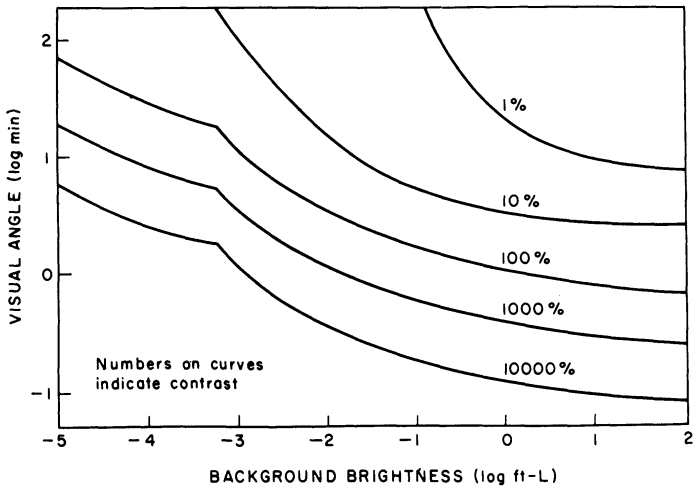


FIGURE 5-4(b). BACKGROUND BRIGHTNESS

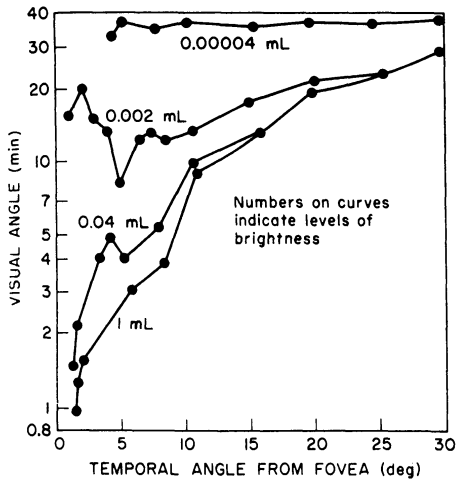


FIGURE 5-4(c). TEMPORAL ANGLE FROM FOVEA

From Human Engineering Guide to Equipment Design by Morgan et al. Copyright 1963, McGraw-Hill. Used with permission of McGraw-Hill Book Company.

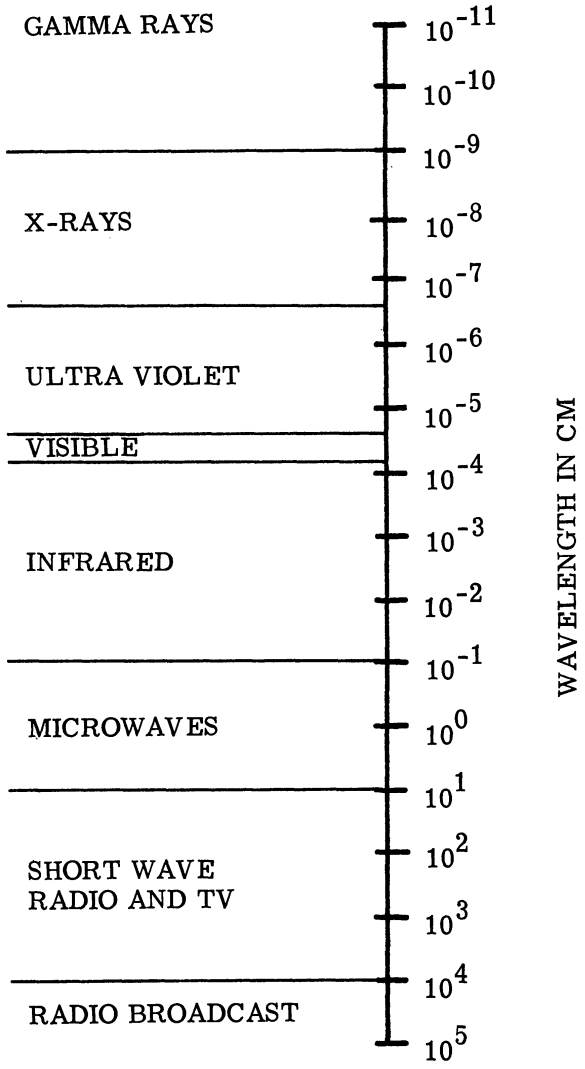


FIGURE 6-1. THE ELECTROMAGNETIC SPECTRUM

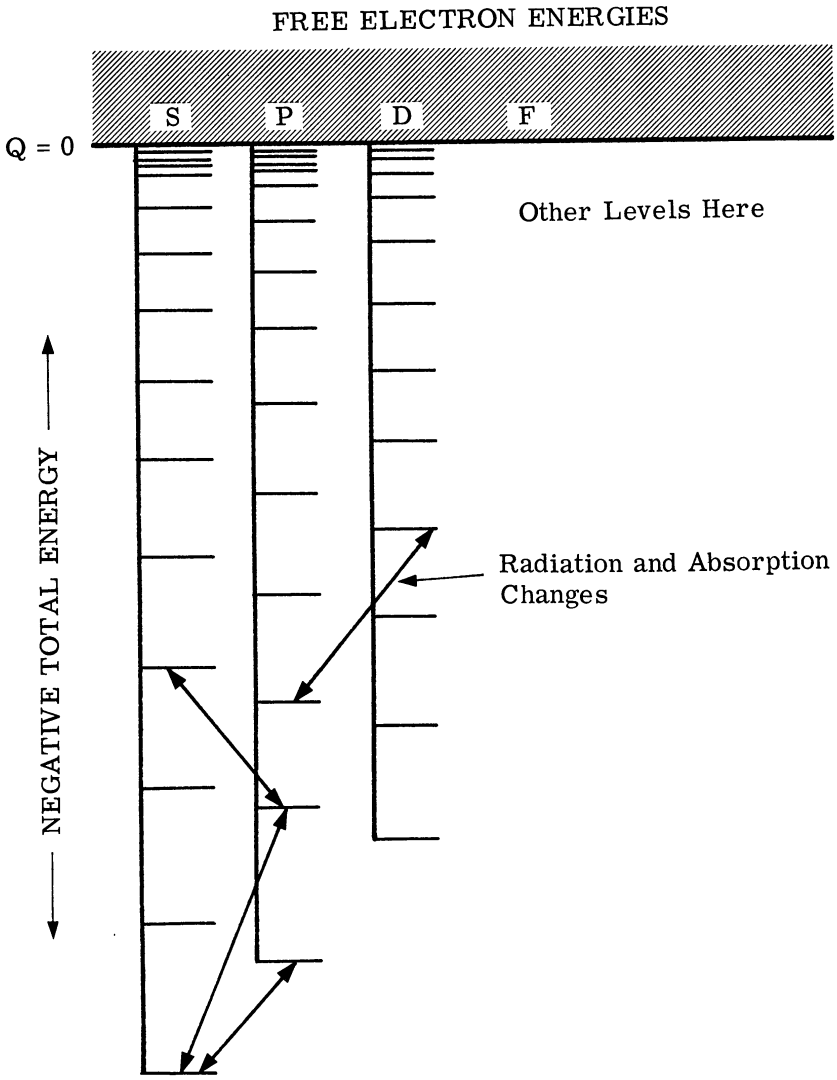


FIGURE 6-2. HYDROGEN-LIKE ATOMIC ENERGY LEVELS

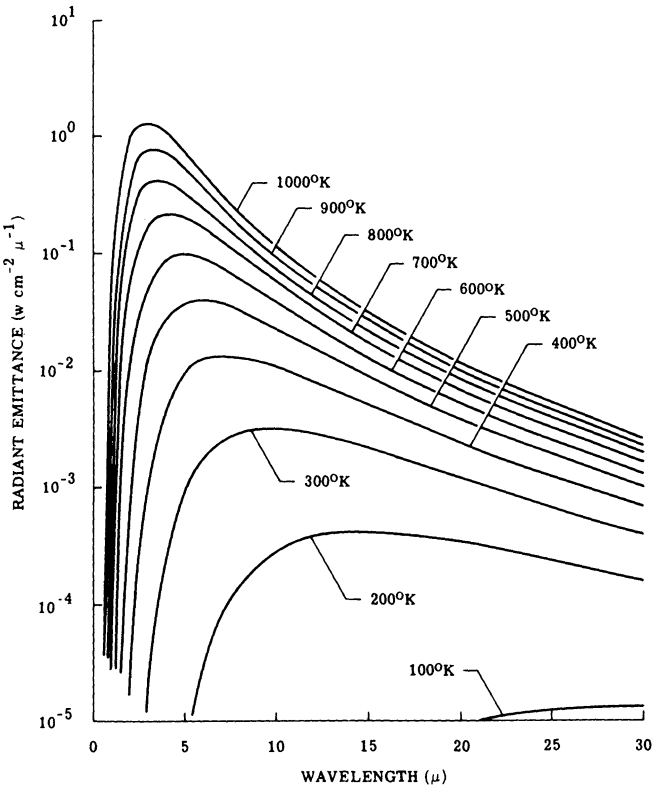


FIGURE 7-1. BLACKBODY CURVES FOR 100°K TO 1000°K

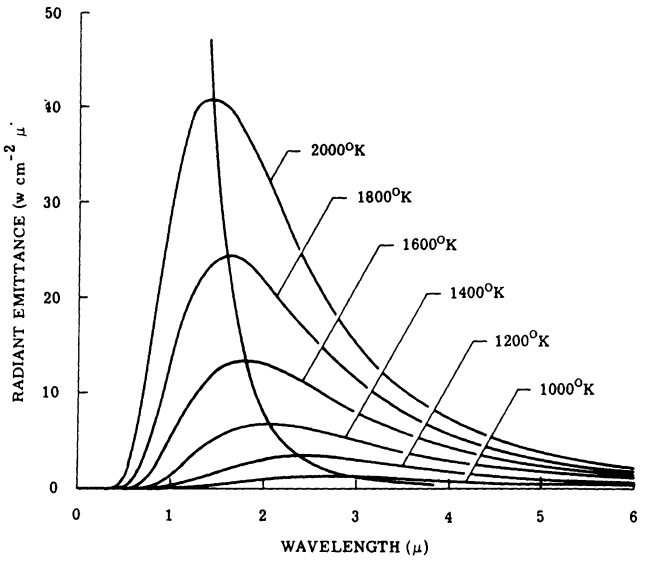


FIGURE 7-2. BLACKBODY CURVES FOR 1000°K TO 2000°K

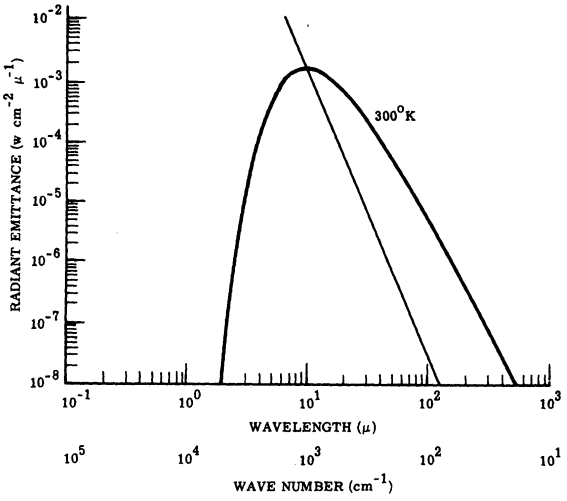


FIGURE 7-3. W_λ VS. WAVELENGTH

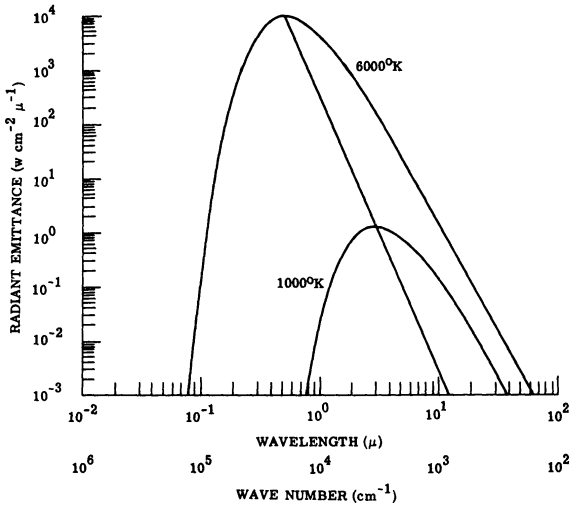


FIGURE 7-4. W_λ VS. WAVELENGTH

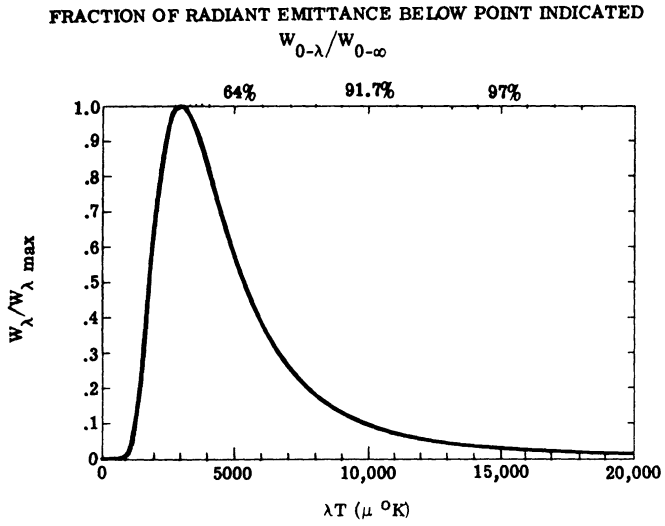


FIGURE 7-5. $W_\lambda/W_{\lambda\ max}$ VS. λT FOR A BLACKBODY

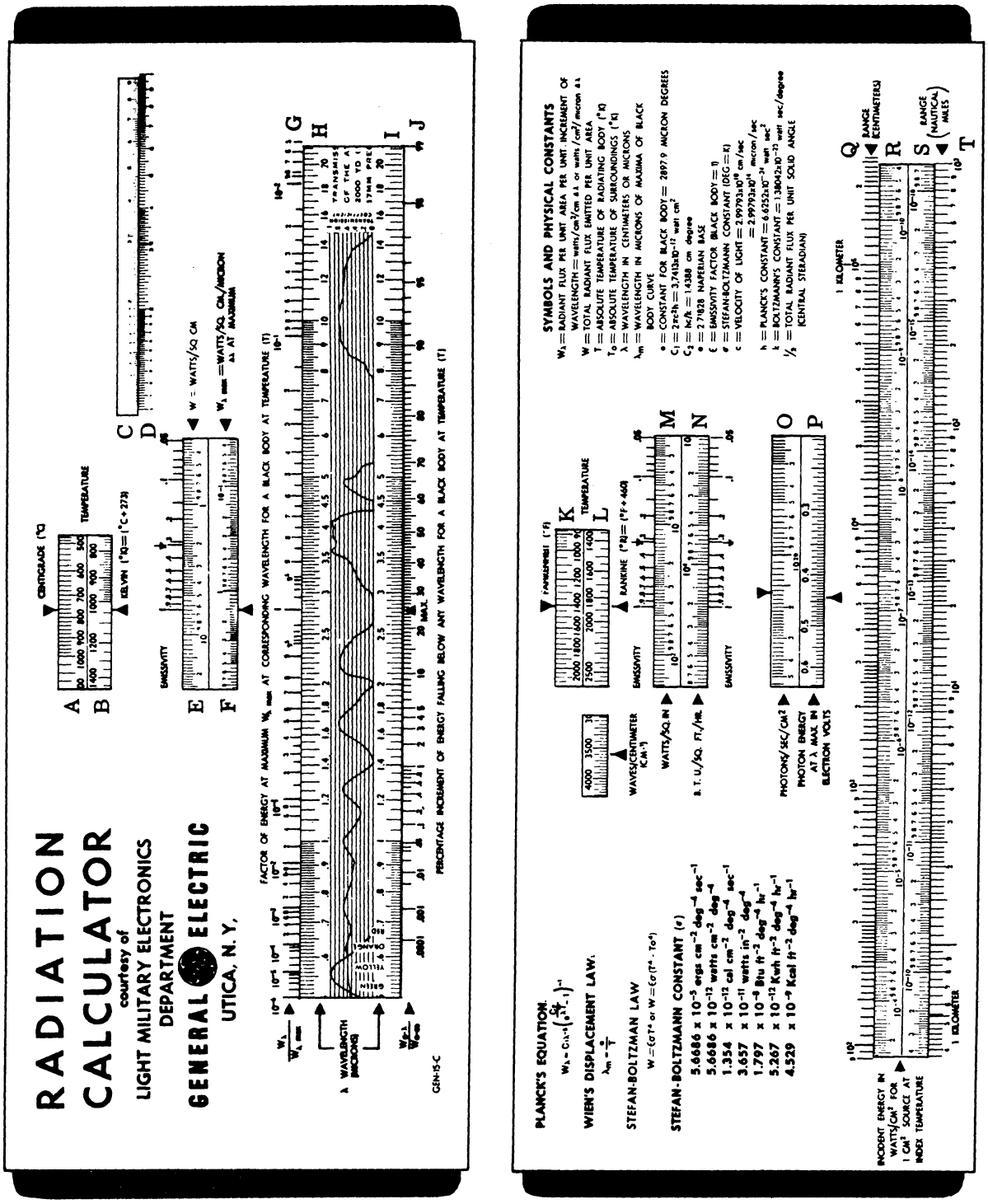

RADIATION CALCULATOR
courtesy of
LIGHT MILITARY ELECTRONICS DEPARTMENT
GENERAL ELECTRIC
UTICA, N. Y.
GEN-15-C
CENTIGRADE (°C)
TEMPERATURE
KELVIN (°K)= (°C+273)
EMISSIVITY
W = WATTS/SQ CM
FACTOR OF ENERGY AT MAXIMUM
AT CORRESPONDING WAVELENGTH FOR A BLACK BODY AT TEMPERATURE (T)
WAVELENGTH (MICRONS)
PERCENTAGE INCREMENT OF ENERGY FALLING BELOW ANY WAVELENGTH FOR A BLACK BODY AT TEMPERATURE (T)
FAHRENHEIT (°F)
RANKINE (°R)= (°F+460)
WAVES/CENTIMETER (CM⁻¹)
WATTS/SQ IN
B.T.U./SQ. FT./HR.
PHOTONS/SEC/CM²
PHOTON ENERGY AT λ MAX IN ELECTRON VOLTS
INCIDENT ENERGY IN WATTS/CM² FOR 1 CM² SOURCE AT INDEX TEMPERATURE
1 KILOMETER
RANGE (CENTIMETERS)
RANGE (NAUTICAL MILES)
PLANCK'S EQUATION
WIEN'S DISPLACEMENT LAW
STEFAN-BOLTZMAN LAW
STEFAN-BOLTZMANN CONSTANT (σ)
5.6686 x 10⁻⁵ ergs cm⁻² deg⁻⁴ sec⁻¹
5.6686 x 10⁻¹² watts cm⁻² deg⁻⁴
1.354 x 10⁻¹² cal cm⁻² deg⁻⁴ sec⁻¹
3.657 x 10⁻¹¹ watts in⁻² deg⁻⁴
1.797 x 10⁻⁶ Btu ft⁻² deg⁻⁴ hr⁻¹
5.267 x 10⁻¹² Kwh ft⁻² deg⁻⁴ hr⁻¹
4.529 x 10⁻⁹ Kcal ft⁻² deg⁻⁴ hr⁻¹
SYMBOLS AND PHYSICAL CONSTANTS
W = TOTAL RADIANT FLUX EMITTED PER UNIT AREA
T = ABSOLUTE TEMPERATURE OF RADIATING BODY (°K)
T₀ = ABSOLUTE TEMPERATURE OF SURROUNDINGS (°K)
λ = WAVELENGTH IN CENTIMETERS OR MICRONS
BODY CURVE
e = 2.7182 NAPERIAN BASE
ε = EMISSIVITY FACTOR BLACK BODY (= 1)
c = VELOCITY OF LIGHT = 2.99793x10¹⁰ cm/sec
= 2.99793x10¹⁴ micron/sec
h = PLANCK'S CONSTANT = 6.6252x10⁻³⁴ watt sec²
k = BOLTZMANN'S CONSTANT = 1.38042x10⁻²³ watt sec/degree
½ = TOTAL RADIANT FLUX PER UNIT SOLID ANGLE (CENTRAL STERADIAN)

FIGURE 7-6

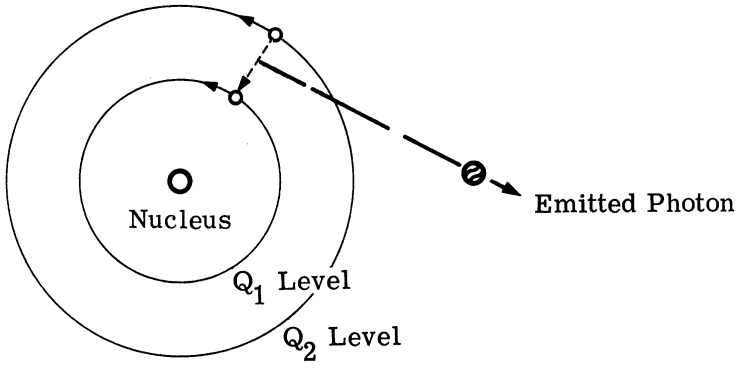


FIGURE 8-1. PHOTON EMISSION

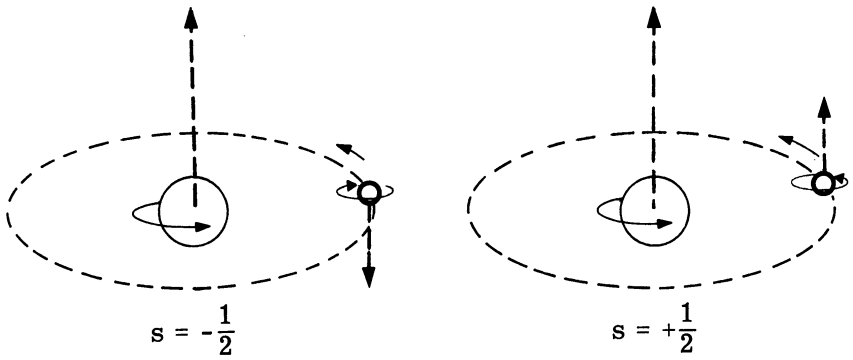


FIGURE 8-2. QUANTIZED ELECTRON SPIN

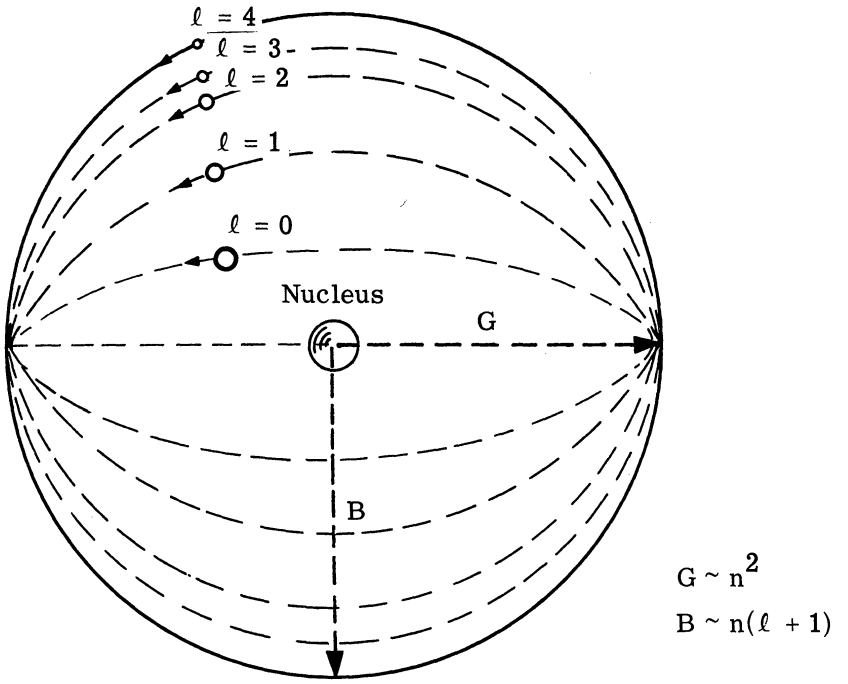


FIGURE 8-3. THE FIVE ORBITALS FOR THE n = 5 ENERGY STATE

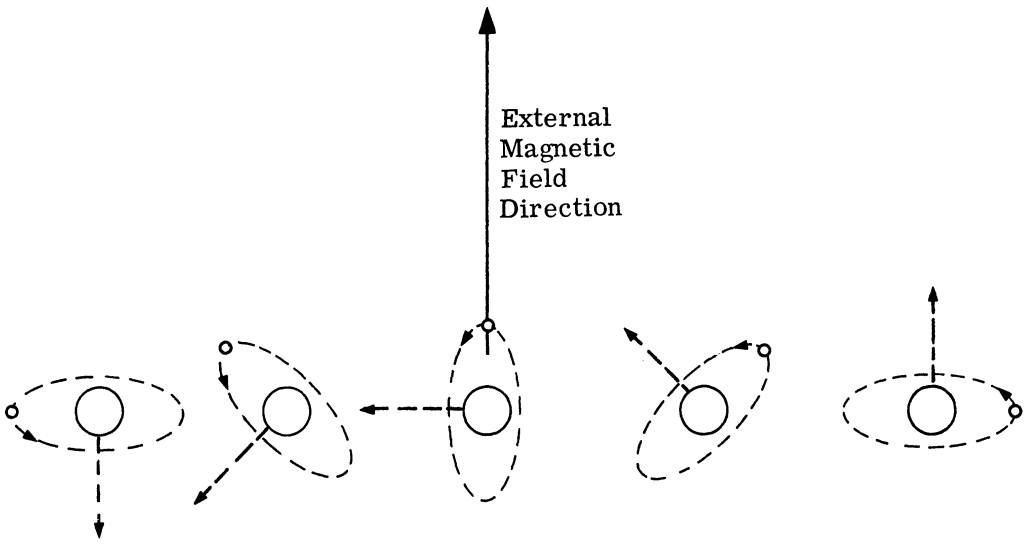


FIGURE 8-4. THE FIVE MAGNETIC QUANTUM NUMBERS FOR THE ℓ = 2 ORBITAL

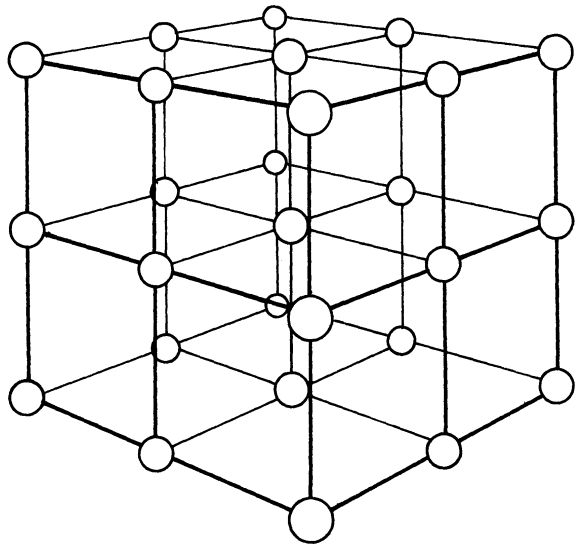


FIGURE 8-5. A SOLID WITH A SIMPLE CUBIC LATTICE. The black spheres represent the atoms and the straight lines joining them are imaginary. For the sake of clarity, the size of the atom has been reduced in comparison with the distance between atoms. In actual fact, the atoms are almost touching. Also, they oscillate rapidly about the positions shown.

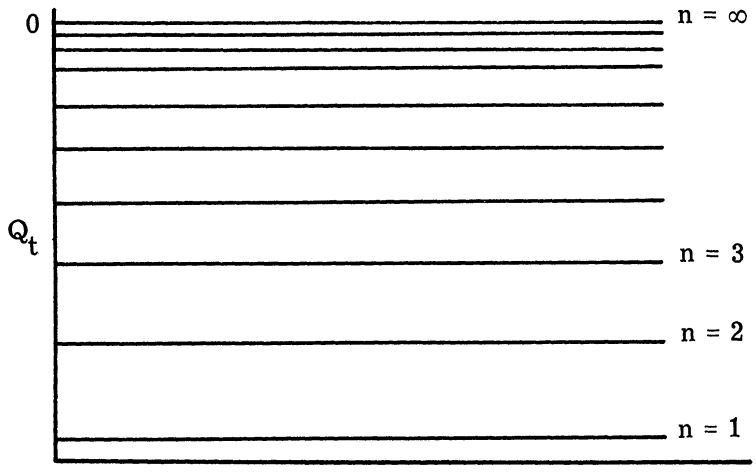


FIGURE 8-6. THE ENERGY STRUCTURE OF A SINGLE ATOM

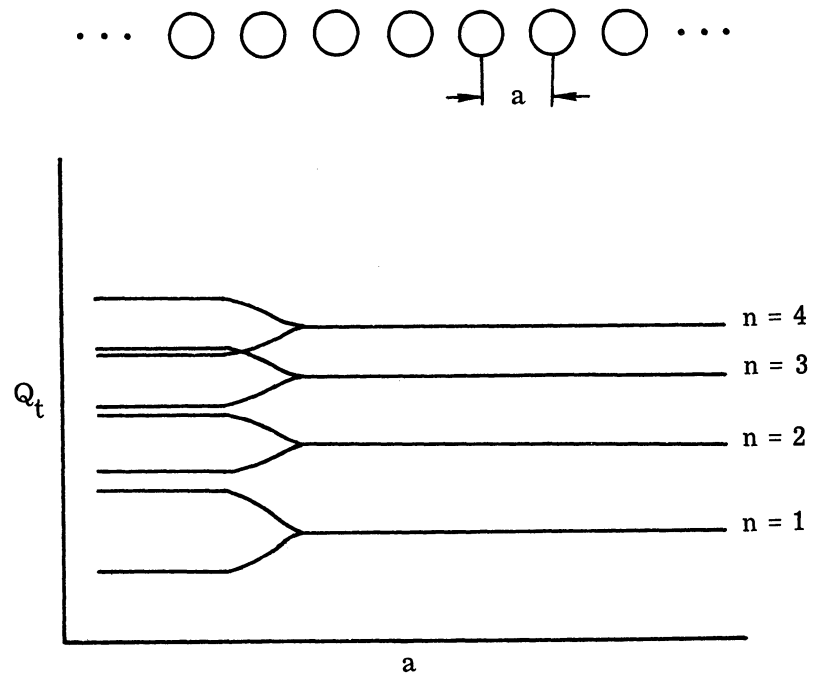


FIGURE 8-7. VARIATION OF ENERGY LEVEL WITH SPACING DISTANCE a

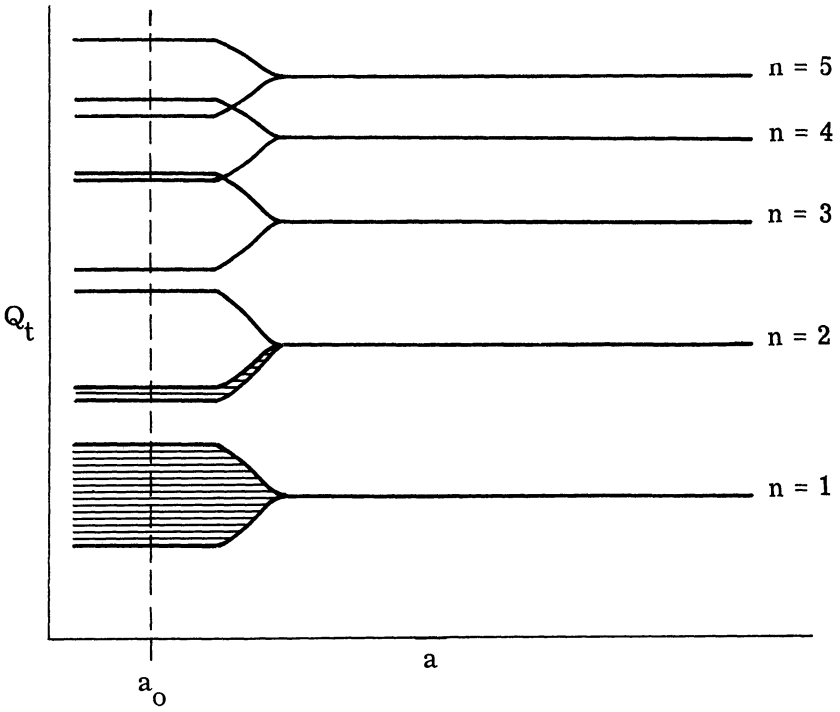


FIGURE 8-8. FILLING OF ENERGY LEVELS IN THE LITHIUM ATOM

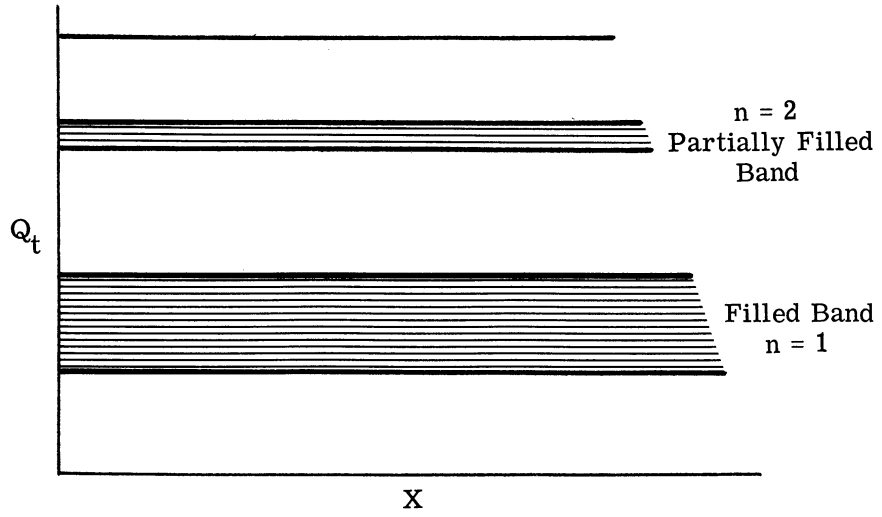


FIGURE 8-9. ENERGY LEVEL DIAGRAM ALONG THE CHAIN

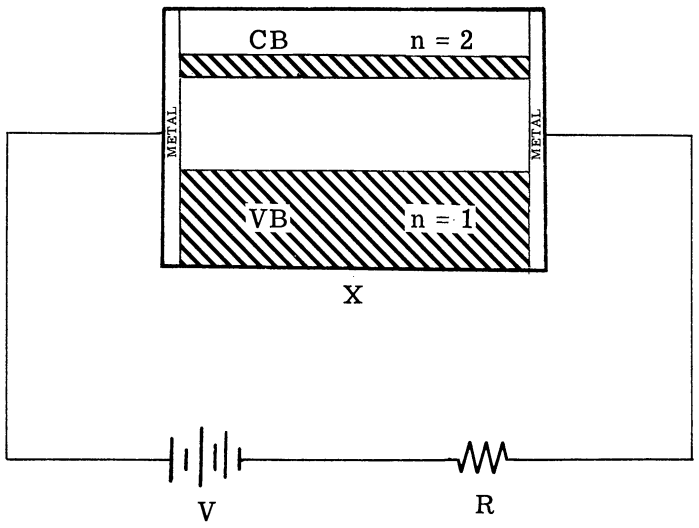


FIGURE 8-10. LITHIUM IN AN ELECTRICAL CIRCUIT

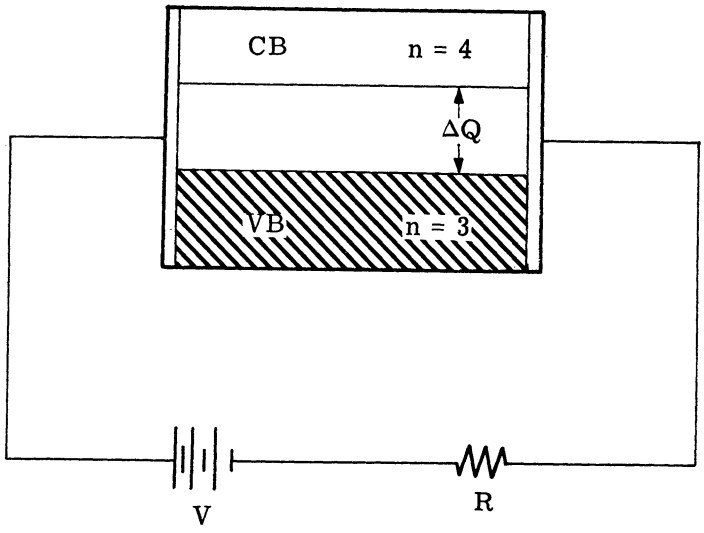


FIGURE 8-11. ENERGY BAND PICTURE OF AN INSULATOR

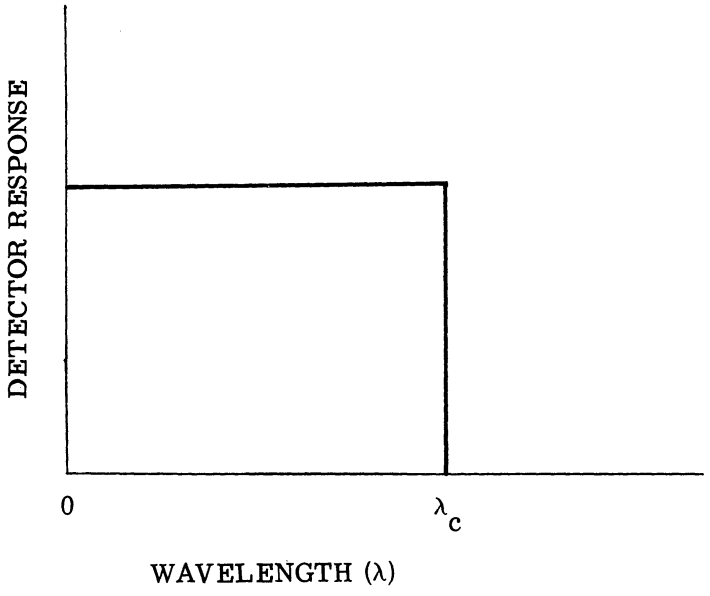


FIGURE 8-12. DETECTOR RESPONSE TO EQUAL PHOTON NUMBERS

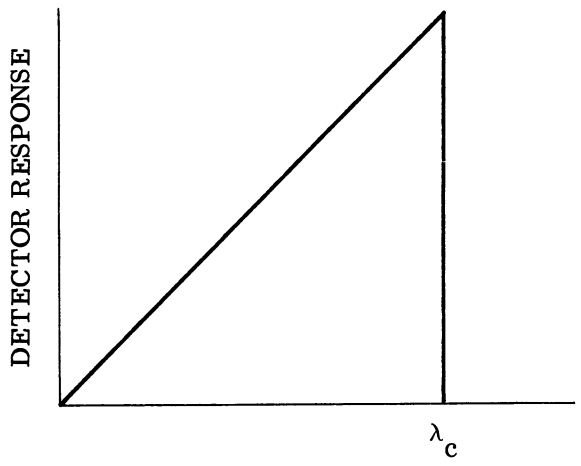


FIGURE 8-13. DETECTOR RESPONSE TO EQUAL RADIATION POWER

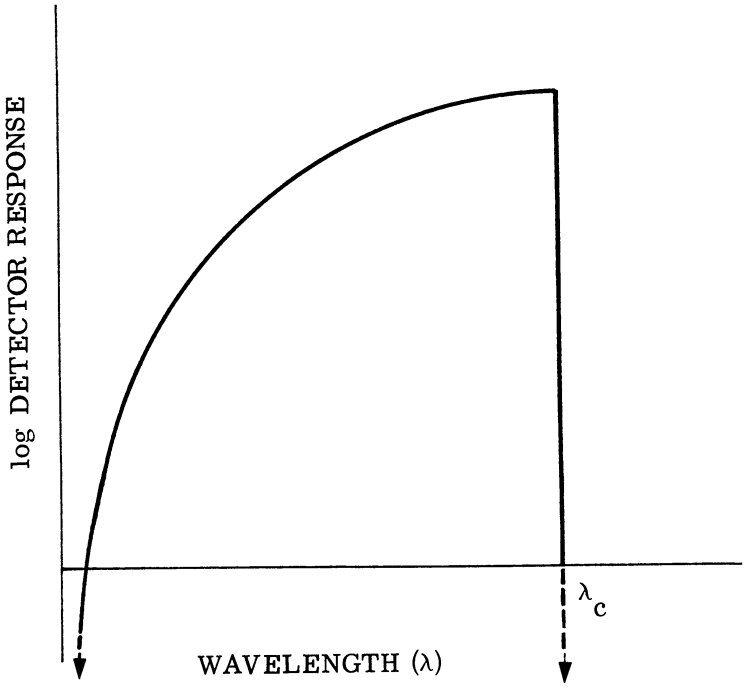


FIGURE 8-14. WAVELENGTH λ DETECTOR SPECTRAL RESPONSE

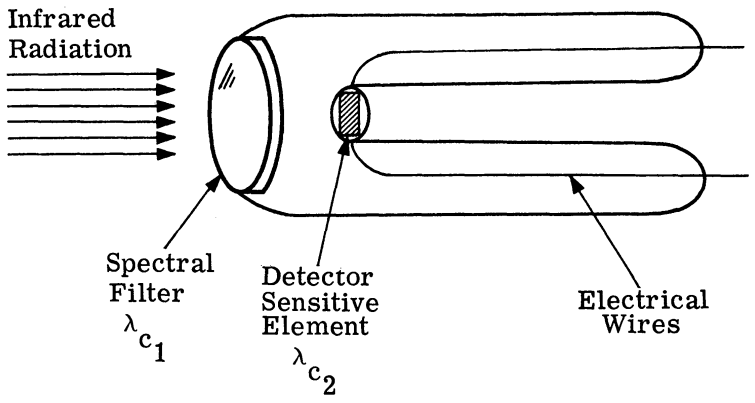


FIGURE 8-15. INFRARED DETECTOR

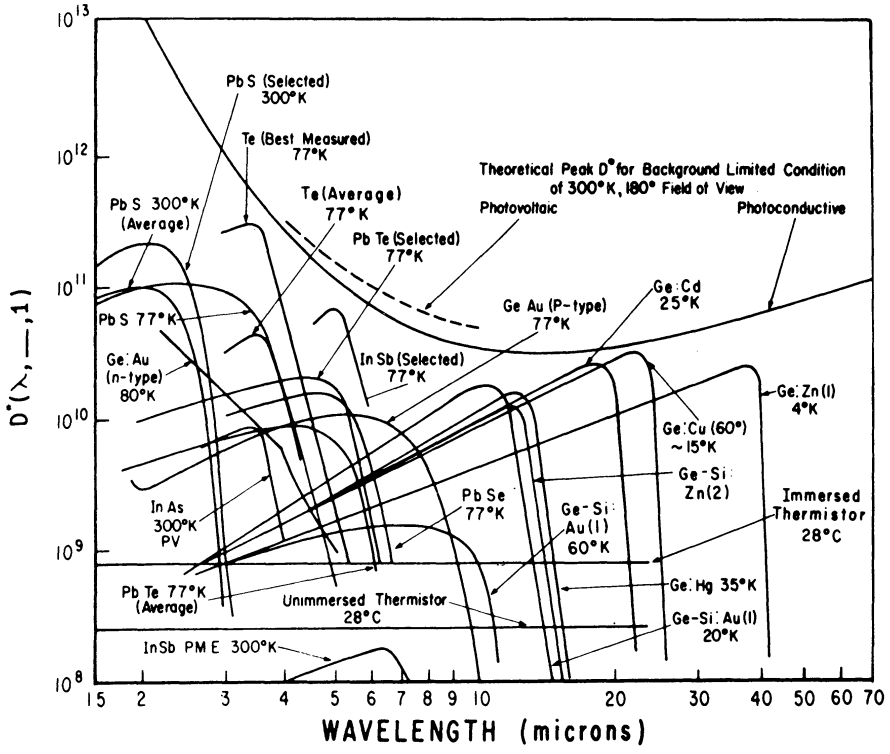
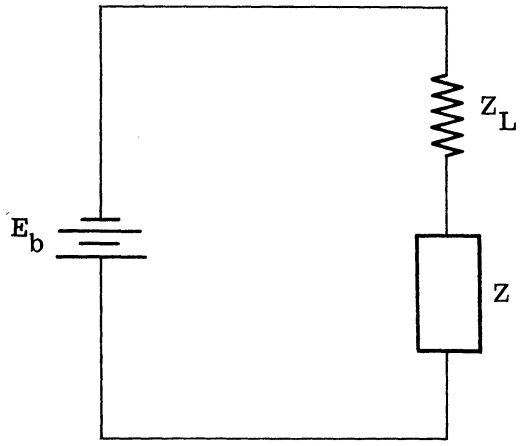
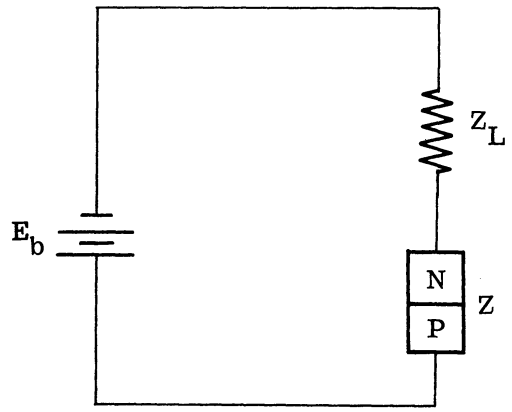


FIGURE 8-16. SPECTRAL D* FOR A NUMBER OF COMMERCIALLY AVAILABLE DETECTORS



(a) For Photoconductive Elements



(b) Backbiasing for Photovoltaic Elements

FIGURE 8-17. TYPICAL BIAS CIRCUITS

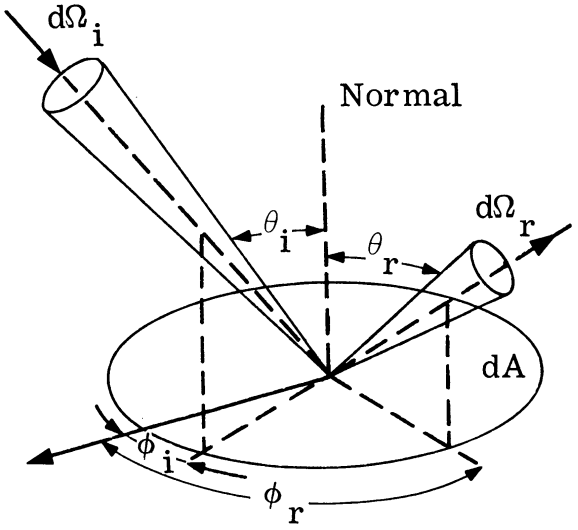


FIGURE I-1. ANGLES FOR INCIDENT AND REFLECTED RADIATION

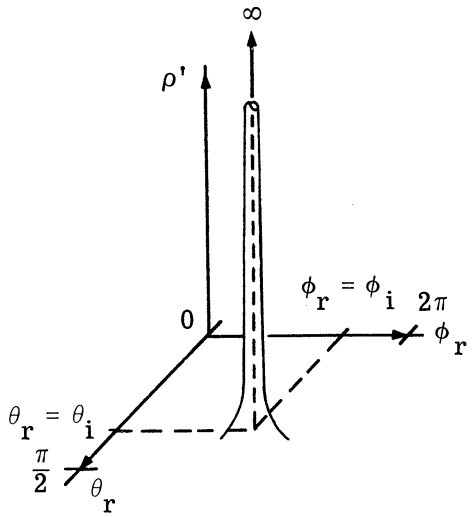


FIGURE I-2(a). PLOT OF ρ' VS. θ_r AND ϕ_r FOR A SPECULAR REFLECTOR

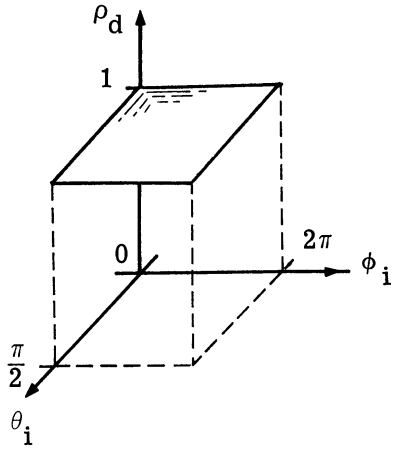


FIGURE I-2(b). PLOT OF ρ_d FOR A PERFECT SPECULAR REFLECTOR

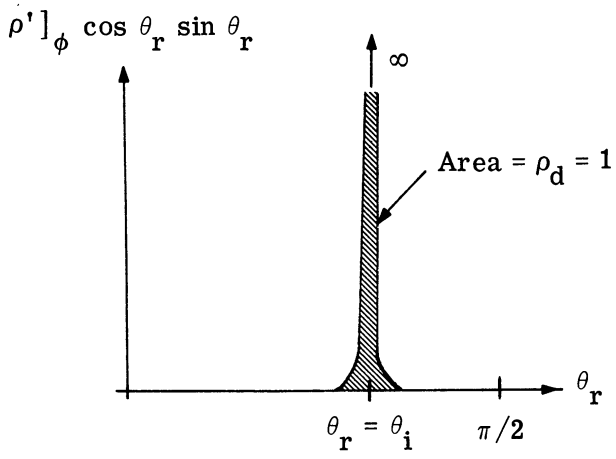


FIGURE I-3. $\rho']_\phi \cos\theta_r \sin\theta_r$ VS. θ_r

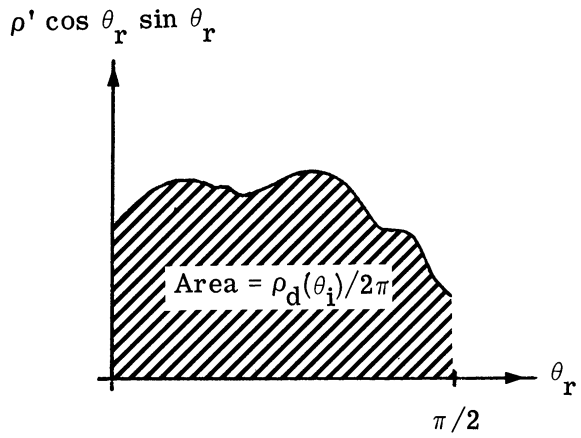
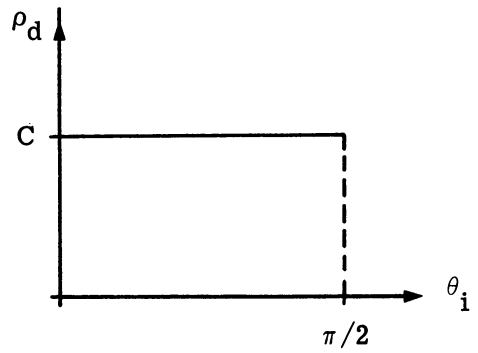


FIGURE I-4. ARBITRARY $\rho' \cos\theta_r \sin\theta_r$ VS. θ_r



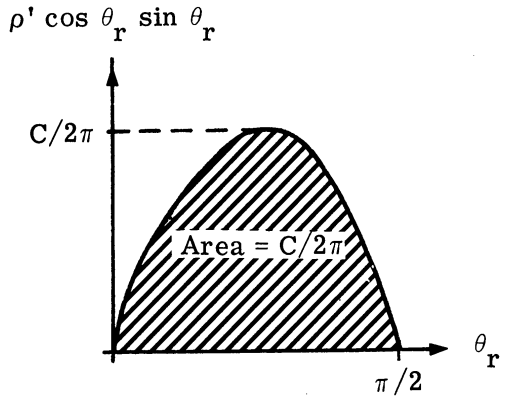


FIGURE I-5. ρ_d AND $\rho' \cos\theta_r \sin\theta_r$ FOR A LAMBERTIAN REFLECTOR

SECTION II
RADIOMETRY FOR REMOTE SENSING

1. INTRODUCTION

Radiometry is the measurement of some radiometric quantity such as E or L. Measurement and calibration are so intertwined in meaning that a review of definitions is valuable. Measurement is the process by which a quantitative comparison is made between two or more states of a physical observable such as length, temperature, or spectral radiance. If we define the performance of an instrument as a measure of its usefulness to make unambiguous quantitative comparison, then we may define calibration as the special measurement process by which one determines all parameters significantly affecting an instrument's performance. Further, if subsequent measurements can be referred to standard values, such as those maintained by the U. S. National Bureau of Standards, then the calibration process is called absolute calibration. It can be said that calibrations are performed to make the results of measurement as independent as possible of the particular measuring instrument employed.

The radiometric measurements of interest in remote sensing include the determination of the intensity, I, or the radiance, L, of some source which is not accessible for direct measurement. However, the output of a radiometric instrument, after calibration has been performed, is determined by the incident radiation which is being measured. The calibration is to allow determination, within experimental error, of the significance of all possible instrumental outputs in terms of the incident radiation that is being measured. This point of view has the following ramifications: first, the phrase "incident radiation that is being measured" logically leads to the selection of the instrument's receiving aperture as the reference position for incident radiation, since this is the radiation that would affect any other instrument similarly placed. Second, "the significance of all possible instrumental outputs," requires a complete determination of the instrument's three performance characteristics, i.e., responsivity, detectivity, and reference radiation. Finally, the reference to experimental error indicates the need to obtain estimates of the accuracy (precision and systematic error) of the results of calibration measurements as the basis for similar estimates of the accuracy of subsequent measurement results.

There remains the problem of calculation of the desired radiometric quantity at some point remote to the instrument by use of the values determined at the radiometer aperture. For example, the determination of the radiant intensity of a star requires, in addition to measurement of the irradiance at the observatory radiometer aperture, knowledge of the effects of the intervening atmosphere and a value of the distance to the star.

The challenge of radiometry stems, for the most part, from the desire to completely state the significance of all possible instrument outputs. Radiometers measure the magnitude

of radiometric quantities within some region of the wavelength-time-space-polarization domains created by the dependency of the radiation upon these parameters.

2. RADIOMETRY NOMENCLATURE

A radiometer is any apparatus which is either designed for or used in the appropriate manner for quantifying some property of incident radiation. An ordinary 35mm camera used to take a picture of Uncle Joe for the snapshot album is not a radiometer. However, the same camera used with radiance standards, spectral standards, or dimensional standards to establish the quantity of radiation being recorded from Uncle Joe would be a radiometer. There are a variety of terms used to describe radiometric devices in the literature. These terms are used purely to indicate design intentions in making these devices. The terms are generally combinations of the prefix and suffix words shown in the following table.

Prefix	Suffix
radio	meter
photo	graph
spectro	scope
monochrom	

The combinations of these terms are not always consistent. Roughly speaking, the implications are as follows:

"radio" refers to electromagnetic radiation in general.

"photo" refers to visible radiation.

"spectro" refers to the division of radiation in accordance with the simple harmonic components of the radiation.

"monochrom" refers to monochromatic radiation.

"meter" implies that a measurement is indicated but not necessarily recorded.

"graph" implies that a measurement is recorded as a picture by the device.

"scope" implies that the radiation is to be viewed by the eye through the device.

Hence, a radiometer as a generic term would be any device used to measure some property of incident radiation. In a more specific sense it is a device used to measure irradiance or radiance in some fixed spectral band. Thus, additional prefixes specify the spectral region of the band, e.g. infrared radiometer, ultraviolet radiometer, microwave radiometer. The term photoradiometer is not used. Photometer is used instead but with the additional implication that the relative spectral response is the same as that of the visibility curve. For definitions of spectrometer, spectroradiometer, and spectrophotometer, see chapter 19, "Infrared Measuring Instruments," Handbook of Military Infrared Technology. A number of illustrative diagrams are shown along with a table of available devices.

A spectrograph is a spectrometer that produces a pictorially recorded spectrum as an inherent part of its function. Usually, the term spectrograph is used to imply that the recording is done specifically by exposure of photographic film to the spectrum in the image plane. The output recording is a spectrogram. A recording spectrometer is a spectrometer which usually makes a graphical plot of power as a function of wavelength as opposed to a photograph of the spectrum. A spectroscope is a spectrometer whereby the spectrum is viewed with the eye.

A monochromator differs from a spectrometer in that it is constructed in such a way that a narrow spectral band of radiation may be selected at will from a broad band spectrum so that only that narrow band is brought to focus outside the cover of the instrument. That power can be used as a monochromatic source for other experiments.

Certain modifying terms are used with these names. Double-pass refers to the use of an optical component twice by causing the radiation to pass through it twice in the same direction. A double-pass spectrometer disperses the radiation by sending radiation through a prism first forward and reflected back through once, then reflected from another mirror so as to enter forward and back through a second time. A double monochromator is a monochromator composed of two monochromators used in series. The output of the first forms the input to the second. A double monochromator is useful in selecting weak spectral bands from a source whose spectral power is very large in adjacent but unwanted spectral ranges. These instruments are all, to some degree, selectively responsive. If one seeks the variation of the amount of radiation as a function of any of the spatial, spectral, or temporal parameters, then an instrument is used which responds to a selectable narrow range of values of the parameters and the total range of interest is scanned. The narrow range is specified by statement of the instrumental resolution. Obviously the specifications are not independent. One cannot, for example, arbitrarily reduce $\Delta\lambda$ and Δf together. Practically, one runs out of radiation! The limits are determined by the signal-to-noise ratio attainable for the specific experiment.

3. BASIC RADIOMETER PROPERTIES

The components which are basic to any radiometer are:

(1) Optics which collect the radiation through an aperture of area A, disperse or filter the radiation spectrally, and focus it on a field stop of area a.

(2) Detector which transduces the radiation which comes from the field stop to a signal, s, (usually an electrical signal of magnitude volts).

(3) Electronics/Recorder which process (amplify) the signal and record it.

The simplest type of radiometer is one in which the detector element forms both the aperture stop and the field stop with the entrance pupil, exit pupil, entrance window and exit window coinciding with these stops. No optical elements precede the detector. No amplification of output signal is used. The output signal value is indicated directly by an electric current meter. The ordinary photographic exposure meters are of this kind. Recall that the responsivity of a detector is defined as the change in output signal, ΔS, divided by the change in incident power, ΔP, on the detector. For radiometers as a complete system it is useful to extend the meaning of responsivity to refer also to the change in output signal, ΔS, divided by a change in irradiance, ΔE, at the entrance pupil of radiometer, or to the change in output signal, ΔS, divided by the change in target-area radiance, ΔL. Thus, for a radiometer three kinds of instrument responsivity can be considered,

$$R_P = \frac{\Delta S}{\Delta P}, \; R_E = \frac{\Delta S}{\Delta E}, \text{ and } R_L = \frac{\Delta S}{\Delta L}.$$

The detectivity of a detector is

$$D = \frac{R}{S_n}$$

where S_n is the rms noise fluctuation of the detector output. For a radiometer, one may wish to consider the radiometer detectivity

$$\frac{R_P}{S_n}, \; \frac{R_E}{S_n}, \text{ and } \frac{R_L}{S_n}.$$

In many radiometers in use today the output signal is an electrical voltage so that in the following discussion the output signal will be symbolized by V.

The three major radiometric performance characteristics are defined as (1) R, the responsivity—the change in output per unit change in input of incident radiation, either $\Delta V/\Delta E$ or $\Delta V/\Delta L$; (2) either D, the detectivity which is the reciprocal of the noise-equivalent incident

(input) radiation or the noise voltage, V_n; and (3) the reference radiation, E_0 or L_0, the level of incident radiation corresponding to a zero reading on the output scale of the instrument. Other more directly obtained characteristics, such as the wavelength calibration of a scanning spectrometer, are also needed for interpretation of the output of more specialized instruments. The output voltage of a radiometer as a function of E at the radiometer aperture is illustrated in figure 3-1. The equation for the function is

$$V = R_E(E - E_0) + V_n$$

where $R_E \equiv \frac{\Delta V}{\Delta E}$

$E_0 \equiv$ reference irradiation.

V_n = output noise voltage.

The reference irradiance, E_0, is measured using an external adjustable calibrated source. The source is adjusted until the radiometer output reads zero volts. The value of E at the radiometer aperture due to the adjustable source can then be calculated to establish the value of E_0, provided that the noise voltage is negligible small. Frequently, the values of E to be measured are quite large so that

$$E >> E_0$$

and

$$V >> V_n.$$

Then the relation may be approximated well enough by:

$$V = R_E E.$$

A similar relation can be derived for extended sources using R_L.

Of the three defined characteristics the responsivity is by far the most significant since it is always involved in transforming an output "reading" into the corresponding value of incident radiation input. Frequently it is sufficient to know the rough order of the detectivity or noise voltage if the measurements to be made are so far above the instrument noise level that noise fluctuations are negligible. Similarly a rough measure of the reference radiation is often sufficient if E_0 is small. However, just because V_n and E_0 are often negligible, these last two characteristics are easily overlooked in those rather rare instances when noise fluctuations and reference radiation may be highly significant factors in the interpretation of an instrument output. For this reason, they should always be checked.

The relationships between R_P, R_E, and R_L are derived from the relation between P, E, and L. Thus:

$$R_P \equiv \frac{\Delta V}{\Delta P} = R_E A^{-1} = R_L [Th]^{-1}$$

where $Th \equiv \int \cos\theta \, d\Omega \, dA$.

("Th" is often called the instrumental "through put".) The $\cos\theta$ factor comes from the definition of $L \equiv \frac{\partial^2 P}{\cos\theta \, \partial A \partial \Omega}$

4. NORMALIZATION AND THE IDEAL RADIOMETER

Since the radiance and irradiance are functions of wavelength, space - location and direction, polarization and time, the radiometer responsivity must also be a function of these parameters. The spatial dependency defines what is called the field of view; the spectral dependency, the spectral response; and the dependency upon time, the fluctuation frequency response or bandwidth. Radiometer designs aim at making these dependencies independent factors so that the responsivity is the product of several relative responsivity factors and a normalization constant, R_0.

$$R = R_0 \cdot r_1(\lambda) \cdot r_2(f). \text{ etc.}$$

In the ideal radiometer case the responsivity is zero outside of the geometrical bounds established by a solid angle Ω, and is constant within these bounds. Similarly,

$$r_1(\lambda) = 1 \text{ for } \lambda_1 \leqq \lambda \leqq \lambda_2$$

$$r_1(\lambda) = 0 \text{ outside of } \Delta\lambda = \lambda_2 - \lambda_1$$

and

$$r_2(f) = 1 \text{ for } f_1 \leqq f \leqq f_2$$

$$r_2(f) = 0 \text{ outside of } \Delta f = f_2 - f_1$$

The ideal radiometer will produce the nearest to the true values of E or L that can be obtained for a given spectral, spatial, and temporal resolution. If we used an ideal radiometer to measure the value of L between two wavelengths, λ_1, and λ_2, then the spectral responsivity, $R_L(\lambda)$, would appear as graphed in figure 4-1a. The value of L between wavelengths λ_1 and λ_2, which we want to know, is shown by the area under the curve for L_λ in figure 4-1b. The voltage output of the radiometer is equal to the area under the product curve shown in figure 4-1c. Mathematically,

$$V = \int_0^\infty R_L(\lambda) L_\lambda \, d\lambda.$$

Since $R_L(\lambda) = R_{0,L}$ (which is constant between λ_1 and λ_2 and is otherwise zero), then

$$V = \int_{\lambda_1}^{\lambda_2} R_L(\lambda) L_\lambda \, d\lambda = R_{0,L} \int_{\lambda_1}^{\lambda_2} L_\lambda \, d\lambda$$

The factor remaining to be integrated is the value of L we want.

$$\int_{\lambda_1}^{\lambda_2} L_\lambda \, d\lambda = L$$

Hence

$$V = R_{0,L} L$$

However, many radiometers are not ideal in spectral responsivity. Yet, it is still required to obtain a direct relation between voltage output and the value of radiance between some two wavelengths. In using a real radiometer we will be able to measure $R_L(\lambda)$ as a property of the radiometer in laboratory tests. Now, turning our radiometer toward some unknown source we will obtain a voltage V. We may estimate from the context of the experiment the qualitative nature of L_λ of the unknown. Suppose that we have good reason to believe that L_λ is nearly constant in value in the spectral range where the radiometer responsivity is not zero. We may then expect

$$V = \int_0^\infty R_L(\lambda) L_\lambda \, d\lambda \cong L_\lambda(\lambda_m) \int_0^\infty R_L(\lambda) \, d\lambda$$

where $L_\lambda(\lambda_m)$ is the value of spectral radiance at the wavelength, λ_m, where the radiometer has the maximum response. Now one can find wavelengths limits

$$\lambda_m - \frac{\Delta\lambda}{2} \text{ and } \lambda_m + \frac{\Delta\lambda}{2}$$

for an ideal radiometer having the same maximum responsivity as the actual radiometer which would give the same value of V by requiring

$$\int_{\lambda_m - \frac{\Delta\lambda}{2}}^{\lambda_m + \frac{\Delta\lambda}{2}} R_{0,L}(\text{ideal}) \, d\lambda = \int_0^\infty R_L(\lambda) \, d\lambda$$

Since

$$\int_{\lambda_m - \frac{\Delta\lambda}{2}}^{\lambda_m + \frac{\Delta\lambda}{2}} R_{0,L} \, d\lambda = R_{0,L} \Delta\lambda$$

then

$$R_{0,L} \Delta\lambda = \int_0^\infty R_L(\lambda) \, d\lambda$$

or

$$\Delta\lambda = \frac{\int_0^\infty R(\lambda)\,dy}{R_{0,L}}.$$

Thus, we could conclude that the radiance is related to the output voltage of the real radiometer by

$$V \cong R_{0,L}[L_\lambda(\lambda_m)\Delta\lambda] = R_{0,L}L$$

but for power in the spectral band width $\Delta\lambda$ centered at λ_m. Notice that the spectral band width $\Delta\lambda$ is not the spectral band width $(\lambda_2 - \lambda_1)$ of the actual radiometer but is the spectral bandwidth of an equivalent ideal radiometer having the same maximum responsivity as the actual radiometer.

This procedure can also be shown graphically. In figure 4-2a, a radiometer spectral responsivity is shown to be triangular in shape.

A triangular shape is common in some spectroradiometers. In figure 4-2b the unknown spectral radiance L_λ of the source is shown. The value of L_λ between λ_1 and λ_2 is not quite constant but it is not changing much. The value of V which the actual radiometer would exhibit is equal to the area under the product curve in figure 4-2c. The product curve is not quite triangular in shape because L_λ is not quite constant. The equivalent ideal radiometer spectral responsivity which has the same maximum responsivity as the actual radiometer is shown as a dotted rectangle in figure 4-2a. It is that rectangle of height, $R_{0,L}$ which has the same area as the triangle. That ideal radiometer would produce a product curve as shown by the dotted line in figure 4-2c. The area under the dotted product curve is equal to the value of V which an ideal radiometer would provide. Notice that these two product curve areas are approximately equal and would be exactly equal if L_λ were truly constant.

It is not difficult to show that nonideal radiometers can provide quite accurate readings compared to their equivalent ideal radiometers in cases where L_λ changes even rapidly but linearly in the interval λ_1 to λ_2 provided only that $R_L(\lambda)$ is symmetrical about the maximum response wavelength, λ_m. Suppose that $L_\lambda = A(\lambda - \lambda_m) + L_\lambda(\lambda_m)$ in the interval λ_1 to λ_2 and that $R_L(\lambda_m + \lambda) = R(\lambda_m - \lambda)$. The voltage V of the radiometer is then

$$V = \int_0^\infty R_L(\lambda)[A(\lambda - \lambda_m) + L_\lambda(\lambda_m)]\,d\lambda$$

where A is the slope of the L_λ curve. It follows that

$$V = \int_0^\infty R_L(\lambda)A(\lambda - \lambda_m)\,d\lambda + \int_0^\infty R_L(\lambda)L_\lambda(\lambda_m)\,d\lambda$$

Now we know that $A(\lambda - \lambda_m)$ is antisymmetrical about the wavelength, λ_m, because when λ is greater than λ_m, the value of $A(\lambda - \lambda_m)$ is the same size but opposite in sign from the value of $A(\lambda - \lambda_m)$ when λ is an equal amount smaller than λ_m. However, the spectral responsivity is symmetrical. It has the same size and sign at equal wavelengths intervals away from λ_m. Hence the product curve in the first integral must be antisymmetrical and have, therefore, an area of zero. The remaining term is then

$$V = \int_0^\infty R_L(\lambda) L_\lambda(\lambda_m)\, d\lambda,$$

and thus just as for an ideal radiometer

$$V = L_\lambda(\lambda_m) \int_0^\infty R_L(\lambda)\, d\lambda = R_{0,L}[L_\lambda(\lambda_m)\, \Delta\lambda].$$

The equivalent ideal radiometer will produce the identical output voltage.

One can see from this that a real radiometer with a symmetrical spectral response curve will provide good (ideal radiometer) readings for many kinds of unknown sources.

The determination of the equivalent ideal radiometer having the same maximum spectral responsivity as the real radiometer is called normalizing the radiometer response to the peak or maximum. It is quite possible to utilize other equivalent ideal radiometers in the normalization process. One can easily determine the peak spectral responsivity of the ideal radiometer whose spectral band width is limited by the wavelengths for which the actual radiometer spectral response has fallen to one half its maximum value, $R_L(\text{ideal}) = \dfrac{\int_0^\infty R_L(\lambda)\, d\lambda}{(\Delta\lambda \text{ for half response})}$. The values would then be stated as that radiance within the half spectral responsivity points.

Normalization to specify the equivalent ideal radiometer responsivity for the other parameters such as field of view and fluctuation frequency bandpass follows analogous reasoning. Likewise, requiring symmetry of the responsivity for these other parameters permits good accuracy for the linear variation of radiance with these other parameters for the same reasons.

5. CALIBRATION STANDARDS

In the previous discussion, it was assumed that a standard source of radiance or irradiance was available so that calibration or accurate measurement of $R(\lambda)$ could be performed. Now, we must turn our attention to the problem of standards. Standards are necessary in order to assure consistent quantification from time to time by one observer and from observer to observer. Thus, a standard must be both independent of the time of use and easily accessible and replicated for use by others. There are two possible ways of providing radiation standardization. One is to provide a standard source, the other is to provide a standard receiver. Clearly, if we have a standard source we can calibrate a receiver and use that as a secondary standard to calibrate other sources. If we have a standard receiver we can specify the properties of a source and use the source as a secondary standard to calibrate other receivers.

The trend has been to seek first a standard source, the blackbody, for standard irradiance. For most remote sensing purposes, such standards are easily replicated and stable in time. The blackbody source is most easily employed for microwaves, infrared and long-wavelength visible radiation and becomes difficult to use for radiation of shorter wavelengths because the amount of short-wavelength radiation produced by blackbody sources at practical temperatures is much too small. A stable high temperature greybody source, the tungsten lamp, is used for visible and ultraviolet radiation standards. The greybody source can be calibrated by the use of a low-temperature blackbody standard in the overlapping wavelength range where both can provide adequate power. For wavelengths shorter than the ultraviolet, standard receivers are commonly used.

5.1. BLACKBODY SOURCES

The basic form of a blackbody standard is a cavity in the form of a sphere, cylinder or cone. Probably because of the difficulty of manufacture and the difficulty of maintaining a uniform temperature over the internal walls, the sphere is not widely used. The cylinder and cone are used, usually in combination, to effect the optimum ratio of internal cavity dimensions to aperture area. The more common shapes among manufactured sources of conventional design (i.e., $T < 2000^{\circ}C$) are: (1) the cone terminating in a cylindrical section at the end to limit the size of the opening; and (2) the cone terminating in an inverted conical section truncated to provide a limited opening.

Blackbody standard sources can be operated theoretically from as low a temperature as desired to high temperatures theoretically limited only by the melting or subliming point of the core material. The core material is selected to withstand the rigors of operation. Most cavity wall materials are made of the oxide of some hard metal. For very low temperatures

the limits on operation are imposed by the point at which frost forms on the cavity walls. To avoid frost the cavity must be used in a vacuum or synthetic atmosphere. The complexity of designing a blackbody for high temperature is imposed by the need for reducing the loss of heat from the cavity and maintaining a uniform temperature across the inner surface of the cavity.

The cavity temperature can usually be read on an instrument which has a temperature-sensing element inserted in a drilled hole. The bottom of the hole is near the inner wall of the cavity and in the conical cavity at some point near the apex of the cone. The accuracy of the blackbody exitance calculation using such a temperature reading depends on many things including the accuracy of the temperature sensing element itself, the degree to which the reading actually represents the wall temperature, and the degree to which the wall temperature at that point reflects the temperature of the whole cavity. It is not difficult to construct a cavity which approximates a blackbody to within a fraction of a percent. What is difficult, however, is to measure accurately the effective radiating temperature of the cavity.

The thermostatic control element controlling power to the cavity is inserted in another drilled hole. It is not often certain what the accuracy of the control to the power input is. Since the temperature is most difficult of all parameters to control, and since the radiation is dependent on the fourth power of the absolute temperature, the temperature of the cavity is often the parameter which is the most critical in determining the accuracy of the blackbody standard.

In order to make blackbody sources which operate above 2000°C, a great deal of effort must go into maintaining constancy of temperature and keeping the surface intact. Usually a noncorrosive atmosphere must be introduced. Above 3000°C the usual techniques of producing blackbody radiation are no longer practical. The carbon arc will provide a reasonably uniform source of radiation at near 4000°C with a sacrifice in constancy in temperature and a precise knowledge of emissivity.

5.2. STANDARD LAMP SOURCES

5.2.1. IRRADIANCE SOURCE. The sources of standard spectral irradiance are tungsten-filament quartz-iodine lamps for the wavelength range of 0.25 to 2.6 μm. The lamps are commercial GE type 6.6A/T4Q/1CL-200-watt lamps having a tungsten coiled-coil filament enclosed in a small (1/2 in. × 2 in.) quartz envelope containing a small amount of iodine.

The spectral radiant intensity of the entire lamp as mounted in the manner prescribed below is measured and recorded for each lamp. The spectral irradiance from these lamps is based upon the spectral radiance of a blackbody as defined by Planck's equation and has been determined through comparison of a group of quartz-iodine lamps with (1) the NBS standards

of spectral radiance, (2) the NBS standards of luminous intensity, and (3) the NBS standards of total irradiance.

The lamp is mounted vertically with the NBS-numbered end of the lamp down and the tip away from the detector. Measurements of distance (from lamp filament) are made along a horizontal axis passing through the center of the lamp filament. The correct vertical position is determined by setting the centers of the upper and lower seals along a plumb line as viewed from one side of the lamp. The plane of the front surface of the lower press seal is set to contain the horizontal perpendicular to the line connecting the lamp filament center and detector.

The lamp is mounted in the supplied holder which is constructed in such a manner as to reflect a negligible amount of radiant flux in the direction of the radiometer or spectrometer slit. A black shield should be placed at a distance of about 3 ft to the rear of the lamp to intercept stray radiant flux along the radiometric axis and adequate shielding should be provided to intercept stray flux from other directions.

If there is excessive water vapor in the laboratory atmosphere, errors may result at the wavelengths of water-vapor absorption bands. In the original calibrations the comparisons of the lamps with the other NBS standards were made in such a manner that the effect of water-vapor absorption was eliminated.

Values of spectral irradiance for these lamps are tabulated as a function of wavelength in microwatts per square centimeter-nanometer at a distance of 43 cm from center of lamp to receiver. Values of spectral irradiance for wavelength intervals other than one nanometer, say x nanometers, may be found by multiplying the tabulated values by x.

Standards of spectral irradiance require no auxiliary optics. If any are employed, proper correction must be made for their optical characteristics. The lamp is simply placed at a measured distance from the detector or spectrometer slit. If a distance other than 43 cm is used, the inverse-square law may be used to calculate the spectral irradiance (the inverse-square law should not, however, be used for distances shorter than about 40 cm).

In measurements wherein two sources (a standard source and a test source) are being compared by the direct substitution method (slit widths kept unchanged, use of the same detector), no knowledge of the spectral transmittance of the spectrometer, nor of the spectral sensitivity of the detector, is required. It is necessary, however, to make sure that the entrance slit of the spectrometer is fully and uniformly filled with radiant flux both from the standard and from the test source; and if at any one wavelength the detector response for the standard is significantly different from that for the test source, the deviation from linearity of response of the detector must be evaluated and taken into account. Furthermore, if the standard and test source differ in geometrical shape, it must be ascertained that the instrument transmittance

and detector response are not adversely affected thereby. Many detectors are highly variable in sensitivity over their surface area and may require diffusion of radiant flux over their surface to insure accurate radiant-energy evaluations.

All lamp calibrations are made by the use of alternating current and it is recommended that they be so used in service. To reduce the line voltage to lamp operating voltage a 10-ampere variable autotransformer may be employed for coarse control. For fine control a second (5 amp) variable autotransformer may be used to power a radio-filament transformer whose secondary (2.5-5 V) winding is wired in series with the primary of the 10-amp transformer. It was found that this method is very effective for accurate control of the 6.50-amp lamp current.

Standards of spectral irradiance are expensive laboratory equipment and it is suggested that they be operated sparingly and with care in order to prolong their useful life. They should be turned on and off at reduced current and great care should be taken so that at no time will the current appreciably exceed 6.50 amperes. It is recommended that for general use, working standards be prepared by calibrating them relative to the laboratory standard supplied by NBS.

These standard lamps operate at high temperature so that the quartz envelope is above the flammable point of organic materials. A hot lamp may cause fires or burn lint or dust on the envelope resulting in optical damage to its surface. In no case should the fingers come into contact with the quartz envelope, either hot or cold, because the resulting fingerprints will burn into its surface during lamp operation.

5.2.2. RADIANCE SOURCE. The sources of standard spectral radiance are ribbon-filament lamps for the wavelength ranges of 0.25 to 0.75, 0.5 to 2.5 and 0.25 to 2.5 μm. The lamps are commercial GE type 30A/T24/7 lamps having a tungsten ribbon filament (SR-8A type) centered about 8 to 10 cm behind a fused silica window 3 cm in diameter.

The portion of the filament whose spectral radiance has been determined is the central portion visible through the fused-silica window. This determination was made by direct substitution of the lamps for working standard lamps which had in turn been calibrated by this substitution method relative to blackbodies (operated from 1400 to 2300^{o}K) through the use of a double quartz prism spectroradiometer and associated electronic equipment.

In operation the lamp is mounted vertically and the beam of radiant flux with a horizontal axis passing through the center of the filament is measured. In the original determination no portion of the beam measured departed from this axis by more than 2.5 degrees. Hence, if an aperture subtending a larger angle is required in any application of these standards of spectral radiance, it should be ascertained that the irradiance is constant over the whole aperture.

If there is excessive water vapor in the laboratory, atmospheric errors may result at the wavelengths of water-vapor absorption bands. In the original calibrations the comparisons of the lamps with the blackbodies were made at the same distance and in such manner that the effect of water-vapor absorption canceled out.

If the calibrations wherein the blackbody was heated within a wire-wound furnace and temperatures around 1400°K were reached, the temperatures were measured with Pt-Pt 10% Rh thermocouples. Tests using couples placed at various positions and observations with an optical pyrometer indicated closely uniform temperatures within the blackbody enclosure. A ratio of blackbody opening to total internal surface area equal to approximately 0.003 and an internal surface reflectance less than 0.10 indicate an emissivity of 0.999 or higher.

For the blackbody temperatures above 1400°K a graphite enclosure heated by a radio-frequency generator was employed and the temperatures were measured by an optical pyrometer. The physical characteristics of this blackbody indicated an emissivity approximating 0.996.

The spectral radiance of the blackbody is based upon the Planck radiation law in which the constants, based upon the most recent atomic and other information, are set down as follows:

$$c_1 = 1.19088 \times 10^{-12} \text{ watt cm}^2 \text{ per steradian}$$

$$c_2 = 1.4380 \text{ cm degree K.}$$

Values of spectral radiance for these lamps are tabulated as a function of wavelength in microwatts per (steradian-millimicron-square millimeter of filament). Values of spectral radiance for slit-widths other than one millimicron, say x millimicrons, where x is less than 100, may be found by multiplying the tabulated values of x.

It is suggested that the auxiliary optics employed with these standards of spectral radiance be composed of two units: namely, a plane mirror and a spherical mirror (each aluminized on the front surface). If the spherical mirror is placed at a distance from the lamp filament equal to its radius of curvature, and the plane mirror set about 1/3 to 2/5 this distance from the spherical mirror, facing it and so placed that the angle between incident and reflected beams is 10° or less, a good image of the filament itself may be focused upon the spectrometer slit. Little distortion of the filament image occurs provided precise optical surfaces are employed and angles between incident and reflected beams are kept to less than 10°. Various optical arrangements may be employed.

The solid-angular aperture of the auxiliary optics should be smaller than the solid-angular aperture of the spectrometer employed so that no loss of radiant energy will result through over-filling the spectrometer optics.

The spectral radiant flux, P_λ, which enters the spectrometer slit is computed from the formula:

$$P_\lambda = \rho(\lambda) L_\lambda s A/d^2$$

where $\rho(\lambda)$ is the spectral reflectance of the combination of mirrors used, L_λ is the reported spectral radiance of the standard, s is the area of the spectrometer slit, A is the area of the limiting auxiliary optic, and d is the distance of this optic from the slit.

No diaphragm or other shielding is required in the use of these standards, except for a shield to prevent radiant energy from the lamp from entering the spectrometer directly without first falling on the concave mirror. An image of the filament should be focused upon the spectrometer slit, and only the power from this image should enter the slit.

In order to calibrate a spectroradiometer with one of these standards of spectral radiance, a knowledge of the spectral reflectance of the mirror surfaces is required. A good aluminized surface should have a spectral reflectance considerably above 0.87 throughout the spectral region of 0.5 to 2.6 μm, increasing slightly with wavelength except possibly for a slight dip near 0.80 μm. In practice the proper reflectance losses can best be determined through the use of a third (similar) mirror (a second plane mirror) which may be temporarily incorporated into the optical setup from time to time.

In measurements wherein two sources (a standard source and a test source) are being compared by the direct substitution method (use of the same auxiliary optics, slit-widths, areas and detector at any one wavelength) no knowledge of the spectral reflectance of the auxiliary mirrors, nor of the spectral transmittance of the spectrometer, nor of the spectral sensitivity of the detector is required. It is necessary, however, to make sure that the entrance slit of the spectrometer is fully and uniformly filled with radiant flux both from the standard and from the test source; and, if at any one wavelength the detector response for the standard is significantly different from that for the test source, the deviation from linearity of response of the detector must be evaluated and taken into account.

Operation of these standards should be on alternating current to obviate filament-crystallizing effects that occur when the operation is on direct current. The filaments are massive and "iron out" all effects of the normal fluctuations present in a commercial ac supply. All calibrations were made by means of alternating current. To reduce line voltage a step-down transformer (1 kva) having a ratio of 10 to 1 or a 50-amp variable transformer may be employed. Then to give fine control a second variable transformer (10-amp capacity) is wired into the circuit to control the input of the heavy duty transformer. For still finer control a third variable transformer may be employed with a radio-filament transformer to add (or subtract) voltage fed into the step-down transformer. It was found that this method is

very effective for accurate control of large lamp currents. The heavy duty (1 kva) step-down transformer is preferred to the 50-amp variable transformer since the latter is subject to contact damage when operated for long intervals of time at high currents.

The lamp standards of spectral radiance are expensive laboratory equipment and it is suggested that they be operated sparingly and with care in order to prolong their useful life. This precaution applies especially to the standards calibrated in the short-wave region and operated at 35 amp. They should be turned on and off slowly and only for short intervals should they ever be operated at or above 30 amp, and then only to calibrate a similar lamp as a working standard. In general even at lower currents a working standard should be prepared and used except for purposes of checking the operation of such a working standard.

5.3. STANDARDS OF WAVELENGTH

Radiation standards for wavelength calibration of spectrometers are sources of gaseous atomic and molecular radiations or are absorption filters using the gaseous absorption characteristics of atoms or molecules. The Handbook of Chemistry and Physics lists large numbers of the wavelengths of known spectral lines with more than adequate accuracy for remote sensing purposes. The common commercial fluorescent lamp utilizes mercury vapor. The gaseous discharge line spectrum for mercury is observable through the fluorescent coating. These lines form excellent wavelength standards in the visible range for remote sensing purposes. Absorption cells filled with ammonia vapor or carbon dioxide serve well for the infrared range. Certain plastics exhibit characteristic absorption bands due to certain chemical bonds which can be identified and used for wavelength standards. Wavelength standards for radar and microwave radiation can make use of molecular absorption, macroscopic electrical resonator absorption, and comparison with National Bureau of Standards radiation from radio station WWV in Colorado.

6. RADIOMETER CALIBRATION METHODS

Calibration may be accomplished in many ways. These may be categorized into two groups—those in which the radiometer field stop is partially illuminated by the calibration source, and those in which the field stop is totally filled with uniform calibration radiation. The first may be achieved by a distant point (small) source, or by a collimator with suitable magnification and a small source. The latter can be realized with a so-called extended source, either distant or close (at the entrance aperture), or by proper placement of a point source very close to the entrance aperture(the so-called "Jones method" of calibration). In any actual measurement situation one may overlook the total significance of some parameter and the size of its effect on the result. For this reason a good general rule is always to carry out a calibration under conditions which reproduce as completely as possible the situation expected when measurements based on that calibration are made. On the other hand, for maximum versatility and as a check on consistency, it is important to conduct additional calibrations covering as wide a range of variation of as many parameters as possible. Thus the calibration should be in terms of E if measurements of E are to be made with the radiometer. Using the distant point source method one would obtain, for an ideal radiometer,

$$\frac{\Delta V s^2}{\Delta I} = \frac{\Delta V}{\Delta E} = R_E \left[\frac{\text{volt}}{\text{watt-cm}^{-2}}\right]$$

where s is the distance from the source to the radiometer aperture. Here we have neglected atmospheric attenuation which is <u>not</u> negligible in actuality.

Similarly with a distant extended source of radiance L,

$$R_E = R_L \left[\int \cos\theta \, d\Omega\right] = \frac{V}{L}\left[\frac{Th}{A}\right]^{-1}$$

Here, in addition to accounting for the intervening atmosphere we also have the problem of determining Ω, the angular field of view of an ideal radiometer. If we are planning for measurements of L, then the extended source gives the desired calibration directly. When the extended source is placed at the radiometer aperture, the adjustment for atmospheric attenuation is minimized.

A near small source of area A_s and radiance L can be used to establish R_L by using the relation

$$R_L = \frac{V}{L}\frac{A}{A_s}.$$

The calibration arrangement using a near small source is shown, together with cautions in the use of this method, in Nicodemus' Chapter 8 of Volume 4 of Kinglake's Applied Optics and Optical Engineering; Academic Press.

For a more detailed discussion of radiometer calibration methods the reader should refer to "Infrared Target and Background Radiometric Measurements: Concepts, Units, and Techniques" (WGIRB), IRIA, The University of Michigan, Report 2386-64-T (AD 275 810) January 1962 or to Infrared Physics, Volume 3, No. 3, pp. 139-169, September, 1963.

7. CORRECTIONS FOR ATMOSPHERIC EFFECTS

An extensive discussion of atmospheric effects in the infrared spectral range is given in Chapter 6 of the Handbook of Military Infrared Technology. It is primarily in the infrared range where corrections for atmospheric effects become most difficult. For a purely monochromatic beam of radiation passing through a homogeneous medium it is generally possible to express the transmittance of the medium by a simple exponential formula (Beer's law).

$$\tau = e^{-k(\lambda)x}$$

The factor $k(\lambda)$ is the absorption coefficient and has units of inverse distance. The value of x is the distance from the source. Many media are characterized by $k(\lambda)$ which varies slowly and smoothly with wavelength in a number of spectral ranges. For these media and spectral ranges it is possible to find an average value, $\bar{k}$, so that

$$\tau = e^{-\bar{k}x}$$

over a modestly wide spectral band. Solids and liquids are generally the kind of medium for which good average values of absorption coefficient may be defined. The close proximity and strong interaction of interatomic forces alter the natural spectrally sharp absorption resonances of the isolated atoms so that the absorption coefficient is spectrally smooth. Gases, on the other hand, retain the spectrally sharp characteristics of the isolated atoms and molecules so that the $k(\lambda)$ can be expected to vary quite rapidly over very small spectral intervals. In the infrared range one may observe molecules having such sharply varying absorption with wavelength that even the highest resolving power laboratory instruments will have difficulty in measuring $k(\lambda)$ in detail. For gaseous media, such as the atmosphere, an average value of k cannot be defined; so, even for relatively narrow radiometer spectral ranges, Beer's law does not apply because attenuation is not a simple exponential with distance. Instead one finds that the greatest attenuation occurs nearest to the source and noticeably less attenuation occurs as the distance from the source increases. The normal course of correcting for atmospheric attenuation in remote sensing is to refer to atmospheric transmission measurements that have been made by others. If the atmosphere were unchanging, then a set of appropriate measurements of $\tau(\lambda)$ as a function of range with modest spectral resolution would be sufficient. The stable and well-mixed components of the atmosphere, O_2, N_2, and CO_2, do not cause a serious problem. The variability of H_2O in the form of water vapor is the major concern. Both the quantity and the spatial distribution of H_2O are highly variable from time to time and from place to place. The procedure for accounting for H_2O is by making a measurement at one position of the local water vapor concentration and then estimating the quantity of precipitable water in the measurement path. From previous measurements by others one may then estimate

the correction to be made. The correction made in this manner is always approximate and is usually the source of greatest variability between measurements of different observers.

The atmosphere does not consist of gases alone. Particulate matter of all kinds is found suspended in these gases. Dust, smoke, pollen, water drops, ice crystals, and salt crystals are among the debris floating in the atmosphere. The distribution and concentration of these vary widely also. The particulate matter both absorbs and scatters radiation transmitted through the atmosphere. A good rule is that particles which are much larger than the wavelength of the transmitted radiation attenuate the power in direct proportion to their cross-sectional area independently of the particular wavelength considered. Particles which are much smaller than the wavelength attenuate the transmitted power in proportion to the inverse fourth power of the wavelength.

Intercepted power which is not absorbed by the particulate matter is scattered in all directions. Radiation scattered by small particles tends to be polarized. Polarization is greatest for radiation scattered at right angles to the direction of the transmitted power. The direction of polarization is perpendicular to the plane containing the direction of view and the direction of transmitted power. Although scattering from molecules is very slight; nevertheless, scattering does occur as it would for any small particulate matter. The vast numbers of such molecular scattering particles tend to make the total scattering noticeable. Many orders of magnitude fewer solid and liquid particles can produce the same degree of scattering, however.

Figure 7-1 shows schematically the regions of transmission of the atmospheric gases as a function of wavelength. The long wave radio waves propagate with little loss. Near the end of the microwave range molecular absorption by O_2 and H_2O begins. (See Kerr p. 663). The atmosphere is not generally considered to be sufficiently transparent for remote sensing purposes between about 1 mm and 20 μm. The region between 20 μm and 1 μm is still subject to molecular absorption but significant ranges of good transmission occur. (See Handbook of Military Infrared Technology, p. 256). Molecular absorption is weak in the visible range. The ultraviolet range soon becomes useless for long ranges through the atmosphere because of molecular scattering and photochemical absorption-producing ozone. X ray transmission is hampered by atomic absorption and photochemical ionization. Low-energy gamma transmission is only good for short ranges. High energy gamma rays tend to penetrate everything to some degree, including the ground and the receiving instruments.

The effect of particulate matter on propagation varies greatly depending upon particle size and distribution. Most particulate matter has very little effect on radiation with wavelengths longer than 10 cm. Between 10 cm and 1 mm wavelength precipitation produces some attenuation and the scattering of radiation passing through precipitation begins to compete in

power level with other reflecting obstacles so that rain squalls can be detected but clouds without precipitation usually are not sufficiently strong scatters to be observed in this spectral range. At wavelengths of 10 μm, however, clouds and fogs generally produce serious attenuation and make quantitative radiometry too difficult to consider. Images of the ground can be made through thin ground fog layers if the ground itself is the principal radiator. The effect is the same as being able to read a typed page covered with several sheets of blank white paper when the typed page is back lighted instead of front lighted.

8. SPECTROMETERS

Spectrometers are instruments designed to allow determination of the wavelength distribution of radiation. Separation of the radiation into its spectral components is accomplished by the use of an optical element possessing a known functional dependence on wavelength. Thus, for example, the dependence of the index of refraction on wavelength is used in prism spectrometers. Such an optical element disperses parallel rays or collimated radiation into different angles from the prism according to wavelength.

8.1. PRISM SPECTROMETERS

First, consider the action of a prism spectrometer for monochromatic radiation, i.e., without consideration of dispersion of radiation with wavelength. Figure 8-1 shows the basic optical schematic. Note that distortion of the image of the entrance slit is minimized by the use of plane wave illumination. Even with plane wave illumination the image of the slit is curved because not all of the rays from the entrance slit can traverse the prism in its principal plane (see fig. 8-2). The prism is shown in the position of minimum angular deviation of the incoming rays in figure 8-1. At minimum angular deviation maximum power can pass through the prism.

In figure 8-1,

D is the angle of deviation
A is the angle of the prism
B is the base length of the prism
S is the prism face width
W is the width of the illuminating beam
i_1 and i_2 are angles of entry and exit
r_1 and r_2 are internal angles of refraction.

For a prism adjusted to the position of minimum deviation, it can be shown (see Jenkins and White, or Sawyer) that:

$$i_1 = i_2 = i$$

$$r_1 = r_2 = r$$

and thus

$$r = \frac{A}{2}$$

$$i = \frac{D + A}{2}$$

Now consider the effects of dispersion. The angular dispersion is defined as $\frac{dD}{d\lambda}$ while the linear dispersion is $\frac{dx}{d\lambda} = f\frac{dD}{d\lambda}$ where f is the focal length of the camera or imaging lens and x is the distance across the image plane. We know that the index of refraction, n, is a function of wavelength. A few examples are shown in figure 8-3.

The angular dispersion of the prism is given by $\frac{dD}{d\lambda} = \frac{dD}{dn} \cdot \frac{dn}{d\lambda}$

Since
$$n = \frac{\sin i}{\sin r} = \frac{\sin\left(\frac{D + A}{2}\right)}{\sin\left(\frac{A}{2}\right)}$$

$$dn = \frac{1/2 \cos\left(\frac{A + D}{2}\right)}{\sin\left(\frac{A}{2}\right)} dD$$

Thus
$$\frac{dD}{dn} = \frac{2 \sin\left(\frac{A}{2}\right)}{\cos\left(\frac{A + D}{2}\right)}$$

However, from figure 8-1 we see that

$$\frac{B}{2} = S \sin\left(\frac{A}{2}\right)$$

So that
$$\frac{B}{S} = 2 \sin\left(\frac{A}{2}\right)$$

and
$$W = S \cos\left(\frac{A + D}{2}\right)$$

so that
$$\frac{W}{S} = \cos\left(\frac{A + D}{2}\right)$$

Thus
$$\frac{dD}{dn} = \frac{B/S}{W/S} = \frac{B}{W}$$

and
$$\frac{dD}{d\lambda} = \frac{B}{W}\frac{dn}{d\lambda}$$

while
$$\frac{dx}{d\lambda} = f\frac{B}{W}\frac{dn}{d\lambda}.$$

We may define the resolving power, RP, of an instrument as the smallest wavelength difference which is resolved according to Lord Rayleigh's criterion divided into the average wavelength in that spectral region to be resolved.

Thus,
$$RP \equiv \frac{\bar{\lambda}}{\Delta\lambda}$$

The limiting resolution is set by diffraction due to the finite beam width, or effective aperture of the prism, which is rectangular. Thus, as shown for lenses, the angular half-width of the central maximum of the diffraction pattern (which equals the separation of the first minimum from the maximum) is given by division of the wavelength by the effective aperture, i.e., $\frac{\lambda}{W}$.

$$RP = \frac{\lambda}{dD}\frac{dD}{d\lambda} = \frac{\lambda}{dD}\frac{B}{W}\frac{dn}{d\lambda},$$

and by appropriate substitutions,

$$RP = \frac{\lambda}{(\lambda/W)}\frac{B}{W}\frac{dn}{d\lambda},$$

so that

$$RP = B\frac{dn}{d\lambda}.$$

If the entire prism face is not illuminated, then only the illuminated base length must be used for B.

8.1.1. NUMERICAL EXAMPLE. It is desired to know if it is possible to spectrally resolve incident radiation 0.01 μm apart near 1.0 μm wavelength with a fused silica prism with a base length of 1 cm. It is assumed that the spectrometer is properly adjusted and the slit widths may be made as narrow as required. A table of the index of refraction as a function of wavelength for fused quartz shows that $n_1 = 1.451808$ for $\lambda_1 = 0.9$ μm and $n_2 = 1.450473$ for $\lambda_2 = 1.0$ μm. We require a resolving power

$$RP = \frac{\bar{\lambda}}{\Delta\lambda} = \frac{1.0\ \mu m}{0.01\ \mu m} = 100.$$

The resolving power of the spectrometer is

$$RP = B\left|\frac{n_2 - n_1}{\lambda_2 - \lambda_1}\right| = 10^4\ \mu m\ \frac{0.001335}{0.1\ \mu m}$$

Thus, the spectrometer is capable of a resolving power of RP = 133.5. Therefore, such a resolution is within the capacity of the spectrometer. The substitution of the prism for one of the identical shape but of a different material may fail to achieve this resolution.

8.2. GRATING SPECTROMETERS

Gratings achieve dispersion by diffraction and interference effects. A particularly useful although simple method to derive some of the features is the so-called "vector method" or

method of complex number addition. We assign to the amplitude of the wave from each slit a value of $a_1 e^{i\phi_1}$. This is shown graphically in table 8-1 for two equal slits (Young's double-slit experiment).

A more interesting example is three equal slits, shown in table 8-2. The beginning treatment of six slits is shown in table 8-3. The resulting intensity distribution can be plotted in the normal way (see Jenkins and White for details). One finds for n slits, n - 1 minima and n - 2 sub maxima. In the limit, when n becomes uncountably large while $\underline{a}$ approaches zero (the sum, for $\delta = 0$, is constant in this limiting process), we find the results of table 8-4, the case for a single slit. The resulting pattern has a central maximum which is twice as wide as the successive sub-maxima. Thus the patterns derived for the cases before, 2, 3, ---, 6, ---, n slits should have single slit diffraction envelopes imposed over them. This is seen to be the case in the equation for the intensity pattern of a grating, below.

A grating is an n-slit system, treated as above, for Fraunhofer diffraction with interference arising from division of the incident plane wave front. If we let d = grating constant or spacing (the distance between adjacent slits) and w = width of any one slit while p is the order number of the principal maxima, then

$$p\lambda = d\ (\sin\phi + \sin\theta)$$

where ϕ = angle of incidence and θ = angle of diffraction. The most common case is for $\phi = 0$.

$$p\lambda = d \sin\theta$$

$$p = 0, 1, 2, 3, ---$$

and

$$E = E_0\left[\frac{\sin\left(\frac{\pi w \sin\theta}{\lambda}\right)}{\frac{\pi w \sin\theta}{\lambda}}\right]^2 \times \left[\frac{\sin\left(\frac{N\pi d \sin\theta}{\lambda}\right)}{\sin\left(\frac{\pi d \sin\theta}{\lambda}\right)}\right]^2$$

where N is the number of lines or slits. This equation is more often written as

$$E = E_0 \frac{\sin^2\beta}{\beta^2} \cdot \frac{\sin^2 N\gamma}{\sin^2\gamma}$$

For a single slit,

$$E = A^2 h^2 w^2 \frac{\sin^2\beta}{\beta^2}$$

where h = height of slits

A = amplitude.

Thus we see that for a grating

TABLE 8-1. TWO EQUAL SLITS

δ (phase difference)	Vector diagram	Irradiance	Designation
0^o	a a A	$E_o = A^2 = 4a^2$	1st Max.
180^o	$\delta = 180^o$	0	1st Min.
360^o	$\delta = 360^o$ A	E_o	2nd Max.
etc.	etc.		

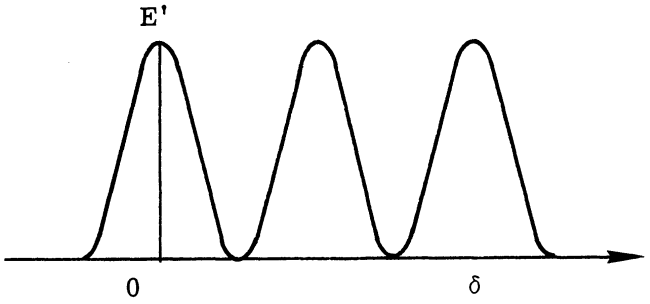
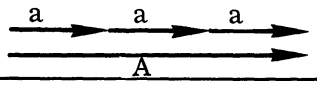
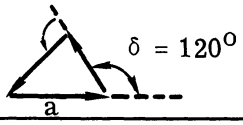
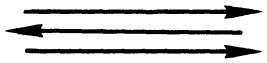
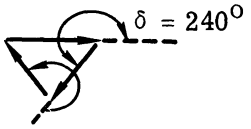
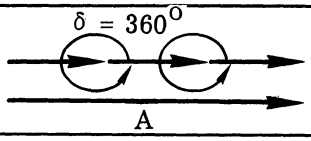


TABLE 8-2. THREE EQUAL SLITS

δ (phase difference)	Vector diagram	Irradiance	Designation
0^o		$E_0 = 9a^2$	1st Max.
120^o		0	1st Min.
180^o		$E = \frac{E_0}{9} = a^2$	1st Sub-Max.
240^o		E = 0	2nd Min.
360^o		$E = 9a^2 = E_0$	2nd Max.

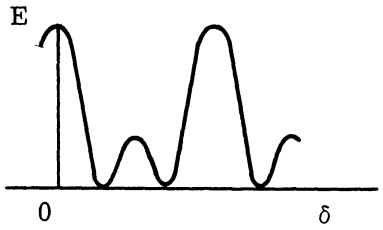


TABLE 8-3. SIX EQUAL SLITS

δ (phase difference)	Vector Diagram	Irradiance
0^o	a a a a a a A = 6a	$E_0 = 36a^2$
60^o	$\delta = 60^o$ $\delta = 60^o$	0
120^o	$\delta = 120^o$	0

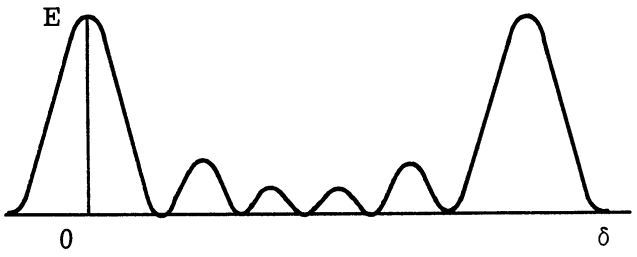
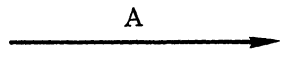
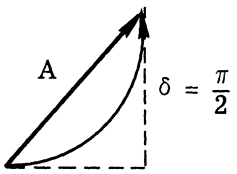


TABLE 8-4. SINGLE SLIT

δ (phase difference)	Vector diagram	Irradiance	Designation
0		$E = E_0 = A^2$	Max.
$\pi/2$			
π			
2π		E = 0	1st Min. Positive
-2π		E = 0	1st Min. Negative
4π		E = 0	2nd Min. Positive
-4π		E = 0	2nd Min. Negative

E = (constant) × (single slit diffraction function) × (N-slit interference function).

These considerations are for unblazed gratings. See the Handbook of Military Infrared Technology, section 7.12.1, page 310, for more details.

For a diffraction grating the angular dispersion is given (for ϕ constant) by:

$$\frac{d\theta}{d\lambda} = \frac{p}{d \cos \theta}$$

The resolving power can be calculated by noting that a grating has N slits and therefore N - 1 minima. Thus, from a principal maximum of order number p to the one or order (p + 1) we have

$$p;\ p + \frac{1}{N};\ p + \frac{2}{N};\ \text{----};\ p + \frac{(N-1)}{N};\ p + \frac{N}{N}$$

The Rayleigh criterion requires that the order p for λ_2 coincides with the order $\left(p + \frac{1}{N}\right)$ for λ_1. Thus we have

$$p(\lambda + \Delta\lambda) = \lambda\left(p + \frac{1}{N}\right)$$

or

$$\left(1 + \frac{\Delta\lambda}{\lambda}\right) = \left(1 + \frac{1}{pN}\right)$$

$$RP = pN = \frac{\lambda}{\Delta\lambda}.$$

8.3. TYPES OF PRISMS AND PRISM SPECTROMETERS

Several examples of specialized prisms and spectrometers are:

1. The Bunsen-Kirchhoff spectrometer (a scale is reflected off of a prism face into the telescope).
2. The constant deviation (Wadsworth) prism-mirror (see Sawyer, page 74 and H.M.I.R.T., page 308).
3. The Pellin-Broca prism (Sawyer, page 75).
4. The Amici prism (Ibid, page 72)
5. The Young-Thollon half-prisms (Ibid, page 71) and arrangements for mirrors (Ibid, page 82).
6. The Littrow-type mount (Ibid, page 117).

8.4. TYPES OF GRATINGS AND GRATING MOUNTS

From what has been said about gratings, it is obvious that they can be used either in transmission or reflection. A more interesting variation comes from their use in plane or concave reflecting form. The latter was treated most completely by Rowland, achieving a useful combination of focussing and grating action. He showed that the radius of curvature of the grating surface is the diameter of a circle (called the Rowland circle) which has the following property. Any source placed on the circle will be imaged on the circle, with dispersion, if the rulings are made so that d is constant on the secant to the grating-blank (spherical) surface. The astigmatism acts so that a point source on a Rowland circle is imaged as a vertical line perpendicular to the plane of the circle.

If the dispersion is sufficiently large, one may find overlapping of the lines from one order with members of the spectra belonging to a neighboring order. A simple construction allows one to determine the location of the orders. This construction is shown in figure 8-4. The sequence is as follows: 1. Draw a semi-circle with radius set equal to d. 2. Draw a line passing through the center of this semi-circle with the selected angle of incidence. This represents a ray in the zero-th order. 3. Drop a perpendicular from the intersection of this ray with the circle to the diameter of the semi-circle. This is the starting point. 4. From this point mark off the divisions along the diameter in units of λ/d. 5. Erect perpendiculars from these points to the circle. 6. Lines from the center to these intersections of the perpendiculars and the circle represent the rays of the higher orders of the grating.

Errors and imperfections in the ruling of gratings can produce spurious images which are called "ghosts." These are discussed briefly in Section 68, Chapter 7, of Sawyer. Also, the grooves in a grating can be shaped so as to send more radiation along a preferred direction corresponding to an order other than the zero-th. Such gratings are said to be blazed in that order.

Mountings of particular interest are:

1. Paschen-Runge (pg. 145, Sawyer) (see pgs. 311-12 Handbook of Military IR Technology)
2. Rowland (pg. 149, Sawyer)
3. Eagle (pg. 151, Sawyer)
4. Wadsworth (pg. 131, Sawyer)
5. Pfund (pgs. 83 & 280, Sawyer)
6. Ebert-Fastie
7. Littrow

One of the most popular grating spectrometer mounts today is the Ebert. It was described by Ebert more than 60 years ago, but was relatively unused until it was recently rediscovered

in the literature by Fastie. It features a single spherical collimating mirror and a grating placed symmetrically between the two slits as shown in figure 8-5. The big advantage of the Ebert system is the fact that it is self correcting for spherical aberration. At first glance it would appear that this advantage is outweighed by a serious astigmatism since a spherical mirror is being used off axis. If straight entrance and exit slits were used, this would be true. However, if curved slits are used, this defect is almost completely overcome. The Ebert spectrometer therefore utilizes slits having a radius of curvature approximately equal to one half the width of the grating. If viewed along the axis, the complete symmetry may be seen as shown in figure 8-6. The slits are aligned so that they lie on a circle having a curvature equal to that of the slits and concentric with the mirror. It should be pointed out that since the slits are on opposite sides of the grating, a complete correction for astigmatism is not accomplished. It is advantageous to have the slits on the same side of the optic axis. If the aperture of the Ebert system is kept relatively small, however, astigmatism is usually not the limiting optical defect.

The Littrow system shown in figure 8-7 has slits on the same side of the grating in accordance with minimizing astigmatism. An advantage of the Littrow mount, therefore, is that straight slits can be used. In fact, straight slits may be used even for a spherical collimating mirror M if the aperture is not too large. Its greatest disadvantage is that it does not correct for spherical aberration, not too serious a defect for long focal length-small aperture instruments. If an off-axis parabola is used to collimate the light, aberrations are greatly reduced.

Although the Ebert system has gained in popularity recently, there is still an open question as to whether it or the Littrow mount is optically superior. Investigations are being carried out in an attempt to either settle the question or to determine the conditions under which one is more favorable than the other.

9. REFERENCES SHOWING APPLICATIONS OF RADIOMETRY

A paper in the reprints by the American Society of Photogrammetry is entitled "Thermal Mapping of Yellowstone National Park," by J. H. McLerran, U. S. A. CRREL and J. O. Morgan, IST, University of Michigan. In it several infrared images are shown together with a map of relative temperature distributions. Another paper by McLerran on a joint U. S. A. CRREL and University of Michigan program of infrared sensing in the arctic shows the detection of crevasses by infrared imagery. The problem of obtaining absolute radiometric data from airborne infrared imaging devices is discussed by England and Morgan of the University of Michigan in yet another paper entitled "Quantitative Airborne Infrared Mapping." The technique uses a chopped radiometer which traces a path through the center of the infrared image in the direction of flight of the aircraft.

Balloon flights made with TIROS and NIMBUS radiation measuring instruments yielded data reported by Bartman et al. in "Infrared and Visible Radiation Measurements on High Altitude Balloon Flights at 34 KM Altitude." Since that time ESSA has reported many radiometric data from its satellite program. NASA-supported programs continue to provide radiometric data although most of remote sensing yield relative rather than absolute values.

The work by R. Colwell, the University of California, in remote sensing is particularly instructive in terms of experiments made from high platforms which are not airborne. A relatively simple radiometer, built with components such as a Newtonian telescope and a thermistor detector, has been used in our laboratory for direct radiometric data of forests and grasses.

The proceedings and the selected papers discussed briefly are available as follows:

Selected Papers	Price
American Society of Photogrammetry 105 Virginia Avenue Falls Church, Va. 22046	$2.00 for members of the society $3.00 for non members

Symposia Proceedings	Price
University of Michigan Publications Distribution 615 E. University Ann Arbor, Michigan 48106	
First Proceedings	$4.00
Second Proceedings	$4.00
Third Proceedings	$6.50
Fourth Proceedings	$6.50

(Orders accompanied by remittance will be sent postpaid).

10. AUXILLARY FIELD EQUIPMENT PRICE LIST

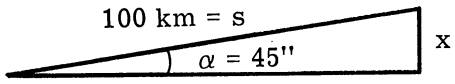
1. Stoll-Hardy HL4 Radiometer	$985.00
a. Battery set (with polyethylene insulating bags)	15.25
b. Leslie Cube, stainless steel with thermometer	145.00
2. Type 2300-B Pyrometer (scale in degrees centigrade)	175.00
a. Type 3378 carrying case	40.00
b. Type 2545 surface temperature thermocouple	35.00
c. Type 2045 air temperature thermocouple	35.00
d. Type 2410 needle thermocouple	35.00
e. Type 248 rigid immersion thermocouple	35.00
f. Type 4025 carrying case for thermocouples	7.00
3. Assmann psychrometer, No. 230	175.00
4. Hygro-thermograph, No. 255	220.00
a. Set of daily gears for No. 255 hygrothermograph	4.75
b. Spare hair element for No. 255 hygrothermograph	7.50
c. Two bottles purple ink	2.25
d. 100 No. 255 W charts	5.00
e. 200 No. 255 D charts	10.00
5. Hand anemometer, No. 444	75.00
6. Rain gauge, No. 503	45.00
a. Rain gauge support, No. 504	7.50
7. Instrument shelter with supports, No. 176	110.00
8. Brunton pocket transit, No. 3215	49.50
a. Ball and socket tripod head, No. 3216	9.00
b. Tubular extension tripod, No. 3217	24.00
c. Instrument case, No. 3218	2.10
9. Chain tape, No. 976-D, 100'	23.80
10. Abney hand level, No. L 646	35.50
11. Lightweight tripot, No. 572	80.00
12. Plane table drawing board, No. 573X	30.00
13. 14 oz. Plummet, No. 454	4.75
a. Plummet sheath, No. 463	3.00
14. Field book, No. 2731	1.50
15. Veriwide V100L cameras with Schneider, super-angular F/8 lens	225.80
a. Eveready carrying case	11.25
b. Leitz optical view finder	51.75

c. K-2 yellow filters for Veriwide V100L camera	4.63
d. Haze filters for Veriwide V100L camera	4.63
e. Sun shades that do not cut off for V100L cameras	3.00
f. Cable releases for V100L camera	1.65

11. RADIOMETRY PROBLEMS

1. Given an angular field of view (f.o.v.) of 45 arc secs, what linear distance is subtended at a range of 100 km?

ANS. 22 meters



$$x = 10^5 \text{ meter} \times \alpha \text{ (in radians)}$$

$$\frac{3'}{4} \times \frac{1^o}{60'} \times \frac{2\pi \text{ radians}}{360^o} = \frac{\pi}{40 \times 360}$$

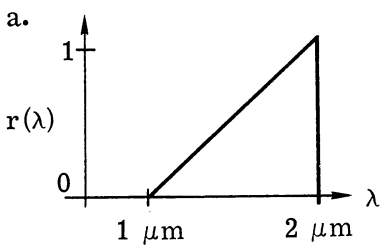
$$x = \frac{3.14}{4 \times 3.6} \times 10^2 = 21.8 = 22 \text{ m}$$

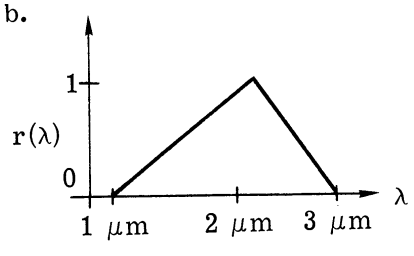
2. If a radiometer with an f.o.v. of 10 mr by 10 mr obtains a signal V from a point source with $I = 10^{-3}$ w ster^{-1} at a range of 10 km, what is the value of isotropic radiance which, on filling the field, will give the same signal?

ANS. 10^{-7} w m^{-2} ster^{-1} or 10^{-11} w cm^{-2} ster^{-1}

Sphere with radius 10 km has area $4\pi \times 10^2$ and covers 4π steradians. Therefore 10^2 km^2/ster = 10^8 m^2/ster.
10^{-2} r $\times$ 10^{-2} r = 10^{-4} ster = f.o.v.. Thus,
f.o.v. ≡ Area of 10^{-2} km^2 = 10^4 m^2.
$I = LA$; $L = 10^{-3} \times 10^{-4}$ w/m^2 ster.

3. The following r(λ) functions, if normalized to the peak, have what spectral bandwidth?

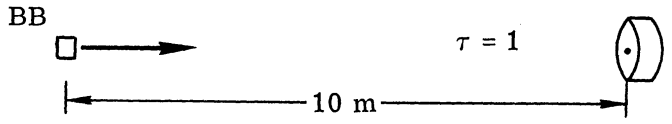




ANS. $\Delta\lambda = 0.5$ μm for a and $\Delta\lambda = 1.0$ μm for b.

4. Given an infinite plane Lambertian emitting surface. With a fixed f.o.v. radiometer, how does the output voltage vary as the viewing angle is changed? Const Why? Invariance of L or $\frac{\cos\theta}{\cos\theta}$.

5. Given a blackbody with an aperture of area 1 cm^2 at 2000°K, and a radiometer at 10 m distance as shown:



If the radiometer has a non-selective thermistor detector with a narrow spectral band-pass filter $\Delta\lambda = 0.1\ \mu m$ at $2.0\ \mu m$, and the output is:

10 volts = V With Source

0 volts = V No Source

Calculate $R_{0,E}$ in volt m^2 $watt^{-1}$

ANS. $R_{0,E} \cong 10^3$ volt m^2 $watt^{-1}$

from $$R_{0,E} = \frac{V}{\frac{A_c}{S^2}\int r_E(\lambda)\tau(\lambda)\epsilon(\lambda)L_\lambda(\lambda, T)\,d\lambda}$$

We note that $r_E(\lambda) = 1$ for $1.95\ \mu m \leqq \lambda \leqq 2.05\ \mu m$

$r_E(\lambda) = 0$ for all other wavelengths

Also, $\epsilon = 1$ for blackbodies.

From the GE slide rule we can find that a 2000°K blackbody radiates 40 watt cm^{-2} μ^{-1} at its peak of 1.45 μm. Since blackbodies are, by definition, Lambertian sources then

$$L(\lambda \text{ peak; } 2000^\circ K) = \frac{40}{\pi} \text{ watt cm}^{-2}\ \mu m^{-1}\ \text{ster}^{-1}.$$

Also from the slide rule we can find that

$$L_\lambda(2.0\ \mu; 2000^\circ K) = 0.8\ L_\lambda(\lambda \text{ peak; } 2000^\circ K).$$

We can assume that in the interval of $r_E = 1$ the spectral radiance is sufficiently constant that

$$\int L_\lambda\,d\lambda \cong L_\lambda \cdot \Delta\lambda$$

Thus

$$R_{0,E} = \frac{10v}{\frac{1\ cm^2}{(10\ m)^2}\left[1 \cdot 1 \cdot 1 \cdot \left(\frac{40}{\pi} \cdot 0.8\right) \cdot 0.1\ \mu m\right]},$$

or

$$R_{0,E} = \frac{10^3 \text{ v watt}^{-1} \text{ m}^2}{\left(4 \times \frac{0.8}{\pi}\right)} \cong 10^3 \text{ v w}^{-1} \text{ m}^2.$$

6. Change filters in problem 5 to a 0.1 μm spectral bandpass centered at 4.0 μm. The output is then:

   4 v = V With Source,

   1 v = V With No Source.

   Calculate $R_{0,E}$.

ANS. This is solved by using 4v - 1v = 3v for the output voltage and finding:

$$L(4.0\ \mu;\ 2000^{\circ}K) = 0.18\ L_{\lambda}(\lambda \text{ peak};\ 2000^{\circ}K)$$

$$R_{0,E} = \frac{3v \times 10^2}{\left(\frac{4.0 \times 0.18}{\pi}\right)} = \frac{3\pi}{0.72} \times 10^2 = 1.3 \times 10^3 \text{ v w}^{-1} \text{ m}^2$$

7. Now we use the radiometer of problem 5 with a $\Delta\lambda$ = 0.1 μm filter centered at 10 μ and perform a radiance calibration using two blackbodies (extended sources) which more than fill the field of view. These calibration measurements are:

   T = 310°K for V = 0.01v, and

   T = 350°K for V = 0.02v.

   Find: $R_{0,L}$.

   Answer: 130v/w cm^{-2} ster^{-1}

   We use $R_{0,L} = \dfrac{V}{\int r_L(\lambda)\tau(\lambda)\epsilon(\lambda)L_{\lambda}(\lambda,\ T)\,d\lambda}$

   but note that:

   $\Delta V = 0.02 - 0.01 = 0.01v$ so that

   $$R_{0,L} = \frac{1 \times 10^{-2}}{\Delta L_{\lambda,BB}(10\ \mu;\ 350^{\circ}K - 310^{\circ}K) \times 0.01}.$$

   From the GE slide rule when T is

   $$310^{\circ}K,\ L_{\lambda} = \frac{3.7 \times 10^{-3}}{\pi} \times 0.99,$$

   when T is

$$350^{\circ}K,\ L_{\lambda} = \frac{6.7 \times 10^{-3}}{\pi} \times 0.91.$$

Thus

$$\Delta L_{\lambda} = \frac{(6.7 \times 0.91 - 3.7 \times 0.99)}{\pi} \times 10^{-3} \text{ and therefore}$$

$$R_{0,L} = \frac{10^{-2}}{\frac{2.4}{\pi} \times 10^{-3} \times 10^{-1}} 130 \text{ Volt } w^{-1} cm^{2} \text{ sr.}$$

8. Given that the two yellow lines of the sodium spectrum are just resolved by a crown glass prism of base = 2 cm, what is the RP of a spectrograph with 3 prisms of this type used in series (in the yellow region) where the prisms transmit at the minimum angle of deviation.

   Use the data,

   $\overline{\lambda} \cong 600$ nm, $\Delta\lambda \cong 0.6$ nm,

   B = 2 cm, n = 1.50.

   Answer: RP = 3000.

   $\frac{\lambda}{\Delta\lambda} = RP = \frac{600}{0.6} = 1000$ for one prism,

   3 prisms in series disperse the spectrum three times as much as one.

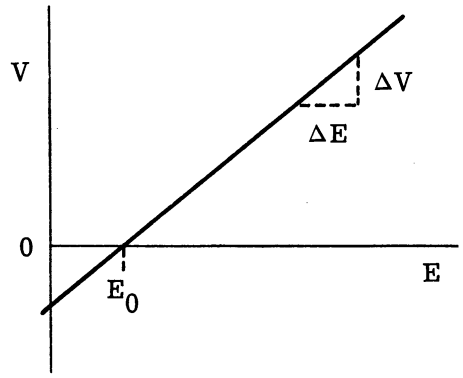


FIGURE 3-1. RADIOMETER RESPONSE TO INPUT IRRADIANCE

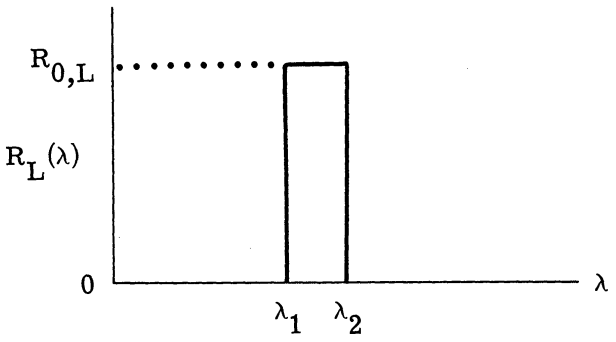


FIGURE 4-1a. IDEAL RADIOMETER SPECTRAL RESPONSIVITY

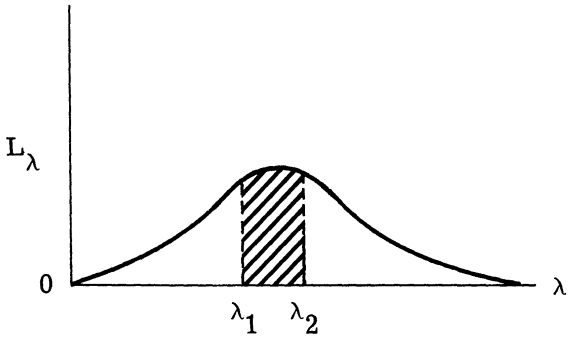


FIGURE 4-1b. A SMOOTHLY VARYING SPECTRAL RADIANCE

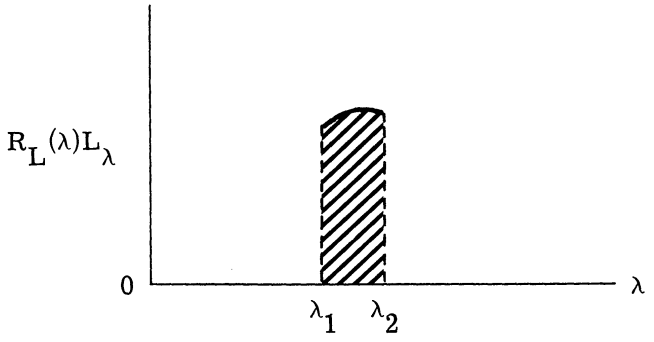


FIGURE 4-1c. THE PRODUCT OF SPECTRAL RADIANCE AND SPECTRAL RESPONSIVITY

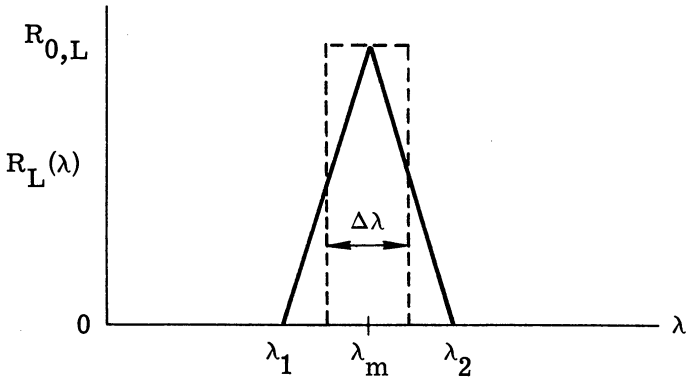


FIGURE 4-2a. TRIANGULAR SPECTRAL RESPONSIVITY

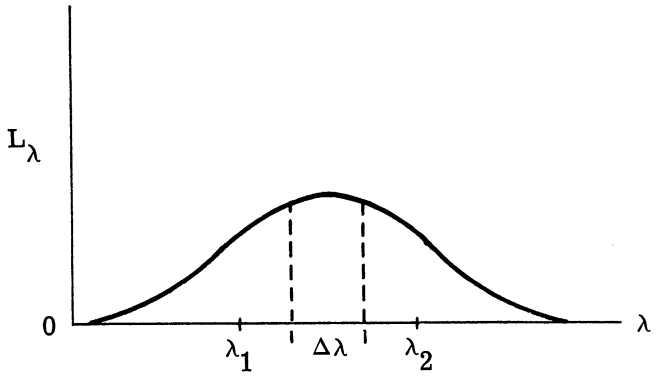


FIGURE 4-2b. SMOOTHLY VARYING SPECTRAL RADIANCE

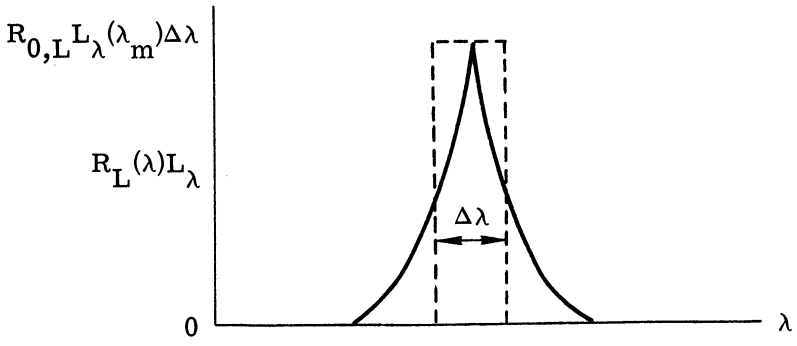


FIGURE 4-2c. PRODUCT OF TRIANGULAR SPECTRAL RESPONSIVITY AND SPECTRAL RADIANCE

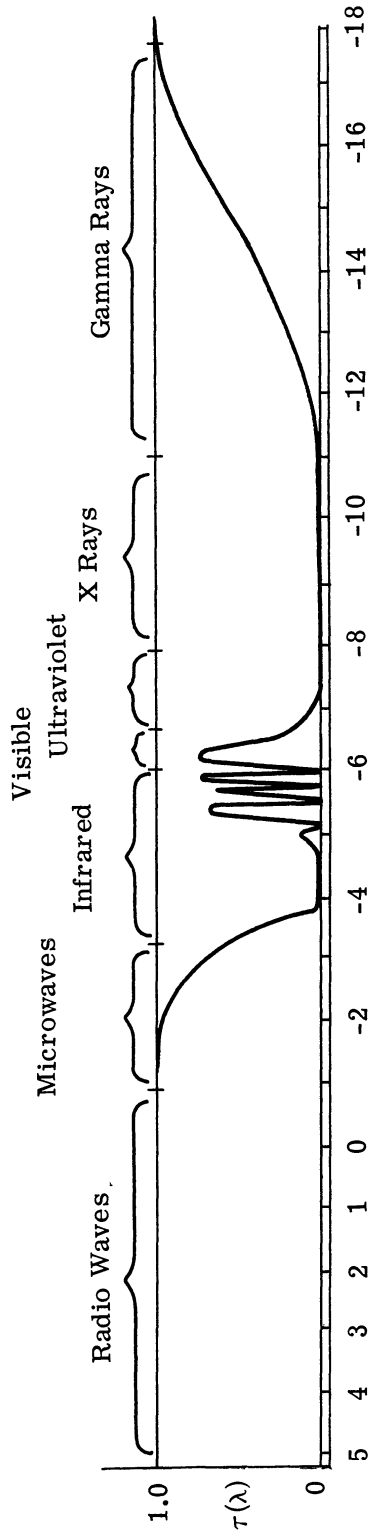


FIGURE 7-1. PROPERTIES OF ATMOSPHERIC TRANSMISSION (WAVELENGTH IN METERS). Transmission is shown schematically and does not represent actual values.

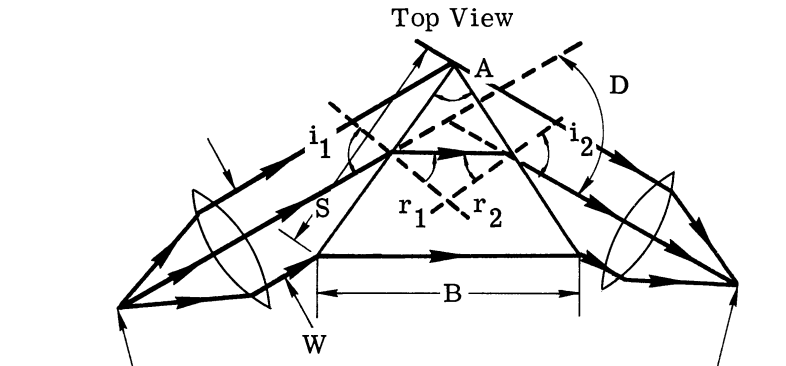


FIGURE 8-1. TOP VIEW

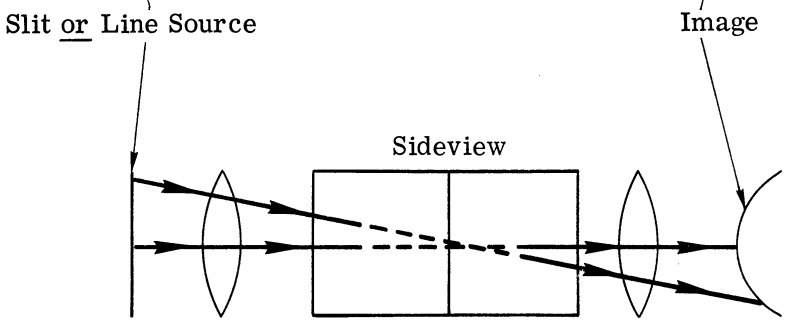


FIGURE 8-2. SIDE VIEW

ELEMENTARY PRISM SPECTROMETER SCHEMATIC

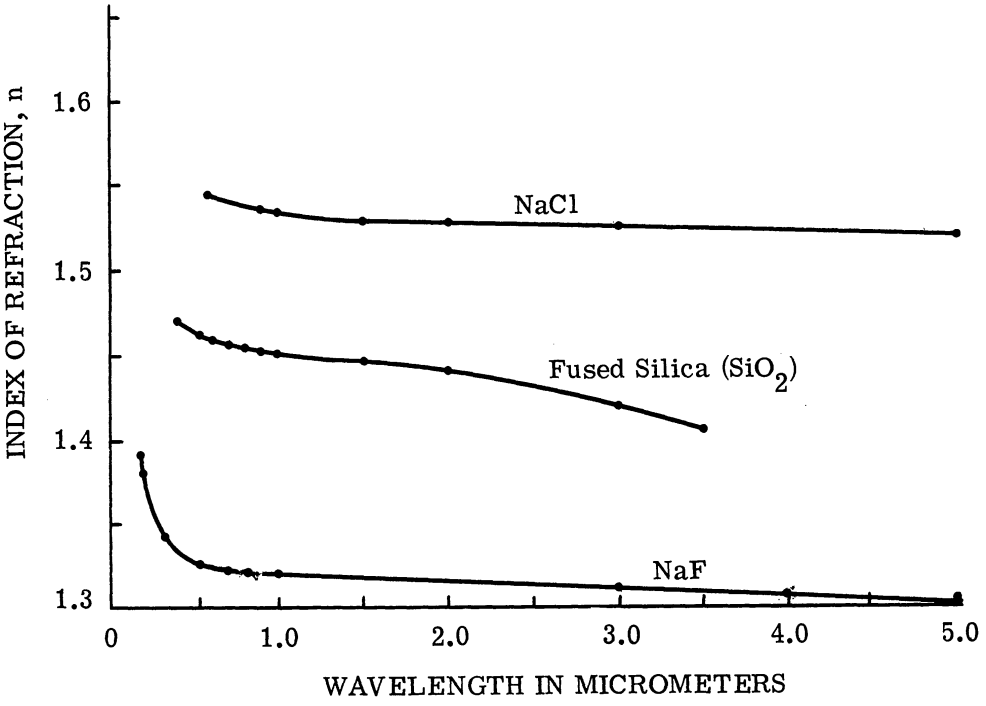


FIGURE 8-3. INDEX OF REFRACTION AS A FUNCTION OF WAVELENGTH

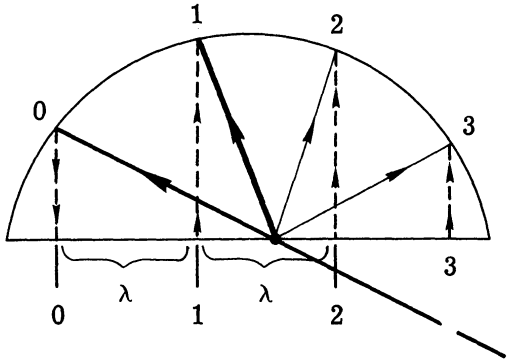


FIGURE 8-4.

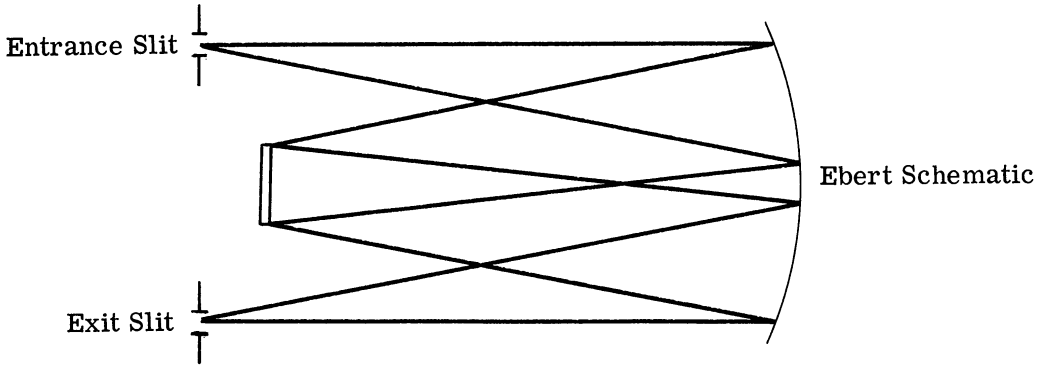


FIGURE 8-5. EBERT SCHEMATIC

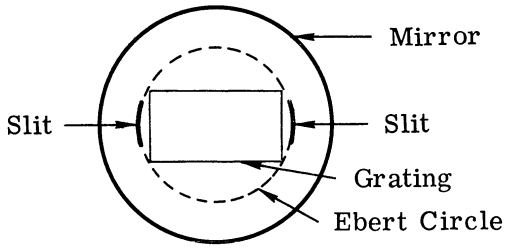


FIGURE 8-6. CURVED SLIT ARRANGEMENT IN EBERT SYSTEM

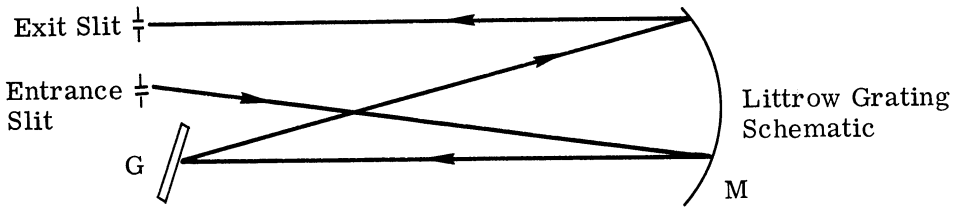


FIGURE 8-7. LITTROW GRATING SCHEMATIC

SECTION III
PHOTOGRAPHY

1. INTRODUCTION

The primary purpose of photography is to replicate sufficiently well in the form of a permanent record the luminous radiance distribution which the human eye would have seen if the eye were located at the camera position at the time of exposure. The development of the photographic process which involved replication of visible light images naturally occurred using visible light illumination units. However, a multitude of secondary applications has been developed where the purpose is to record radiance distributions which are not normally perceived in human vision and which transcend the limitations of vision in a number of ways. In some important instances, such as photography in the ultraviolet and infrared spectral ranges (which have important remote sensing applications), the use of illumination units makes little sense. Because of the previous use of such units over many years, the illumination units were forced into service where they do not apply directly, so that confusion in reading photographic literature is made very likely.

The importance of photography to remote sensing cannot be overestimated. Not only are direct photographic recordings of remotely located objects frequently employed but also photographic recordings are made where the first transducing action is performed by some other sensor system with photography making the output data available for analysis by human vision.

The principal advantage of the photographic system is its capability of rapidly sorting and recording received radiant power in accordance with the direction of propagation of the component waves. Such accuracy and speed for this purpose is unparalleled by any other sensor system and it can be accomplished at a comparatively very low cost. The capability for directional accuracy is used in the science of photogrammetry and cartography. The replication of visual radiance distributions is used in photointerpretation where most of the inferential relations between what is seen and what is in the field of view are derived from the great store of previously learned inferences using normal human vision. To be sure, the aerial photograph produces views from unusual vertical aspects and long ranges that are not provided by normal human experience, but, like an Indian tracker who is trained to perceive the meaning of particular details, the photo interpreter must also be trained to make inferences by using the available details which that aspect and range may afford.

The amateur photographer may be very well pleased in spite of the lack of precision in the use of his photographic system. His need for replication may be satisfied if he is able to recognize good old uncle Joe or to recall a scene on his vacation trip. The person who is seriously interested in photography as one of his set of radiometers must require all that his

photographic radiometer can consistently deliver. In order to utilize the photographic system to its full value and to apply it to new remote sensing problems, an understanding of the limitations and properties of the photographic process is essential. The comparatively low cost and availability of photographic systems is a compelling reason for the use of such a system in remote sensing tasks, either as a primary sensor or as a secondary auxiliary sensor.

In the text which follows the essential features of the photographic process, the photographic sensor system, and photogrammetric and photo-interpretation applications are summarized.

2. CHARACTERISTICS OF PHOTOGRAPHIC MATERIALS

2.1. FILM BASE MATERIALS

Photographic film consists of (1) a thin layer of material, usually in roll form, which includes a base material, (2) one or more emulsions, and (3) a subcoat which causes the emulsion to adhere to the base. In addition to the photosensitive emulsion, other coatings may be applied to the surface for such purposes as minimizing abrasion or friction, preventing the buildup of static, or preventing halation effects.

The film base may be made of one of several materials, depending on the desired characteristics. Since 1950, triacetate or high acetyl base film has found major use. More recently, polyesters, such as polyethylene terephthlate and polycarbonate, have been used as film base. They are more expensive but have certain desirable characteristics.

Both types of base qualify as safety film, since neither supports combustion. Cellulose triacetate has relatively low long-term shrinkage. It meets the stability requirements of ASA standard PH 1.28-1957 for use in permanent records such as microfilm. However, the dimensional stability of the polyester base with respect to changes in temperature and humidity exceeds that of the triacetate base. Polyester film has greater tear strength than triacetate, but has the disadvantage that it cannot be spliced with commonly used cements. Instead, a pressure-sensitive adhesive tape, ultrasonic splicer, or induction heating can be used for splicing.

Films are designed to remain flat at a relative humidity of about 50%. They will curl in one direction or another as the humidity departs from this value and will again flatten out as the humidity returns to the value. Curl amplitude is lower with polyester base films.

Film is subject both to reversible temporary shrinkage and nonreversible long-term shrinkage. Temporary shrinkage is caused by changes in temperature or humidity. Between 20 and 70% relative humidity at a temperature of $70^{\circ}F$, triacetate base film will undergo a change of 0.005 to 0.009% for each 1% change in relative humidity. A temperature change of $1^{\circ}F$ will cause about half as much change. Shrinkage of polyester films is about half the amount mentioned above.

Long-term shrinkage of cellulose triacetate base may reach 0.3 to 0.4%, of cellulose acetate butyrate base about 0.15%, and of polyester base less than 0.1%.

Polyester film can have a thinner base at the same strength, so that more film length can be obtained for a given volume or weight. Also, polyester film can be exposed in a vacuum, as in the case of electron beam recording or in outer space.

Plate glass is also used as a film base to provide the highest degree of dimensional stability. Such a dimensional stability is needed only for unusual photographic mapping accuracies and some holography or lensless photography applications.

2.2. THE PHOTOGRAPHIC EMULSION

The photographic emulsion is a specially prepared gelatin containing the photosensitive materials in suspension. The emulsion is carefully spread onto the film base in liquid form so that a very thin layer hardens and adheres to the film base. The emulsion is porous in microstructure so that chemical solutions can diffuse in and out through the structure. The photosensitive crystalites of silver halide are dispersed and entrapped by the gelatin structure. These crystalites of silver chloride, bromide, and iodide are microscopic in size. The selection of sizes of crystalites to be used has important consequences in the film properties to be discussed later. Each of these crystalites is to function as an individual quantum detector, amplifier and recorder. Most emulsion now contain certain selected dyes which also take part in the quantum detection process in conjunction with the silver halides to extend the response of the silver halide to longer wavelengths.

2.3. THE PHOTOCHEMICAL PROCESS

The primary photochemical process occurring during exposure is the absorption of photons by a crystalite. If the photon energy is equal to or greater than the energy band gap of the silver halide, a mobile electron is set loose to move about in the crystalite. The electron may become stuck temporarily at some defect position in the crystal, possibly near the crystalite surface where defects are likely to be found. Within the crystalite some of the silver ions will be out of the proper lattice positions. These are called interstitial ions and they result from imperfect crystal formation as a natural event. An important property of silver halides is that these interestitial ions can move about between the lattice ions under the influence of ordinary thermal agitation of the lattice. The excess negative charge of the trapped electron and the excess positive charge of the silver ion attract so that the silver ion tends to move toward the electron. The eventual proximity of electron to interstitial silver ion causes the electron to orbit the silver ion to make it a single atom of neutral silver. If a few more electrons released by photo absorption repeat the process in that same locality, then a small group of neutral silver atoms is formed. This small group of silver atoms is called the latent image since no visible effect can be seen even under a high-powered microscope. The function of this small group of atoms is to catalyze the precipitation of all silver atoms in the crystalite when a developing solution is present. Roughly speaking, there are about 10^{12} silver atoms in a 1-μm cube of silver halide. If only 100 latent image silver atoms are needed to "seed" the precipitation of all silver in the crystalite, then we could argue that the crystalite and developer combination provides a gain of about 10^{10} over what a straight unseeded photochemical reaction

would provide. A fixing solution is applied to the emulsion after the precipitation process in order to remove the unprecipitated silver halide to avoid further photochemical action.

2.4. BASIC FILM PROPERTIES

One can predict several important consequences from the foregoing discussion. With a constant irradiance a large crystalite will capture more total energy in a given exposure time than will a small crystalite, simply because of its greater size. In addition, the same size precipitation nucleus in a large crystalite will cause more atoms of silver to precipitate because there is more silver per grain, so the gain will be higher. Hence, less power per unit area will cause larger silver precipitation in an emulsion containing large crystalites, so the large-grain film will tend to be more sensitive or "faster."

Exposure at unusually low temperatures, such as occur in high altitude and arctic environments, will slow the thermal diffusion rates of the latent-image forming interstitial ions so that the temporarily stuck electron may be gone before the ion can join the group of others at that location. Thus, more irradiance may be required to achieve the same number of precipitation centers. The film is less sensitive or "slower" under such low-temperature exposure conditions.

If the crystalite sizes are selected by some means to be very nearly the same size, then the threshold exposure to cause precipitation will also be very nearly the same. Almost all crystalites precipitate or very few precipitate, depending upon only a small change in irradiance value. The irradiance pattern will be recorded as either black or white with very little between. Such an emulsion would be called a high-contrast emulsion.

The large crystalites precipitate correspondingly large clumps or grains of silver, so that a sharp change in irradiance along the film will be represented by a line of larger silver grains at the edge of the high irradiance region presenting a jagged edge to a region having a few isolated large grains in the low irradiance region. Thus, sharpness of detail in the irradiance pattern will not be well recorded with large-grain emulsions.

The number of "grey levels" or distinguishable transmittance values achievable for a very small area of developed emulsion will also be determined by grain size. Large grains provide a coarser grey scale than small grains for the same small area.

The precipitation process induced by the developer will occur, given enough time, even without catalytic latent images. For that reason the chemical process requires very careful

control in order to precipitate only the catalyzed crystalites. Chemical process rates are sensitive to reaction temperature. The development process is more so. It is not unusual for professional photographers to hold their chemical bath temperatures constant to within 0.25 to 0.10^{o}F to achieve consistent results. In addition, reaction rates depend upon the concentration of reagents so that a developing solution changes its characteristics as more film is developed by it or as the developer reacts with dissolved oxygen from the air. Since the chemical by-products generated by the precipitation diffuse out at the emulsion surface, they must be carried away so that the concentration of active reagents can be maintained at the emulsion surface.

Thus, the manner and frequency of developer bath stirring or agitation will change the development rate. Frequent stirring should increase the development rate by maintaining a higher concentration of active reagent at the emulsion surface.

2.4.1. QUANTITATIVE PHOTOGRAPHIC FILM PROPERTIES. There are many good explanations in extant literature of quantitative photographic film properties. It would not be profitable simply to restate them here. What is needed is a translation of the language used in this literature to the common language of remote sensing and an amplification of those aspects which pertain particularly to the use of photographic materials for remote radiometry. Some redundancy with extant literature is required, however, in order to maintain continuity of thought. This discussion is presented with the assumption that the reader has the following references:

1. Kodak Data for Aerial Photography, 1967, First Edition, Dept. 454, Eastman Kodak Co., Rochester, New York, 14650.
2. Infrared and Ultraviolet Photography, 1955, Fifth Edition, Eastman Kodak Co., Rochester, New York, 14650.

2.4.2. SENSITOMETRIC CHARACTERISTIC CURVE. The product, $E_v t$, of luminous irradiance times the time of exposure is defined as the exposure, q_v. The units of q_v consistent with those presented in the discussion of basic principles are lumen seconds per square meter. In figure 2-1 is shown a typical characteristic curve showing the blackening effect of q_v after development. The transmittance, τ_v, of the film will change for different values of q_v. The density, D, of the film is defined as $\log_{10} \frac{1}{\tau_v}$. Note first that the units for exposure are given in photographic literature as meter-candle-seconds. These units are not appropriate for unit analysis. However,

1 meter-candle-second = 1 lumen sec/m^2

so that the size of the quantity is the same in either unit system. The curve is plotted linearly in D and linearly in $\log_{10} q_v$. Note that the value of the $\log_{10} q_v$ depends upon the unit system. The line over the numbers on the $\log_{10} q_v$ scale represents a negative value. The value of gamma for this example curve can be read from the curve. For $\log_{10} q_v = -2.0$ a density, D, should be about 0.40; for $\log_{10} q_v = -1.0$ the density should be about 1.40. Thus $\gamma = \frac{\Delta D}{\Delta \log_{10} q_v}$ so that

$$\gamma = \frac{1.40 - 0.40}{-1.0 - (-2.0)} = 1.0$$

A larger value of γ results in a larger density difference for a given exposure change. The value γ can be varied for a given film type by a change in the time and conditions of development. Knowledge of the characteristic curve allows one to infer the value of q_v from a measure of density. In turn, since $q_v = E_v$ (at the film) $\times$ t, and E_v and t are related through camera properties to L_v(scene), an inference can be made about the magnitude of L_v(scene).

The characteristic curves published by the film manufacturers represent average values for the film type and are not intended to be sufficiently accurate to use directly for quantitative radiometry. The curves are suitable for an exposure guide. More accurate work requires the determination of the characteristic curve for the particular sample of film being exposed.

2.4.3. PHOTOGRAPHIC FILM SPEEDS. The characteristic curve is the fundamental data concerning film exposure characteristics. Since most film users do not refer to the characteristic curve, a single number is derived from the curve so that such a number can be used with an exposure meter to assist in adjusting the exposure appropriately assuming a fixed development procedure. Films used for popular photography are given an ASA number derived by subjective evaluation of pictorial rendition. The exposure index number is chosen so that the most pleasing pictorial rendition results.

The relative sensitivity value is defined as $\frac{1}{q_v(\text{tungsten lamp}) \text{ meter-candle-seconds}}$ with exposure time set at one second to produce $\Delta D = 0.6$ above gross fog level. These values are the S_A values. Referring again to the characteristic curve (figure 2-1) we can determine the S_A value of that hypothetical film provided the curve was measured using tungsten lamps. Gross fog level is clearly D = 0.2 since less exposure no longer reduces density. To achieve $\Delta D = 0.6$ we must read the log q_v value for D = 0.8. From the graph, log $q_v = -1.60$. Thus, $q_v = 10^{-1.60}$. In order to evaluate the quantity $10^{+1.60}$ one must use a $\log_{10}$ table. We must find that value in the table which has a $\log_{10}$ value of 1.60. The resulting number is 40. Therefore, the S_A speed is 40.

The aerial exposure index is defined as the reciprocal of twice the value of q_v (in meter-candle-seconds) at the point on the toe of the characteristic curve for which the slope is 0.6γ. We can find the aerial exposure index of the hypothetical film shown in figure 2-1 as follows: The value of γ was previously found to be 1.0. Thus

$0.6\gamma = 0.60$

A slope of 0.60 can be made on tracing paper by making a right triangle with one leg length along the D direction 0.60 and the other leg length along the $\log_{10} q_v$ axis of 1.0. The hypotenuse completing the triangle makes an angle whose tangent is 0.6γ with the $\log_{10} q_v$ axis. By slipping the $\log_{10} q_v$ side of this triangle along the $\log_{10} q_v$ axis, with apex to the left, one finds the point on the characteristic curve which has that new slope. Reading the value of $\log_{10} q_v$ at the point of tangency, one finds $\log_{10} q_v = -2.20$. Thus $q_v = (1/160)$ meter-candle-seconds; hence $\frac{1}{2q_v} = 80$ which is the aerial exposure index. Exposure meters make use of ASA, S_A, or aerial exposure indices to provide an indication of the appropriate exposure value to use.

2.4.4. INFRARED AND ULTRAVIOLET EXPOSURES. Characteristic curves are also published for infrared sensitive film. Since this film is sensitive to radiation which is not visible, some qualifications apply to these curves. Characteristic curves are shown for Infrared Aerographic Film, Type 5424 in reference 1, page 56. First note that these curves specify the type of illumination as daylight passing through a #25(A) filter. Thus, most of the irradiance on the film is longer in wavelength than the visible. The term daylight refers to an irradiance having the spectral distribution of a 6000°K blackbody source. The irradiance upon the film is that portion of a 6000°K blackbody irradiance which can pass through the #25(A) filter. The infrared irradiance after filtering will be proportional to the luminous irradiance before filtering for such a fixed spectral distribution. It is the luminous irradiance of the 6000°K blackbody before filtering which is used to obtain $\log_{10} q_v$ for the exposure axis in infrared film. The units of q_v are meter-candle-seconds. The infrared exposure q_e(infrared) watt-seconds/m^2 which actually activates the film will be proportional to q_v for a 6000°K blackbody spectral distribution. The aerial film will be used to photograph a natural scene of materials that exhibit spectrally dependent reflectances so that E_e(at the camera) will not be that of a 6000°K blackbody. Therefore, the use of E_v measured by an ordinary exposure meter will not always provide a good guide to the value of q_e (infrared). A silicon solar cell filtered by the #25(A) filter would produce a short circuit current more nearly in proportion to E_e (infrared) and hence be a better guide to exposure, q_e (infrared). Characteristic curves are not usually published for ultraviolet photography. Recommended exposure times and light sources are offered for starting trials.

2.4.5. SPECTRAL SENSITIVITY CURVES. The spectral sensitivity of a film for monochromatic light is defined as the inverse of the exposure, $q_e(\lambda)$, to produce a specified density above gross fog density. The units used in photographic literature are ergs/cm^2 rather than watt seconds/m^2 or joules/m^2. Since there are 10^7 ergs per joule and 10^4 cm^2 per m^2, no problem in unit conversion is likely. In reference 1 page 24, the sensitivity curves for monochromatic $q_e(\lambda)$ are shown for several film types. The term, log sensitivity, means $\log_{10}\left[\frac{1}{q_e(\lambda)\ \frac{\text{ergs}}{\text{cm}^2}}\right]$. The wavelength of the center of the narrow spectral band is in millimicrons or nanometers (10^{-9} meters). Notice that none of the films shown are equally sensitive for all wavelengths in the visible range. Notice also that the separation of the two sensitivity curves for D = 0.3 and D = 1.0 for plus-x aerographic film, type 2401 is not equal. They are closer together at 670 mμ than at 400 mμ. One may conclude that the γ of the film is wavelength dependent. The implications of this fact are that exposures $q_e(\lambda_1)$ at wavelength λ_1, and $q_e(\lambda_2)$, at wavelength λ_2, do not accurately produce a density, D, corresponding to $q_e(\lambda_1) + q_e(\lambda_2)$. It is for this reason among others that the photographic process makes a relatively low-accuracy quantitative radiometric system. In addition to the variation in γ, there is some evidence of spectral "cross-talk" or an interference of one spectral component with another, so that an exposure at one wavelength changes the spectral sensitivity value for another wavelength. Under most circumstances these effects may not be serious provided that high radiometric accuracy is not required.

2.4.6. VARIATIONS WITH EXPOSURE TIME. The reciprocity law for photographic exposure states that the density produced by an exposure q_v is independent of the irradiance value E_v and the time of exposure, t, provided only that $E_v t = q_v$. It is found, however, that the density does not depend solely upon the exposure value, q_v, but also depends upon the time of delivery of that energy per unit area. In figure 2-2 is shown a plot of $\log_{10} E_v$ against $\log_{10} q_v$, for constant density. If the reciprocity law held, then the density would depend only on the value of q_v so that a straight horizontal line would result. Note that the line is nearly straight and horizontal only in the region where $\log_{10} E_v$ is between -1.0 and +1.0 corresponding to exposure times somewhere between 1/2 second and 1/500 second. Other exposures times require more energy per unit area to get the same density. Reciprocity law failure is most important for astronomical photography through a telescope with which faint galaxies are photographed by exposures lasting for hours during the night while tracking the starlit sky. For remote sensing photography during the daytime, the reciprocity law holds well enough for practical purposes.

Other time-dependent and nonlinear effects are described in the photographic literature, but these are not usually of concern to most remote sensing problems.

2.4.7. PHOTOGRAPHIC PRINTS. When photographic prints are made, the reflectance of the surface in dark areas reaches a minimum when the area is completely covered by a silver deposit. Under these conditions, a small amount of light is still reflected, so that there is an upper limit to the obtainable density as measured by reflection. The characteristic curves of printing papers have a much more limited range of linearity between log exposure and density than does film.

The major problem in obtaining satisfactory prints from negatives stems from this characteristic of print paper. If the range of densities on the negative transparency exceeds the density range capacity of the print paper, then it is clear that the film density range will not be replicated in the print. The frequent result is to find that those objects producing a high film density are all rendered in the print as white areas without detail. Those objects producing a low film density are rendered as equally black in the print. Only those objects producing intermediate film densities are rendered in the useful grey scale range of the print. The print is used because it can be handled like any other sheet of data and read without auxiliary apparatus. The transparency requires backlighting. The use of print paper has little other virtue than the convenience of use. Whenever print paper is used, it is clear that information contained in the grey scale of the print image is less than that which could be obtained in a transparency.

2.4.8. THE LOG E PROCESS. In the printing of a negative transparency to make a positive print, the backlighted transparency forms the source of light to expose the print. If the backlighting is performed on a time serial point by point basis, as can be done by a raster scan on a cathode ray tube in back of the transparency, it is possible to alter the backlighting point by point in accordance with the density of the transparency at each point. If the density range of the transparency is too great, it is possible to reduce the print exposure range through control of the backlighting. When the backlighting spot is behind a transparency region of high density, the spot brightness can be automatically increased in proportion so that details on the print will not be rendered equally white. When the density of the film is low in front of the spot, the spot brightness can be automatically reduced so that details will not be rendered equally black. This process is the Log E process. It enables the successful printing of negative transparencies within the useful range of the print paper.

2.4.9. IMAGE DEFINITION. The resolving power of a film is a measure of its ability to record fine details of an image. It is frequently expressed in terms of a pattern of alternating dark and light lines of equal width and good contrast recorded on the photographic surface, the resolution being expressed as the maximum number of dark lines per unit length which can just be distinguished. The resolving power of a given film or paper depends, among other factors, on the characteristics of the emulsion and the size of individual grains of photosensitive material contained in the emulsion. Granularity of the recorded image is usually coarser for fast films than for slow, and will, in general, be much coarser than indicated by the grain size. The nature of the developer and the development time also affect the graininess of the image. Fast negative emulsions normally have a resolving power of 40 or 50 lines per millimeter; positive emulsions may resolve 80 to 100 lines; and very slow materials may be good for 150 to 200 lines per millimeter.

The definition obtainable in a photograph is improved by the use of film with a good modulation transfer function, low granularity, and good contrast. Factors other than the photographic material also affect the resolution. Reduced contrast of the target reduces attainable resolution. In a complete camera system, resolution may also be limited by the optical system of the camera. These factors affecting image definition are further discussed in following sections.

2.4.9.1. Measures of Image Definition. If a small spot of light is focused on a photographic emulsion, the resulting image will not be limited to the original area of the image. The image will appear to have been spread out to some extent. Part of the reason for the spread is the turbidity of the emulsion but other factors, such as chemical processing effects, may also contribute. The resulting density distribution may be viewed as due to the response of a nondiffusing emulsion to an equivalent irradiance with spreading in the original image. The equivalent irradiance will be a function of the distance away from the original spot. The ratio of the equivalent irradiance to the actual irradiance of the spot is called the spread function.

Figure 2-3 shows a hypothetical spread function. Two spots falling on the emulsion close together will produce an image which is fused together because of the spread function so that the two spots may not be resolved. An emulsion which exhibits a wide spread function will not be capable of resolving details as well as one which has a narrow spread function.

Figure 2-4 shows how the spread function effects the images of two stars photographed by an emulsion. The part of the image between the two stars is close to the peak value so that the contrast is reduced. In a similar way the details in an extended image lose contrast due to the proximity of other parts of the image.

Another and related measure of image definition is the modulation transfer function.

If one were to photograph a target whose radiance varied in a sinusoidal fashion, then the irradiance across the film would vary similarly. Suppose the radiance of the target were described by

$$L(x) = L(max)\left(\frac{1 - M \sin kx}{1 + M}\right)$$

where L(x) is the radiance at position x

L(max) is the maximum radiance

k is $2\pi/\Lambda$ the "spatial frequency."

Λ is the spacing between radiance maxima

M is a modulation factor establishing the contrast between maxima and minima.

The photographic film will respond to the illumination as if E_{eq} were present so that

$$E_{eq}(x) = E_{eq}(max)\left(\frac{1 - M' \sin kx}{1 + M'}\right)$$

It will be found that for targets with a large value of k that $M' \neq M$. The value of M' can be computed by measurements on the film as

$$M' = \frac{E_{eq}(max) - E_{eq}(min)}{E_{eq}(max) + E_{eq}(min)}$$

The ratio, M'/M, as a function of the k value of the target (assuming a scale of one to one in imaging) is called the modulation transfer function, MTF. Figure 2-5 shows how a target with continuously increasing k value with position across the target would be imaged by a photographic emulsion. The measurement of M' can be made from such a result. The resulting MTF might appear as shown in figure 2-6. It is clear that any imaging device or system may have an MTF defined for it by following the same procedure as for photographic film. For instance a camera lens can be tested using a microphotometer in the image plane instead of a photographic film. Hence a lens MTF can be established. The great value of the MTF as a measure of image definition is that the MTF of a complete imaging system can be computed from the MTF of each component used in series - i.e.,

MTF (system) = MTF (component 1) × MTF (component 2).

The reduction in MTF is illustrated in figure 2-7. Component 1 shows a high image definition property at the threshold contrast level. Component 2 is somewhat poorer than component 1. When these two components are used together, the definition is poorer than either of the

components separately. The position of the threshold contrast level will depend upon the target modulation factor. One may expect the human eye to have a threshold contrast of 0.02 in an image. Thus, when M' = 0.02 the threshold contrast is reached. Thus MTF (threshold) = $\frac{0.02}{M}$. For low-contrast targets (M = 0.1) the MTF (threshold) will be 0.2 and the image definition will be significantly poorer than for the poor component 2 alone. The more components or processes through which an image is taken, the more degradation can be expected. In particular, if 10 components which had MTF's identical to that of component 1 were used in order to image a target with 0.1 contrast, the largest value of k which could be seen would be about 0.19 times the largest value of k for component 1 alone.

The shape and scale of the MTF for emulsions are dependent to some extent on factors other than the physical and chemical characteristics of the film itself and the test procedures used in determining the MTF. The wavelength of the light will have some effect on the MTF. For Plus-X film, as an example, the spatial frequency producing 50% response ranges from about 20 cycles per mm or $2\pi \times 20$ radians per mm for red light to 40 cycles per mm or $2\pi \times 40$ radians per mm for blue light. Another influence is the adjacency effect, which causes the edge of a heavily exposed area of a negative to develop greater density than points farther away from the edge. The adjacency effect influences the shape of the MTF. However, the magnitude of this effect will be reduced to some extent during the developing process by strong agitation and full development. It is also reduced by low contrast at an edge. MTF may also vary with exposure level, decreasing with increasing exposure for thick emulsions and increasing with increasing exposure for thin emulsions.

2.4.9.2. Granularity. Granularity is one of the contributing factors in limiting image definition. Granularity of a film is usually measured by recording the variation in density of the film as an aperture of specified dimensions is moved over an area of the film which has been subjected to uniform exposure. If the dimensions of the aperture are moderately larger than the size of individual grains, the effect will be to cause a random variation of density which has a normal distribution about a mean value. The root mean square value of this deviation can be taken as a measure of the granularity of the film. The standard deviation tends to be inversely proportional to the square root of the aperture area, so this area must be specified to provide meaningful data.

Typical rms granularity values can be quoted for film exposed to obtain a density of 1.0 above base plus fog, using a circular aperture of 24-μm diameter and processed for 8 minutes in D-19 developer. The values range from 0.016 for SO-243 film to 0.088 for 4401 (SO-102)

film. In general, high-speed emulsions are coarse grained, and slower emulsions are fine grained.

Granularity depends not only on the physical and chemical characteristics of a film, but on the type of developer used, the development time, etc. For a given set of conditions, granularity increases with density. Since increased granularity tends to decrease the definition of an image, overexposure should be avoided where maximum resolution is important. Granularity can be reduced with some types of developers, but this may be done at the expense of contrast attainable in the final result. Decrease of contrast will tend to counterbalance the effect of the finer grain in obtaining high resolution.

2.4.9.3. Resolving Power using a Bar Pattern. The resolving power of a photograph or photographic system is usually tested by the use of a standard target pattern consisting of sets of three black bars mounted on a light background. These sets of bars vary in size usually by the ratio of $1:\sqrt{2}$ over a large range, and the resolving power of the system is determined by noting the size of the bars which can just be distinguished from each other.

The use of sinusoidal methods of analyzing photographic systems, by measurement or analysis of modulation transfer functions has certain advantages over tests involving visual observation of test patterns, but is not as generally accepted as the use of the test-bar target system.

Resolving power of a film material is increased by having a narrow spread function and fine grain. An increase in gamma also aids resolving power, since it increases contrast in the film for a given contrast of the target. The ability to resolve an image is reduced by about half when the contrast is lowered from 1,000:1 to 2:1.

Although the resolving power is increased by increasing gamma, it tends to decrease as the granularity of the image increases with increasing exposure. As a result of these counteracting effects, the optimum resolving power occurs when the exposure of a film falls on the lower part of the straight portion of the characteristic curve. At the limit of resolving power, the change in image contrast between dark and light parts of the test pattern can just be distinguished from the variation due to the granularity.

The resolving power is usually stated as a single value for specified conditions, but in actual fact varies about this mean value. For example, a test of type 4404 film under certain test conditions indicated that it had a resolving power of 316 lines per mm, with a standard deviation of about 20.

The standardized bar target system for determining and expressing resolving power has been used for many years. It measures the combined effect of all parts of the system which affect the definition of the image. In addition to the emulsion, it accounts for the effect of lens imperfections, image motion compensation error, vibration, film characteristics, amount of exposure, variations in the developer or developing process, and limitations of human vision. On the other hand, the method has the disadvantage that it is frequently not possible to determine how much each factor contributes to the degradation of the system. Sinusoidal MTF analysis is better for this purpose

2.4.10. COLOR FILM. The manner by which color film utilizes the basic photographic emulsion in three layers to reproduce the color of the image is covered in good detail in "Color as Seen and Photographed," Kodak Color Data Book, 1950, and in the Manual of Color Aerial Photography, American Society of Photogrammetry, 1968. The reader is advised to consult these references. Only a few additional remarks are needed here.

In the case of color film, the principal intention is to replicate what the eye would have seen in place of the camera. For that purpose only three dyes and three emulsions are required. A fourth dye and emulsion layer would not be useful for normal viewing. If one should measure the spectrum of a small area of a color film, one would observe only the characteristic transmittance properties of the three dyes in certain proportions in the film. A spectrum of the object in the field would not be observed. Any number of spectral distributions can create the same color sensation to the eye.

The use of stereocolor permits the recording of essentially all that the human eye requires for making the normal inferential connections between what is perceived and what is there from one viewing position and time. However, the vertical aspect and small scale of aerial photography require the use of new object attributes by the photointerpreter to make these connections.

2.4.11. COLOR TRANSLATION. The use of Infrared Ektachrome presents an additional alteration in the attributes available to the photointerpreter. Infrared Ektachrome film is one example of many possible color translation systems. In ordinary color film the reception of blue light activates the appropriate dye system to replicate the blue light exitance from the processed film. Similarly, the received green light activates the dye system for replicating the green light exitance of the processed film and likewise for the received red light. In a color translation system the received color activates a different dye system, so that the exitance color corresponding to the received radiation is either radiation of always shorter or always longer wavelength than that of the received radiation. In the case of Infrared Ektachrome, the blue light is to be filtered from the light falling upon the film and the received green light is made to produce the blue exitance from the film, the received yellow and red

produce the greeen exitance from the film and the received near-visible infrared produces the red exitance from the film. Thus, the color perception via the film is equivalent to what the eye would have seen if the color response of the eye were translated toward the long-wavelength portion of the spectrum. The inference concerning material properties must be intellectualized at first in order to make use of the film colors. One can see that any number of color translation schemes can be considered. Modified color translation might be made from the ordinary Ektachrome or Infrared Ektachrome by using appropriate filters. Ordinary Ektachrome should be used with an ultraviolet filter because ultraviolet radiation will activate the blue exitance of the film along with the blue. One could conceive of using a filter which allows all radiation to pass except the visible blue, in that case only the ultraviolet will cause the blue exitance from the processed film while green and red would be replicated as received. Using Infrared Ektachrome with a filter which stops only blue and green light would produce a combination where ultraviolet would cause blue exitance, red and yellow would cause green exitance, and infrared would cause red exitance. (Here one would have to assume that the dyes used in the second and third layers will not be penetrated by the ultraviolet).* The manufacturing of multiple layer films requires extensive research and development to result in a reliable product. While a large number of color translation schemes and modified false color films can be conceived, it is not likely that they will be manufactured unless a large market for these is visualized by a manufacturer. Experimentation is best done using black and white films in multispectral cameras.

*According to "Kodak Data for Aerial Photography," three layers of Ektachrome Infrared type 8443 will be activated by the visible blue. Whether or not ultraviolet will be stopped in the first layer is not stated.

3. CAMERAS

A camera is the system of components for creating the real image at the surface of the photographic film so that it may be exposed and recorded as needed. It consists of an optical system of lenses to produce a real image; a platen film holder to hold the film in the focal plane of the optical system; an iris or aperture stop to adjust the value of E_v(at film) to an appropriate range of values; a shutter to produce the exposure $q_v = E_v t$ at the appropriate time and for the required time interval; a camera body to hold the optical components and film platen at precise positions as well as to form a light tight enclosure to prevent unwanted radiation from exposing the film; a film transport or Drive mechanism to bring fresh film onto the platen for a new exposure, and a boresighted view finder to select visually the subject to be recorded. On some cameras additional parts are an exposure meter to determine the appropriate aperture setting and time interval of exposure; a range indicator to ascertain the proper focus setting, and a motion compensation system to hold a moving image fixed relative to the film during the exposure interval. In aircraft camera installations the camera may be mounted upon a shock-absorbing stabilized platform to reduce the vibration of the camera due to aircraft frame vibrations and to maintain as near to vertical aspect as possible during minor aircraft motions. Some means for measuring the ground speed must be used to set the image motion compensation correctly. If the camera is located at an advantageous site in the aircraft, frequently the operator may be located at some other position. A remote-control system would be required to set the camera into operation.

The use of photographic cameras at high altitudes introduces some possible problems which are not of consequence at ground level. The temperature and water vapor content of air at high altitudes is similar to that of the air over arctic regions. Temperatures of -60°F can be expected. The reduced temperature makes some film stiff and brittle, reduces the speed rating for exposure, and the low water vapor content of the air may permit the electrostatic discharge of accumulated frictionally generated electric charges in film. Such discharges expose the film in a characteristic dendritic pattern at various positions along the film. If the camera optics are not thermally controlled, they may retain the low temperatures of the upper atmosphere during rapid descent to warmer, more humid air. Condensation may occur on the optics so that the image-forming capability of the optics are at least temporarily spoiled. It is also possible for the shutter speed to change with extreme changes in temperature.

3.1. OPTICAL SYSTEMS

Almost all camera optical systems consist of more than a single lens element. Even relatively low-cost amateur cameras of the 35 mm film size variety frequently use more than a

single lens element. Expensive high-precision mapping cameras require optical systems containing as many as 12 separate lens elements in order to provide low distortion and high resolution over the full field of view of the camera. The resolution, as measured by bar pattern or by MTF, changes from one position in the field of view to another. Usually the resolution is best along the optic axis of the lens system. The resolution of a lens system or lens system and film may be given as the area weighted average resolution (AWAR) in lines per millimeter at the film plane. Figure 3-1 shows the specifications of the Geocon I lens manufactured by J. G. Baker. The distance from the back principal plane to the focal point is 6 in. Since the principal planes are not material parts of the lens, the location of the back focal point relative to physical parts of the lens structure is stated. In figure 3-2, the resolving powers of the lens-film combinations are given. The resolving power depends both on the angular position of the test pattern relative to the optic axis as well as upon the orientation of the test pattern, radial (test pattern lines perpendicular to the radial direction) or tangential (test pattern lines perpendicular to a circle at that angle from the axis). Notice that the AWAR is also stated but that the AWAR depends upon the film type and target contrast.

The distortion of the image from a perfect geometrical projection of the field of view is also shown. Recall that distortion is caused by a variable lateral magnification as a function of the angle off axis from the optic axis. The displacement measured in the film plane as a function of the angle off axis is the amount that the image is misplaced from where it should be along a radial direction from optical center of the film plane. Here, the lateral magnification is less at 15^{o} than it is at the optical center. At 45^{o} the lateral magnification is greater than at the optical center. If an aerial photograph were taken with this lens at an altitude of 10,000 feet, the mislocation of the image at 40^{o} off the optical axis would correspond to a mislocation of 15 cm on the ground. However, at 40^{o} the best resolution will separate lines $\frac{1}{40}$ mm apart or 50 cm on the ground so that correcting for distortion in location may be unnecessary.

In the graphs of figure 3-2 the distortion and the relative illumination as a function of angle off axis are plotted. If a uniform Lambertian reflector were photographed, the irradiance in the film plane would be greatest on axis and reduced as indicated off axis toward the edge of the field of view. This fall off of irradiance is inherent in a camera because of the geometry of the camera optics. A camera with a perfect simple lens viewing normally a Lambertian reflector surface will produce an irradiance in the image plane proportional to $\cos^{4}\theta$ where θ is the angle off the optic axis. Figure 3-3 shows the geometry of the camera illumination. If the Lambertian reflector has a radiance, L, then on the optic axis the power entering the camera from area A is $P = LA\left(\frac{A_0}{R^2}\right)$. This power is imaged evenly over the image area $\underline{a}$ of A so that the irradiance at the film plane on axis is

$$E(\text{axis}) = \frac{P}{a} = \frac{LA}{a}\left(\frac{A_0}{R^2}\right).$$

The power from area A at angle θ from the axis entering the camera is $P(\theta) = (L \cos \theta A) \times \frac{(A_0 \cos \theta)}{(R/\cos \theta)^2}$.

$$P(\theta) = \frac{LAA_0}{R^2} \cos^4 \theta.$$

This power is spread evenly over image area a of A so that the relative irradiance is

$$\frac{E(\theta)}{E(\text{axis})} = \cos^4 \theta.$$

An actual camera with a long optical system will likely have a somewhat different reduction of irradiance as a function of angle of view from the optic axis due to internal apertures which may become effective at large angles. The relative illumination is plotted as shown in figure 3-2. Mapping cameras are designed to achieve geometric fidelity as the primary goal. The use of the mapping camera as a quantitative radiometer may not be profitable because the variable irradiance in the film plane must be taken into account at each angle off axis. Insertion of a suitable grey filter over the film plane could correct for the irradiance variation to permit semi-quantitative radiometric measurements. The variation of irradiance as a function of off axis angle is sometimes referred to as the cosine fourth fall off.

3.2. IRIS OR APERTURE STOP

A lens which is designed to have a low f number (a high speed lens) may have more distortion and less resolution than one with the same number of lens elements but with high f number. Insertion of a variable aperture stop allows for the use of a high-speed low f number lens as a low-speed high f number lens with improved resolution. In addition, for photography of subjects which are located both near to and far from the camera the low f number lens may not be capable of focusing both subjects at one setting while a high f number lens might do so. Hence, in the amateur camera used principally for photography of scenery, a variable aperture stop is an important addition. Aerial photography, on the other hand, is rarely done where subjects are both near and far in the same field of view. Furthermore, resolution in aerial photography is usually limited as much by image motion and vibration effects as by the lens film capability. Resolution tests in the air rarely produce as good results as the bench-tested resolutions listed in the specification sheets. Therefore, the aerial camera is more likely to be operated at full aperture with the shutter speed being set at the highest value which will permit proper exposure, q_v. The variable aperture is not a critical component if shutter speed is continuously variable.

3.3. SHUTTERS

There are two popular positions for a shutter. The "between-the-lens" shutter is located close to the aperture stop of the optical system and may be an integral part of the lens system. The aperture stop is the smallest hole through which all power flows so that the smallest linear motion of a shutter is required at this position. A shutter near the aperture stop permits the exposure of all parts of the film during exactly the same time interval. High-speed between-the-lens shutters may not work reliably without elaborate engineering. Exposure times of 0.001 seconds are not uncommon. When one realizes that a certain time interval is required for the shutter to move from the fully closed to fully open position and return to the fully closed position, as well as to remain open for a very brief time, one can see that for exposure times of 0.001 seconds most of the exposure may occur during the opening and closing interval.

The second most popular position for the shutter is immediately in front of the film plane. Such a shutter is called a focal plane shutter. It usually consists of an opaque thin sheet material with an open long narrow slot in it. As the shutter is set in motion along the narrow dimension of the open slot, the slot traverses the film plane exposing the film along a line at each position as the slot passes by. See figure 3-4. The "speed" of the shutter or the exposure time is set by the width of the slot and the velocity of the sheet. The action resembles the motion of a window shade on its roller. In order for the exposure to be uniform the slot must be up to the top speed before it reaches the first edge of the film and the speed must be kept reasonably constant across the film. Focal plane shutters can achieve very high shutter speeds; however, the full film is not exposed during the same time interval. The focal plane camera shutter reduces the degradation due to image motion by making the slot sufficiently thin, but the geometric fidelity may suffer if the left half of the film is exposed at a significantly different time than the right half.

4. PHOTO INTERPRETATION

The interpretation of aerial photographs is the art of inferring correctly the quality and quantity of objects on the surface of the earth from photographs taken by the aerial camera. Since the principal use of the aerial camera is to replicate what the eye would have seen if the eye were at the position of the camera at the time of exposure, photo interpretation will be able to draw upon the normal inferences which are used by the observer using his eyesight directly. Because so many of the normal inferences can be made in photo interpretation, there is a possibility that the need to make completely new inferences from aerial photographic images will be overlooked. The change in scale and aspect introduce considerable difficulty and require the use of new attribute measures for establishing an object's identity. For example, a person may identify the species of a tree from the detailed structure of the leaf, twigs, and bark. At long range on the ground an observer may use the shape of the tree as the identifying attribute without resolving any of these details. From the aerial photograph the observer may have none of these attributes to aid him. He must seek out new attributes that are available from that aspect and range.

A discussion of the fundamentals of photo interpretation is given in Chapter 3 of the Manual of Photo Interpretation. The reader is advised to study chapter 3.

It is not uncommon to consider that photo interpretation is a game in which it is not fair to utilize any information not associated directly with the photograph. Possibly, this impression is given strength unintentionally by the methods used in explaining the principles of photo interpretation from photographs where the ancillary information is withheld for tutorial purposes. Rarely is it necessary to undertake photo interpretation tasks in such a vacuous circumstance. The geographical context is almost always known for each aerial photograph. The culture of the society, the major means of livelihood, the common means of transportation, the reported geological history of the region, the reported rainfall and prevailing wind directions, the reported extremes in temperature, and the common items of trade and manufacture connecting the economy of the area with others are generally available in published literature. To attempt photo interpretation without any of the available ancillary data would be an egregious error.

The common way in which one learns to recognize an object in an aerial photograph is to examine many aerial photographs of known objects of importance to one's mission so that the features of size, shape, shadow, tone and color, texture and pattern can be recognized in order to classify a similar unknown object. The mind abstracts the salient features from specific examples so that an unknown object viewed with the aspect and range of aerial photography can be recognized. The process relies heavily upon the capacity of the mind to generalize. An additional aid to generalization, however, is to establish a conceptual model of the cause for

the appearance of the object in the photograph so that fewer specific examples are required to make a successful generalization. One such model for explaining the reduced infrared reflection of diseased plants is shown on page 614 of the Manual of Photo Interpretation. The model is developed for an oak leaf but has been generalized to all leaves having a similar structure. This model appears to work well for explaining the reduction in tone of diseased trees in infrared aerial photography. However, recent reflectance measurements at The University of Michigan on a leaf under moisture stress showed an increase in infrared reflectance rather than a decrease, as the model would predict. Until the contradiction is resolved, it is wise to use the model with caution.

The availability of a model for the cause of tone, color, or texture in the photograph may make possible the reliable generalization extending the inferential connection between cause and effect beyond the obvious one. For example, the model to explain the cause of crop marks is shown on page 727 of the manual. It can be extended to explain the markings on the surface due to archeological artifacts when the artifacts have long been buried. The cause of the crop mark is preferential crop growth. The cause of the preferential crop growth is uneven subsurface conditions. Uneven subsurface conditions can be caused by, among other things, buried artifacts.

A treatment of photography from satellites is given in two NASA technical notes: "Terrain Photography on the Gemini IV Mission Preliminary Report," NASA TN D-3982; and "A Review of Photography of the Earth from Sounding Rockets and Satellites," NASA TN D-1868.

5. INTRODUCTION TO PHOTOGRAMMETRY*

5.1. INTRODUCTION

This chapter is designed to present a broad overall picture of the theory and practice of the science of photogrammetry with an almost exclusive emphasis on its use in the cartographic field. In considering the application to the cartographic field the following subject matter will be treated: aerial photography, orientation, rectification, photogrammetric instruments, determination of control points, and production of maps. Brief mention will also be made of some of the noncartographic uses of photogrammetry and of sources of information on the photogrammetry field. For those who are most probably not engaged in a career in photogrammetry it is hoped that this chapter will not only present the overall picture of the theory and practice of photogrammetry but perhaps awaken and generate an interest and enthusiasm for a field which this author has found fascinating as well as rewarding.

5.1.1. DEFINITION OF PHOTOGRAMMETRY. Let us first define the science which we plan to investigate—photogrammetry. If we investigate it from the standpoint of semantics we see that the word is composed of two parts, PHOTOGRAM and METRY. Photogram is simply another word for photograph and metry signifies the science of measuring. Thus this approach would yield the following definitions: "Photogrammetry is the art or science of obtaining measurements from photographs." There are of course many variations of this basic definition. For example, a more general definition is found in the book Basic Metrical Photogrammetry by Lyons—"Photogrammetry is the art or science of obtaining information from photographs." On the other hand a more specific definition is found in Webster's New World Dictionary where it is defined as "The art or process or surveying or mapping with the help of photographs."

5.1.2. USES OF PHOTOGRAMMETRY. The definition found in Webster's Dictionary is an indication that photogrammetry finds a wide use in the cartographic field. This is indeed the case and at the present time it is by far the most extensive use of this science. In fact photogrammetry has revolutionized the surveying and mapping field with a large percentage of maps produced in the world today being done so through the use of photogrammetry. However, photogrammetry is also finding use in situations that are not associated with surveying and mapping. This application is often called non-cartographic photogrammetry and involves a wide variety of data determination from photographs.

---

*Prepared by Paul E. Norman, NASA, MSC, Houston, January 1967.

5.2. CARTOGRAPHIC APPLICATION

5.2.1. INTRODUCTION. When the term cartographic application is used it means the use of photogrammetry in the surveying and mapping field. Specifically, we are talking about the production of planimetric maps, topographic maps, and the execution of other surveys. A planimetric map is one which presents only the horizontal position for the features represented. The natural features shown on a planimetric map usually include rivers, lakes, seas, mountains, valleys, cities, highways, buildings, boundaries, etc. In contrast, a topographic map portrays, in addition to the planimetry, data regarding height or altimetric information. There are many ways to portray this altimetric data, one of the most common being that of contours. Thus from the standpoint of the measurements necessary to produce the maps, a planimetric map will require the determination of two-dimensional data (X and Y) and a topographic map three-dimensional data (X, Y and Z). Obviously, under at least the traditional concepts of maps, what must be done therefore is to measure the required data and plot it to scale on a graphic medium; the result is a map. In considering this cartographic process the key word from the preceding sentence is no doubt "measure." To what precision must one determine the data which will be portrayed, the result of which will be our map? There are of course specifications which cover the measurements, or surveying as it is commonly called, and upon investigation these are found to be of high precision. For example, the national map accuracy standards of the United States state that 90% of all well-defined points must be located on the map within 0.5 mm for maps smaller than a scale of 1:20,000. For a scale of 1:20,000 this is 10 meters, and the survey of the feature to be mapped must be at a much greater precision. The same high standards apply to height measurements. What this means in practice is that the surveying must be done very carefully and to a high precision.

Let us first look at the means of obtaining the necessary data for the production of a map before the advent of photogrammetry. Once the basic control was established then the details to be mapped were surveyed by ground methods such as theodolite (transit) and tape, and the data plotted to form the map. Such procedures have been used for many years and in fact continue in use. However, with the advent of photogrammetry it became obvious that its use in the production of maps and in the surveys necessary for these maps could offer a number of definite advantages—namely speed, economy, and greater ease in mapping remote and difficult features.

Thus far we have mainly considered the use of photogrammetry in the surveys necessary for the production of maps. There are, of course, many other surveys carried out which, although they do not result in the production of the maps, are nevertheless possibilities for the use of photogrammetry. I speak now of surveys for engineering projects such as highway and

railroad construction. These surveys are very similar to those needed to produce maps and thus we will deal with them in this section on cartographic applications.

5.2.2. INVESTIGATION OF USE OF PHOTOGRAMMETRY IN SURVEYING AND MAPPING

5.2.2.1. Introduction. We have set the stage for a consideration of the possible use of photogrammetry in the surveying and mapping field. We will now go through a step-by-step development of a general process whereby these surveys and maps may be produced through the use of photographs—i.e., photogrammetry. We will first start with an aerial photograph, consider how it is obtained and if it can be used in its unadulterated form to produce our surveys and maps. It will then be shown that the photograph itself can give only two-dimensional data, and thus is suitable only for planimetric maps. Furthermore, a study of the geometry of the aerial photograph will show that the errors contained in the photograph render it, in its unadulterated form, unsuitable except in special cases for even planimetric maps. The development then proceeds to show that by taking our photographs under different conditions (60% forward lap) and by reprojecting parts of successive photographs we can not only add a third dimension but can remove the undesirable errors previously mentioned. This treatment then will lead us into interior, relative, and absolute orientation. With these steps we will have the basis to be able to use the photographs to perform the measurements necessary for our surveys and maps. However, we will need facilities to physically put our theory into operation and thus the next step will be to develop the instruments necessary to carry out this theory. Once this is completed we then consider photogrammetric means for determining the control, and finally we will put everything together and see how we will actually produce our surveys and maps.

5.2.2.2. Aerial Photography

5.2.2.2.1. Cameras. Obviously we cannot use photogrammetry without the photogram or photograph; thus the logical place to commence is in the acquisition of our photographs. Furthermore, since we want to map the earth it would seem logical that we take our photographs from above the earth, or in other words aerial photography. To acquire such photographs we will need a camera, so let's start with that item. The basic principle of the camera is that images pass through the lens and are brought to focus by that lens in a plane perpendicular to the optical axis. This basic concept can be seen in figure 5-1. Here we see an object (an arrow) which passes through the lens and is imaged in the plane B-B. Furthermore, it is also well known that if we put a piece of light-sensitive photographic material in plane B-B we can capture

the image of the arrow. If we apply these principles to our case then we see that our aerial camera must obviously have a lens, a body housing this lens, and an image plane. It will further be necessary to provide facilities for film, its exposure, and its transport to the next frame. This camera, it will be seen must be of very high precision, rigid construction, foolproof, and at the same time mobile enough to be used in an aircraft. It will further be subjected to large temperature changes, vibrations, and other physical phenomena. All of these requirements mean that the aerial camera is a complex instrument that will bear little resemblance to the cameras you and I are used to using. This camera for example has the dimensions $70 \times 60 \times 50$ cm, weighs about 500 lbs, and costs approximately $18,000.

5.2.2.2.2. Photographic Operations. Once we have our camera we can begin to consider the photographic operations. First it must be decided what type of aerial photographs will be needed. It will be found that the choice is one of three: high-oblique, low-oblique, or vertical photographs. The high oblique is characterized by the fact that the horizon appears in the photograph. This, of course, means that the optical axis of the camera was inclined when the photo was taken. An exaggerated sketch of a high oblique can be seen in figure 5-2; a good example of a high-oblique photograph is the trimetrogon photography. By comparison a low oblique will not contain the horizon, though the optical axis will still be inclined. This can be seen in figure 5-3. The vertical photograph, as the name implies, is the case when the optical axis is vertical, at least in theory. This can be seen in figure 5-4. By far, vertical photography is the most widely used for surveying and mapping; thus our discussion will be limited to the use of vertical aerial photography. In further consideration of the type of aerial photography, black and white photography is by far the most commonly used, with other types such as infrared and color at the present time finding use in special cases.

With respect to photo coverage we will for the present time restrict ourselves to simply getting complete coverage. Thus what we will do is to fly our photography in bands in such a way as to have just enough overlap to insure coverage (see fig. 5-5). Of course we will later find that this type of coverage will not suit our purposes, but for the time being we shall be content with this configuration. What we now have is complete coverage of an area with vertical aerial photography. The logical question then is: To what extent can this photography be used to produce planimetric and topographic maps?

5.2.2.3. Geometry of Aerial Photographs.

5.2.2.3.1. Introduction. In order to answer the above questions we will investigate the geometry of the aerial photograph. First let us take a brief look at two types of projections,

the orthogonal projection and the central projection. The orthogonal projection consists of parallel rays which are perpendicular to the projection plane (see fig. 5-6). This type of projection is used to plot the details which appear on planimetric and topographic maps. A central projection is one where all rays pass through a common point or vertex, and can be seen in figure 5-7. Of particular interest is the fact that the photograph is the result of a central projection between the object and the image.

For the study aerial photograph geometry some definitions will be necessary and will be given with the assistance of figure 5-7. The point through which all rays pass is the projection center (O) and in practice is assumed to be the center of the lens of the aerial camera. The photo nadir point (n) is defined as the intersection between the image plane and the vertical line passing through the projection center. The intersection of this vertical line with the ground is the ground nadir point (N). The principal point (h) is the intersection between the image plane and a line passing through point O and perpendicular to that plane. The distance from the projection center to the image plane is the principal distance, often erroneously called the focal length.

5.2.2.3.2. Scale. One of the first things that will be of interest if we are to use photogrammetry for mapping is the determination of the scale of the photograph. Scale is expressed as the relationship or fraction between the same horizontal distance in the terrain and in the photograph. Let us first consider scale as the relationship photograph/ground.

In figure 5-8, by similar triangles, we see that:

$$\frac{a}{A} = \frac{c}{Z}$$

where A = ground distance

a = photo image of the ground distance

$\left(\frac{a}{A}\right)$ = the scale relationship

c = the principal distance

Z = the flying height

What is usually done is to select "a" equal to one unit and the corresponding equation then becomes:

$$\frac{1}{A'} = \frac{c}{Z}$$

Thus we can compute the scale by knowing the principal distance and the flying height.

Examples

(1) c = 20 cm (0.20 meters), Z = 5,000 meters

thus $\left(\frac{1}{\text{scale}}\right) = \frac{1}{A'} = \frac{c}{Z}$

$$\frac{1}{\text{scale}} = \frac{0.20}{5000} = \frac{1}{25{,}000}$$

What this means, of course, is that one unit on the photograph is equal to 25,000 units on the ground.

(2) c = 15 cm, Z = 5,000 meters

$$\frac{1}{\text{scale}} = \frac{1}{A'} = \frac{c}{Z} = \frac{0.15}{5{,}000}$$

$$\frac{1}{\text{scale}} = \frac{1}{33{,}333}$$

(3) c = 10 cm, Z = 5,000 meters

$$\frac{1}{\text{scale}} = \frac{1}{A'} = \frac{0.10}{5000}$$

$$\frac{1}{\text{scale}} = \frac{1}{50{,}000}$$

We should remember that our relationship is photo/ground and is characterized by 1/A which for clarification purposes has been written as 1/scale. We can also express the scale as the relationship ground/photo which simply means that our mathematical development would then read:

$$\frac{A}{a} = \frac{Z}{c}$$

Now if we let a = 1 we have:

$$\frac{\text{scale}}{1} = \frac{A'}{1} = \frac{Z}{c}$$

Let us now redo the same examples which were previously done

(1) c = 20 cm, z = 5000

$$\text{scale} = \frac{A'}{1} = \frac{5000}{0.20} = \frac{25{,}000}{1}$$

(2) c = 15 cm, z = 5,000

$$\text{scale} = \frac{A'}{1} = \frac{5000}{0.15} = 33{,}333$$

(3) $c = 10, Z = 5000$

$$\text{scale} = \frac{A'}{1} = \frac{5000}{0.10} = 50{,}000$$

There is of course no difference between whether we use 1/A' or A'/1 to express the scale. However, the two cases often give rise to considerable confusion.

It will, of course, have been noticed that the scale development is valid only for flat terrain. Let us now consider the general case, as illustrated in figure 5-9.

Our first glance shows us that something is wrong; two equal photo distances correspond to unequal ground distance. Mathematically speaking we have

$$1/\text{scale}_1 = 1/A'_1 = c/Z'_1$$

$$1/\text{scale}_2 = 1/A'_2 = c/Z_2$$

$$\text{but since } A'_1 \neq A'_2$$

$$1/\text{scale}_1 \neq 1/\text{scale}_2$$

In other words we would not get the same scale because $Z_1 \neq Z_2$. The conclusion is that when our terrain is not flat only points which lie at the same elevation will have equal or constant scale in the photograph. Thus the photograph will have a variable scale except in special cases. What we do in practice is to compute an approximate or average scale which is based on the flight height above the nadir point or mean terrain and designated as Z-h (see fig. 5-10). Let us now take a practical example, a photo whose approximate scale is desired. What we need is c, Z, and h. The first two values will usually be found in the margin of each aerial photograph; in fact, we will normally find the following information in the margin:

Date
Time
Camera serial number
Principal distance
Flying height
Photo number

In the example to be used the altimeter reading is 9,880 ft (3011 meters) and c = 151.455 mm. In this case we have a control point which falls in the photograph. Its elevation is 393, so Z - h = 3011 - 393 = 2618. The scale then is 2618/151.455 = 17,286. Had there not been a control point, we would probably have been forced to use just the altimeter reading and the scale = 3011/151.455 = 19,880.

Thus far, our scale determination has been by means of the flying height and principal distance. If we return to our basic equation we had

$$\frac{\text{scale}}{1} = \frac{A}{a} = \frac{Z}{C}$$

What we can also do is to use A/a and use a known distance on the ground and its photo image. For example, if

A = 155.0 meters

a = 9.89 mm

then the scale = $\frac{155.0}{.00989}$ = 15,672

In summary, we can determine the scale by knowing Z - h and c and using the relation $\frac{c}{Z - h} = \frac{1}{\text{scale}}$ or $\frac{Z - h}{c}$ = scale.

We can also determine the scale by comparing the values for the same distance in the ground and in the photo:

$$\text{scale} = \frac{A}{a}$$

The scale will be exact and constant only when the photograph is exactly vertical and the ground is flat. In all other cases, e.g., accidented terrain, the scale is variable. Thus any map that we make from this single photograph will most likely be approximate.

5.2.2.3.3. Relief Displacement. We will now continue with the study of the geometry of the aerial photograph and investigate why for accidented terrain we have a variable-scale photograph. It will be found to be due to a characteristic often called relief displacement, which can be seen in figure 5-11. What we have is a vertical photograph of an area exactly flat except for a vertical needle A-B. If seen in an orthogonal projection point A would be projected on top of B. However, due to the central projection A and B will occupy different positions. It is common to say that in the photograph, B is displaced from A, thus the name relief displacement. Reverting to the scale a minute, the horizontal distance NA = NB but using the point "a" in the photo would give us one scale while using point "b" another. Why? Because

$$\frac{NA}{na} = \frac{Za}{c} = \text{scale}'$$

and

$$\frac{NB}{nb} = \frac{Zb}{c} = \text{scale}''$$

and obviously scale' $\neq$ scale''

Thus our variable scale situation is due to relief displacement caused by the central projection. Now if one thinks about the situation it is clear that there is one ground point that will not displace, the nadir point (ground). All other points will be affected with the amount of displacement increasing as one approaches the edge of the photograph. Thus it should be obvious that the relief displacement is radial from the photo nadir point. It can be expressed mathematically as

$$\Delta r = r\frac{\Delta Z}{Z}$$

where Δr = the relief diplacement

r = the radial photo distance to the top of the image

ΔZ = height difference

Z = the flying height above the lower point of ΔZ

This can clearly be seen in figure 5-12. This equation can be easily derived from similar triangles; the student is invited to do so. In order to see if the amount is significant we compute Δr for two cases.

Case 1. Z = 6,000 meters
r = 11 cm
ΔZ = 100 meters

Note: these values correspond to S = 40,000 C = 15

$$\Delta r = 11\frac{100}{6000} = 1.8 \text{ mm}$$

Case 2. Z = 3,600 meters
r = 11 cm
ΔZ = 100 meters

$$\Delta r = (11)\frac{(100)}{(3,600)} = 3.1 \text{ mm}$$

It can be seen that the relief displacement is certainly significant, especially in view of the fact that we are allowed less than 1mm maximum error in the map. An interesting additional note is encountered if one analyzes the relationship between Δr and c for cases of final photo scale. It can be seen that as c increases, the flying height for a constant scale increases, which decreases the relief displacement.

The conclusion about relief displacement is that it causes a variation of scale that makes any map produced from a single uncorrected photograph quite approximate. However, we should not lament too much the evils of relief displacement because it is this characteristic of a central projection which later will be seen to give us our third dimension.

5.2.2.3.4. Tilt Displacement. We have yet a third aspect of the geometry of an aerial photograph and that is tilt displacement. We have been assuming that our camera will be exactly vertical when the photos are to be taken. However, if the camera axis is inadvertently inclined,further image displacement will be caused, this can be seen in figure 5-13. Here one sees a square figure ABCD which is imaged as a trapezoid in the photo. The deformation is due to tilt and the displacement of the images is called tilt displacement. Thus we find another source of error in our single photo, which will cause the photo to have a variable scale. However, it will subsequently be seen that there is some benefit to be derived as well from the fact that an inclined camera causes images to be displaced.

Thus the final conclusion is that except in very special cases (vertical camera and flat ground) we will have image displacements due to relief and to tilt. These displacements cause our photograph to have a variable scale and any map made by direct tracing would have rather large errors. We conclude therefore that we must look further for a way to produce our maps using photogrammetry and aerial photography.

5.2.2.4. Use of Photo Position

5.2.2.4.1. Reprojection and Intersection of Rays. Let us now review the situation. The single photograph will at best only produce a planimetric map. Furthermore, the precision of any such map will be largely diminished by relief displacement and perhaps tilt displacement. Our relief displacement, of course, is caused by the central projection and thus it would seem logical that a reverse process, involving reprojection of the aerial negative, could be used to remove its influence. If we disregard, for the moment, the possible optical problems, it would seem quite likely that this could be done. But what good would be done? We would have a line in space whose direction is correct, but from it we could not easily determine spatial position. However, let us change the photographic procedure in such a way that each point on the ground is imaged on two consecutive photographs (see fig. 5-14). Now if we project two consecutive photographs it would seem that we could intersect the rays a'O' and a''O'' and locate the position A. Let us analyze the situation to make sure we can do this.

If we were to place the cameras in the sky in the position occupied at the moment of exposure and reproject the aerial negatives, it is obvious that we could locate point A by intersection. In practice, we cannot do this as we would need an instrument several miles high and wide. However, what would happen if we moved our two cameras together in x direction only. It would be found that the rays continue to intersect and thus we would have a reconstruction of the terrain not on a 1:1 scale; but a scale determined by various factors including the principal

distance, the projection distance, and the X distance between our projecting cameras. This can be seen in figures 5-15 and 5-16. The student should construct a figure similar to 5-15 and from it develop figure 5-16 on a smaller scale. Aside from a relation in scale, the two figures 5-15 and 5-16 have a common factor—the principal distance of both sets of projectors is equal. This, of course, is not an accident; it is required. Why? Because when the image is exposed its direction was established by the line O'a' (see fig. 5-14). In the reprojection the direction will be correct only if the same principal distance is used. Furthermore, the principal point of the photograph must be over the optical axis of the camera (projector). These two requirements, centering the photograph and the establishment of the correct principal distance are called the interior orientation.

We might further add a screen (tracing table) to pick up the projected images and now we can expect some results. What we will do is to first take our photography so that it overlaps at least 50% as in figure 5-14. We will then reproject consecutive photographs, being careful to properly reconstruct the inner orientation and provide a screen to pick up the images. The screen will be movable in height and when both images are superimposed we will have the intersection and, hopefully, the correct height. Still disregarding the optical problem what will we find? What we find is that by raising the height of the screen we can remove the x difference. But when we do we have a y difference because we have not considered the tilt displacement and have put the negatives in cameras whose axis is vertical. Thus if our camera axis was not vertical at the moment of exposure, which is normally the case, we were not justified in placing the camera axis vertical. If we revert to our illustration of tilt displacement, figure 5-12, we see we can only transform the trapezoid back into the square by projecting from an inclined position. Thus each photograph must be placed the same absolute position it was in when the photos were taken. If we do this, the y difference (y parallax) will disappear.

What then are the possibilities or directions in which the camera could be inclined at the moment of exposure? The aircraft and camera could rotate around the axis of flight, often called omega or tilt (see fig. 5-17). The aircraft can rotate around an axis perpendicular to the line of flight, often called phi or tip (see fig. 5-18). The aircraft can rotate around a vertical axis, often called Kappa or swing (see fig. 5-19). These three angles ϕ, ω, κ, along with the exact position of the camera station (X_o, Y_o, Z_o) define the exterior orientation of a single photograph. The problem then is to determine what was the exact position and altitude of the camera at the moment of exposure. The most logical solution would be to measure the six values ω, ϕ, κ, X_o, Y_o, and Z_o as each photo is taken. However, the lack of precision and the expense involved have resulted in the fact that the method is not widely used. Further analysis of the

situation yields a solution which is much more precise and easier to apply. It can be proven that if the photographs are reprojected, under the condition of a correct inner orientation, that any y differences (y parallax) will be due to an incorrect relative position of the two photographs. By relative position is meant the position of one photograph relative to the other without regard to the ground. Thus if we reestablish the inner orientation, project the pairs of overlapping photographs, and have a screen to capture the images at the height where there is no difference in x, we can be sure the y parallax is due to an incorrect relative position of the two cameras. We can furthermore establish the correct relative position by a systematic removal of this parallax. This removal of y parallax is commonly called relative orientation.

Once we have a relative orientation, then what? The photos are correctly positioned one to another, but they will not be correct in the absolute sense. The second step, then, in the determination of the correct exterior orientation is the establishment of the correct relationship between the formed model and the reference coordinate system. This operation, often called absolute orientation, consists of rotating both projectors simultaneously about three mutually perpendicular axes and changing the model scale so that control points in the model coincide with their plotted coordinates. Once this absolute orientation has been carried out then we are in a position to determine the correct spatial position of any image point in the model.

Before examining in closer detail the orientations, let us briefly summarize. It has been shown that if we obtain our photography with at least 50% overlap we have the basis for making three-dimensional measurements of the earth's surface. It was further shown that if the inner orientation of the taking camera is reconstructed that we have the basis for establishing spatial position by reprojection and intersection. However, in order to be able to establish the correct intersection of the rays we must put each of the two photographs in the same exact position as when the photo was taken. This position is defined by three rotations (ϕ, ω, κ) and three coordinates (X_o, Y_o, Z_o). Since we have an overlapping pair of photos it was shown that the most expedient solution was to first determine the relative position (relative orientation) by removing y parallax through proper projector rotation and translation. The absolute position of both photographs is obtained by simultaneously rotating both photographs (and hence the formed model) into their correct absolute position by means of known control points. The absolute orientation is completed with the scaling and connection to a coordinate system. It is also well to note that there are, of course, a number of optical problems connected with the reprojection of the photographs. Such items will be discussed when the subject of instruments is treated.

5.2.2.5. Interior and Exterior Orientations. We will now take a closer look at the orientation procedures and consider how, in general terms, the inner orientation, relative orientation, and the absolute orientation can be carried out. Our purpose, of course, is to establish the correct absolute direction to rays carrying the images.

5.2.2.5.1. Interior Orientation. The interior orientation, as has previously been seen, reestablishes the same relative position in the projector that existed between the images and the projection center (the center of the lens) in the aerial camera. This is accomplished by centering the photographic negative so that its principal point lies over the optical axis of the projector. The second step is to set the proper principal distance so that it is equal to the principal distance of the taking camera. The centering of the negative or photographic plate is accomplished in a number of ways. One of the most common ways is to provide an instrument with a plate holder which has fiducial marks which locate the principal point of the projector. Thus the negative, when placed so that its fiducial marks coincide with those of the plate holder, is properly centered. The setting of the principal distance establishes the correct distance between the projection center and the principal point. There are generally two ways its setting is accomplished:

a. In projection instruments a projector for each class of photography is required and they have facilities for small changes. For example, a set of projectors would be provided for wide angle photography with a principal distance of 152 mm with a variable range of ± 3 mm. This is the case for the Kelsh Plotter and the Stereoplanigraph C-8.

b. In mechanical instruments the principal distances are determined mechanically and thus a wide range of change can be obtained. For example the Wild A-8 has facilities for setting principal distances from 100 mm to 210 mm and the Kern PG-2 from 85 mm to 170 mm.

Once the interior orientation has been established, then a point has the same coordinates x, y, c, with respect to the projection center in the projector as it did in the aerial camera (see fig. 5-20).

5.2.2.5.2. Exterior Orientation. The relative direction of the projected rays is reestablished by the interior orientation. The absolute direction of the ray, with respect to the coordinate system used for measurement, is established by the exterior orientation. We have seen that the exterior orientation of each camera is defined by the values ω, ϕ, κ, X_o, Y_o, Z_o. If we could measure the values and reduce the linear values to the desired scale we could directly establish the proper exterior orientation. However, the lack of precision and the expense of

such measurements have resulted in the fact that they are not widely used. Instead the exterior orientation is established in two stages—the relative orientation and the absolute orientation.

5.2.2.5.2.1. Relative Orientation. The relative orientation is the first of the two steps in the establishment of the proper exterior orientation. Returning to our example with the Kelsh Plotter, it will be remembered that we projected our photographs with the correct interior orientation and found that the images from each projector had an x and y displacement for an arbitrary position of the screen. (The x displacement we call x parallax and the y displacement y parallax.) More important, we can remove the x parallax by raising or lowering the height of our screen (tracing table) but when the px = 0 we found that the images were displaced in y, i.e., there was y parallax. Now we come to a very important point. It can be proven that this y parallax is due only to the lack of correct relative position between the two projectors. Conversely, if we manage to establish the correct relative position there will be no y parallax when px = 0 for any point in the overlapping pair. In other words, all rays will intersect. It has been shown that the relative position (relative orientation) is defined by three rotations and three translations. Stated mathematically, we would have:

$$\omega' - \omega'' = \Delta\omega$$

$$\phi' - \phi'' = \Delta\phi$$

$$\kappa' - \kappa'' = \Delta\kappa$$

$$X'_o - X''_o = \Delta X$$

$$Y'_o - Y''_o = \Delta Y$$

$$Z'_o - Z''_o = \Delta Z$$

These then are the parameters which define the relative orientation, and they are determined by eliminating y parallax in the model. The problem now is how to eliminate the py, obviously by moving one projector through the five degrees of freedom (ΔX will not cause a py thus is not considered for the relative orientation) until there is no py. Obviously we need a systematic way of carrying this out. We are assisted in this problem by an axiom of projective geometry which states that under our conditions, if we succeed in making five rays intersect, we can be assured that all rays will intersect. In other words, if we remove the py from five points the entire model will be clear. To be on the safe side we use six points of which one will be used as a check. Four such points will be located in the corners of the model and the other two at the principal points of the two projectors (see fig. 5-21). By analyzing the y parallax caused by small movements of each of the five orientation elements we see that the amount and pattern

of the parallax caused by each element is different (see figs. 5-22 and 5-23). In figure 5-22 we see the effects of $\Delta\omega$, $\Delta\phi$, $\Delta\kappa$, ΔY (by), and ΔZ (bz), and in 5-23 just the y conponent (py) and its effect on the six points. By analyzing the characteristic of each element, routines are developed which isolate the parallax, due to each of the five elements, so it can be eliminated. The result is a systematic way of eliminating the parallax which upon completion insures that the correct relative orientation has been established. It should be pointed out, however, that the routines are such that they are actually iteration procedures and must be repeated several times before all the parallax will be removed. Furthermore, the left projector may be oriented to the right projector, the reverse procedure, or both projectors may be used.

A commonly used orientation routine is:

a) make $p_1 = 0$ with by'
b) make $p_2 = 0$ with κ'
c) make $p_3 = 0$ with bz'
d) make $p_5 = 0$ with ω' and overcorrect (continue on in the same direction) and amount $(1/2)(Z^2/d^2 - 1)p_5$ where z = projection distance in mm $d = y_1 - y_3 = y_2 - y_4 = y_5 - y_1 = y_6 - y_2$
e) make $p_1 = 0$ with by'
f) repeat a to e until $p_1 = p_2 = p_3 = p_5 = 0$
g) make $p_4 = 0$ with ϕ'
h) check for parallax in point 6.

It will be noticed that in the above case the left projector is being oriented to the right projector. Another common routine using elements of both projectors is:

a) make $p_1 = 0$ with κ''
b) make $p_2 = 0$ with κ'
c) make $p_3 = 0$ with ϕ''
d) make $p_4 = 0$ with ϕ'
e) make $p_5 = 0$ with ω' and overcorrect an amount $(1/2)(Z^2/d^2 - 1)p_5$
f) make $p_1 = 0$ with κ''
g) make $p_2 = 0$ with κ'
h) repeat a to g until all parallax is clear

5.2.2.5.2.2. Absolute Orientation. Once the relative orientation is completed the model will be formed but its position with respect to the coordinate system will not be correct. Thus, we lack the absolute orientation. We can illustrate this idea with the help of figure 5-24a in which we see a model that is formed with two exactly vertical photographs, it should be obvious

that $\Delta\phi = 0$. In figure 5-24b we see the same model rotated around an axis perpendicular to the line of flight. We further see that the relative position is correct, i.e., $\Delta\phi = 0$. The difference between the two cases, as represented by figures 5-24a and b, is purely that of absolute position with respect to the measuring system. The effect of this lack of absolute orientation is apparent if we once again examine these two sketches. In figure 5-24a we see two points, A and B, being projected onto the reference plane producing points A' and B'. The distances A-A' and B-B' represent the heights and in figure 5-24b it is obvious that the height of point A is different from that in figure 5-24a. Furthermore, the x distance between projectors determines the model scale, and, since an x movement is not part of the relative orientation, the proper model scale will have to be determined. There is still a third effect that is not apparent in these sketches, that of azimuth. Since we will want to plot the points in the model with respect to a coordinate system, our plotting sheet must be properly lined up. Furthermore, since our model is in an arbitrary position it is most likely that such condition does not exist. Thus, in conclusion, the lack of absolute orientation results in errors in heights, and an incorrect scale and azimuth.

First let us consider the scale. In figure 5-25, we see a model formed at projection distance Z by using a value B for the base. We further see the established position of points A and B. If we assume that points A and B fall at the model nadir points and that the ground distance between the two points is 1000.00 meters, then by measuring the model distance we can get the model scale. In figure 5-25 we assume that A-B is 6 cm and we get a model scale of 0.06/1000.00 = 1/16,666. Next we move projector number two from O_2 to O_2'. This gives us a model formed at Z' with a base of B', and the points A and B are now formed at A' and B'. If this movement is carried out along the line of the air base (O_1 - O_2), we will not disturb the relative orientation. In fact we will remember that a movement in X of one projector (bx) is not even included in the elements of relative orientation because it should not cause a py. Thus we can assume that a proper X movement of one or both projectors will not affect the relative orientation, but will alter the scale. In fact if we assume that the model distance A' - B' is 10.7 cm we find the new scale of our model is 0.107/1,000.00 = 1/9346. In practice, scaling is done either by a numerical method or a graphic procedure. If it is done numerically, then we measure the model coordinates of two control points, compute the model distance, and compare it to the ground distance. If we do not have the desired model scale, which is the most likely case, then we alter the base in X until we obtain the desired scale. In the graphical procedure we simply plot the ground distance at the desired model scale and then alter the bx until the model distance between our control points is equal to that of the plotted distance.

In considering the azimuth we are, of course, dealing with the orientation of details on the map, normally carried out with respect to North. Since we will, no doubt, be using a coordinate system, we simply plot our control points and rotate either the model or the plotting sheet so that the azimuth of the model coincides with that of the plotting sheet. In practice the operations of scaling and azimuth determination are carried out simultaneously. We start out by lining up the plotting sheet so it coincides with one of our control points. We then rotate the plotting sheet, which is much easier than rotating the model, and change the bx until we get the other control point to coincide. Needless to say we will require several repetitions until we have our model in scale and in azimuth.

We have also seen that the lack of an absolute orientation affects the heights. However, we also know that we cannot alter the relative position of the two projectors, thus the only solution is to either change both projectors with respect to the reference system, or to move the reference system with respect to the projectors. We call these operations leveling. If we choose to move the projectors we have two choices. The first is to mount the two projectors on bars, with the simultaneous movements of the projectors being carried out by moving the bars. The reference plane remains fixed. The principle of this method can be seen in figure 5-26. This figure illustrates a rotation around the Y axis, called common phi, however, a similar type rotation is afforded around the X axis, which is called common omega. The concept of rotations using the instrument frame is used on many stereo-compilation instruments including Multiplex, Balplex, and Kelsh. The second possibility accomplishes the same final result as that of the frame but does it in a slightly different way. It consists of moving each individual projector in such a way so as to retain the relative orientation but at the same time change the absolute orientation of the model. In common omega this is nothing more than rotating each projector around the Y axis by the same amount, this can be seen in figure 5-27. In leveling in common phi using the elements of the individual projectors a combination of both phis, one bz, and bx are used. From figure 5-28 it can be seen that if we alter both phis the same amount, followed by a bz and bx change of one projector, we can accomplish the desired result. In figure 5-28a we see a pair of projectors with an inclined base which represents a lack of leveling in common phi. The required position is that as indicated in figure 5-26. The first step in our procedure is to rotate each projector the same amount in phi. This is indicated in figure 5-28b. The next is to raise the right hand projector, which is of course a change in bz. The final step is to apply a bx movement which can be seen in figure 5-28d. It can be seen that the final position of the two projectors is the same as obtained by the previous method using the bars. One also finds in use a combination of the two methods. For example in the Wild A-8 autograph, a stereocompilation instrument, there is a device that allows common movement in phi, but for leveling in omega individual movements of the projectors

are required. There is yet another possibility for accomplishing the leveling of the model and that is to move the reference plane and leave the projectors fixed. Such a solution for common phi can be seen in figure 5-29; a similar arrangement is used for common omega.

It was previously pointed out that the proper scale and azimuth were established in the same operation. There is also a relation between the leveling of a model and the scale. Such relationship is evident in figures 5-24a and b where it can be seen lack of leveling changes the projected distance A-B and hence the scale. These facts cause us to use a repetition process in the leveling, scaling, and azimuth determination of a model. What is first done is to approximately determine the scale and azimuth. Next the model is approximately leveled in common phi and common omega, which, of course, changes the scale. The process is thus repeated until the determination of scale, azimuth, and leveling are complete and within the required tolerances.

5.2.2.5.2.3. Summary. Since our theory of orientations is now complete let us summarize. It will be remembered that single aerial photographs were largely unsuitable for the production of our topographic maps because they did not readily produce the necessary third dimension. However, it was found that by altering the taking of the aerial photographs, to provide suitable forward overlap, there was a solution to the problem. Such solution lies in reprojecting the photographs and determining the three-dimensional position of points by intersecting corresponding rays. It was further found, however, that the determination of such intersection was dependent not only on the establishment of the same relation between the negative and the projecting lens (interior orientation) but also on the establishment of the same spatial position which each camera had at the moment of exposure (exterior orientation). The interior orientation, it will be remembered, consists of centering the negative so that the principal point falls over the optical axis of the projecting lens and establishing the same distance from the negative to the projection center. Finally, it was found that the most expedient way of establishing the exterior orientation was to first determine the relative position by matching projected images (relative orientation) and then to establish the proper absolute position of the formed model by leveling, scaling, and azimuth determination (absolute orientation). Once the model is properly formed and oriented then the necessary measurements can take place and the topographic map can be produced. Thus, we have the theoretical base to use photogrammetry to produce the measurements necessary for the surveying and mapping field. On the other hand, our task is far from finished. We must develop instruments and procedures for the stereocompilation. We will also develop methods by which we can use photogrammetry to, at least, partially determine the control points necessary for the absolute orientation.

5.2.2.6. Rectification.

5.2.2.6.1. Introduction. Before continuing in our basic investigation of the application of photogrammetry to the production of topographic maps, let us return for a moment to the possibility of using the single photograph. It is entirely possible that there may be cases in which the single aerial photograph could be of considerable value either singly or collectively in the form of mosaics. In such context it will be most probably desirable to remove at least part of the errors which have been previously found to be objectionable. It will be remembered that the main problem is centered around height differences which cause relief displacement, flying height differences which cause scale differences, and displacement due to the inclination of the camera at the moment of exposure. There are basically two solutions:

(1) to remove the effect of the tilt displacement and to establish a common scale for all photographs (this is called optical-mechanical rectification) or

(2) convert the photograph from a central projection to an orthogonal projection, which removes the relief displacement, tilt displacement, and establishes a constant scale (this is called the orthophoto).

5.2.2.6.2. Optical-Mechanical Rectification. In this form of rectification our task is to remove the tilt displacement and establish a common scale for all photographs. If we have flat terrain the only difference between our rectified photo and a planimetric map would be the amount of detail. In case there are height differences then there will be relief displacement in the photo which cannot be removed by this rectification process. For the time being we will assume flat terrain and later on discuss the matter of rectification in accidented terrain.

The basic concept of optical-mechanical rectification is to project the aerial photograph from its proper position and expose the projected images on a photograph medium. This, of course, requires the establishment of the correct interior and exterior orientation. The desired scale is determined by proper selection of the projection distance. It will furthermore be seen that the optical projection presents certain problems. Since the interior orientation and the projection distance do not present problems of a special nature, optical-mechanical rectification is basically a matter of solving the optical projection problems and the problem of the establishment of the proper exterior orientation.

The problems involved in the optical projection of the images revolve around an optical law, often called the Gauss Equation, that governs the formation of images passing through a

lens. If we let "a" be the object distance, "b" be the image distance, and "f" be the focal length then the Gauss Equation is:

$$1/a + 1/b = 1/f$$

The formation of an image can be seen in figure 5-30. If we apply this equation to the case of our aerial photography it can be seen that the flight height is the object distance, and the principal distance (c) of the aerial camera is the image distance. Using our regular notation the Gauss equation as applied to the aerial photograph would be:

$$1/Z + 1/c = 1/f$$

If we take the photograph from a flying height of 5000 meters with a camera whose principal distance is 152.00 mm then in order that the images be in focus the lens would have to have a focal length of 151.995 mm; in other words $1/5000 + 1/0.152 = 1/0.151995$. Another way of stating this case is that if we take an aerial photograph from an altitude of 5000 meters with a camera whose lens has a focal length of 151.995 mm, the images will be formed at a distance of 152.00 mm (see fig. 5-31). This same equation can be applied to the reprojection of the same aerial photograph. If, for example, we place the negative in the same aerial camera and provide a suitable light source we could reproject through the same camera lens. Now the distance from the negative to the lens (principal distance) is our object distance and our focal length remains the same so that it can be seen that the images would be formed at a distance equal to the flying height. This is logical since we are simply repeating in reverse direction the process which occurred when the photograph was taken. This points out one of the basic problems of photogrammetry. We cannot change the image distance in our reprojection since it is part of the interior orientation. On the other hand if we were to use a projector with the exact same characteristics as the aerial camera we will have a sharp image formed at a distance equal to the flying height. It was this same problem that confronted photogrammetrists in the early 1900s. There are of course various solutions to the problem, only one of which will be presented here. Said solution is to change the focal length of the projecting lens, which will not affect the geometry of our system, but which will at the same time afford the formation of a sharp image at a finite projection distance. For example, if we were to take our same aerial photograph whose c = 152.00 mm and project it through a lens whose focal length is 126.67 instead of 151.995, we find from the Gauss equation that the image will be formed at a distance of 760 mm. Thus, the solution to the optical problem is to choose a projecting lens whose focal length allows the formation of a sharp image at a finite distance.

In order to retain the geometry of our photographs it has been seen that we must establish the same interior and exterior orientation that existed in the moment of exposure. The interior

orientation, of course, presents no problem and is simply a matter of establishing the proper principal distance and of centering the negative. The exterior orientation must be established in a manner different to that used for the stereo model, because we are now working with only one photograph. Since our photograph is a two-dimensional object, and since in theory we are dealing with flat terrain, our problem is basically that of establishing the relationship between two planes. If our interior orientation has been established, this relationship can be determined if we have three known control points in the photograph. The procedure to be followed is merely to plot these control points on a template and then project the photograph onto this template. Our projector must have six possibilities of movement: a shift in X, Y, and Z, and three rotations around the X, Y, and Z axis. We therefore change the position of our projector, in accordance with these movements, until the projected images of our three control points coincide with their plotted images. If our ground is flat, we will then have established the proper exterior orientation and our task is finished.

A simple graphical experiment will now be performed which demonstrates this method of determining the exterior orientation. In figure 5-32 we see a cross section of an aerial photograph, made using a vertical plane through the nadir point. The points A, B, and C represent ground control points with the points a, b, and c being their respective photo images. If the interior orientation is recreated and the photograph projected from an arbitrary absolute position, we have a situation similar to that in figure 5-33. Here we see the same bundle of rays because we have used the same interior orientation; however, the rays do not pass through the control points. It can be seen, though, that the projector, with the photograph, has three degrees of movement, assuming that we do not disturb the interior orientation. These three degrees of movement, or degrees of freedom as they are often called, are a shift in X, a shift in Z, and a rotation phi. If the control points are kept fixed and the bundle of rays moved in these degrees of freedom, we can arrive at a position where the three projected rays pass through the three control points. It will be found that this position is, in fact, that which is contained in figure 5-32. The student should verify this by tracing the bundle of rays in figure 5-33 and changing its position until the rays pass through the control points. This example, of course, only involves two dimensions (X and Z). Our projected photograph will contain three dimensions (X, Y, and Z). However, the principle is the same. The three control points would be plotted to the desired scale on a control template and the projected photograph moved in the six degrees of freedom (a shift in X, Y, Z, and a rotation in ϕ, ω, and κ) until the photo images of the control points coincide with their plotted positions.

Thus, we have a way of establishing the interior and exterior orientation of our aerial photograph, so we now have a way to rectify our photographs. However, this method has certain disadvantages including the need for a different projector for each substantially different principal distance. For this reason the method just developed is further modified to make it more practical. Without going into detail it can be proven that by adding a fourth control point and further degrees of freedom we do not have to worry about the principal distance. This concept is valid only between two planes but since we are assuming flat terrain we are safe in using said concept. Thus, at the expense of an additional control point and a somewhat more complicated routine we can develop a method of rectification that will be valid for vertical photography of virtually any principal distance. In fact some rectifiers are constructed in such a way that they accept any principal distance and yet only require the three control points.

Let us examine briefly some of the common instruments used for rectification. We have seen from the Gauss equation that the focusing of projected images is dependent on the projection distance. It so happens that the magnification between the photograph and its projection is also dependent on the projection distance. Since it is desirable to have a variable magnification in a rectifier, one is faced with the problem of satisfying the Gauss equation for varying values of both the object distance and the image distance. As there are still further focusing problems connected with what is known as the Scheimpflug condition, the modern rectifier is so constructed so that mechanical devices automatically maintain the proper focus regardless of the magnification and position of the image and projection planes. These rectifiers will of course have the necessary degrees of freedom to be able to establish the exterior orientation as well as facilities for the exposure of the projected negative once its proper position has been determined. The result is that modern rectifiers are a combination of optical and mechanical, and in some cases electronic, components. They are, therefore, rather complex expensive instruments and normally allow an enlargement up to 6-7 times the photo scale. Examples of some of the better known rectifiers are those produced by the Bausch and Lomb Company, the Wild E-4, and the Zeiss SEG V. These three rectifiers cost between $15,000-$20,000. Examples of the Zeiss, and Bausch and Lomb rectifiers can be seen on pages 835 and 833 of the ASP Manual.

We have now arrived at the point where we can consider the actual stops involved in rectification by the optical-mechanical process. The first step, of course, is to obtain the aerial photography. The second step is to obtain the necessary ground control. This can be done by various methods which will be treated in more detail in section 5.2.2.8. For the time being it is to be assumed that we have at least four ground control points located in the corners of each

photograph. The coordinates of these points are plotted at the desired scale on a template; there is one template for each photograph. The next step is to put the aerial negative in the rectifier and project it on the control template. Next using the degrees of freedom, or movements, of the rectifier, the projected images of the control points are fit to their position on the control template. Finally the photographic medium is placed on the easel and the projected images are exposed. The result is rectified photos whose scale will be constant, for flat terrain of course, and which will not contain tilt displacement. Such photographs are then normally assembled into a mosaic. Because rectified photos are used, these mosaics are normally called controlled mosaics.

It will be remembered that all of our procedures have been based on the assumption that the terrain is flat. This, of course, will not be the case in practice, so that it behooves us to consider the effect of accidented terrain upon our rectification. The most obvious effect will be that our photo images will be affected by relief displacement. Since the images at the control points will be affected, just like any other point, it means that we will get an erroneous solution to our rectification. The amount of error will, of course, depend upon the amount of displacement due to relief. We may choose to accept this error, in which case our rectification is pretty much normal. Another possibility is to determine the height differences between the control points, compute the relief displacement, and displace the control points by this amount. We have in this manner removed the effect of this relief displacement on the control points and our rectification solution will be a correct one. In either case the rectification will not remove the effect of relief displacement, even though our rectification may be correct; we should remember that our motive in rectification is to remove the effects of tilt displacement and to reduce all photographs to a common scale. Thus, with optical-mechanical rectification in accidented terrain we must accept the errors introduced by relief displacement.

Let us suppose that we cannot accept the errors introduced by relief displacement as described in the preceding paragraph. Is there still a solution? The answer, of course, is the orthophoto. In optical-mechanical rectification we expose all projected points at the same elevation; for this reason we must assume flat terrain. Further still it is for this same reason that we cannot remove the effects of relief displacement. If we were to compile a stereo model, we first, in some way or another, determine the intersection of two rays and hence the spatial position of the point, and then orthogonally plot its position on our map sheet. We have already seen this in theory and will see how it is put into practice in section 5.2.2.9. We may use this same concept to produce our orthophoto, which is nothing more than a photo which has been reduced from a central projection to an orthogonal projection. In theory then what we should do

is to reproject the aerial photograph and then instead of exposing all points at the same elevation we should expose each image point at its proper elevation. It is obvious that only in the case of flat terrain will the two processes (optical-mechanical rectification and orthophotography) produce the same results. Since we are talking about the exposure of images at their correct height, their spatial position must therefore be determined. The result is that for the production of orthophotos we return to the stereo model. What is done is to establish the interior and exterior orientation of the stereo model and then theoretically we should expose each image point at its correct spacial position. Practically speaking this is rather difficult, so that some practical modifications must be made. One system, for example, divides the model into bands and exposes all points in a small area at the elevation of the central point. This can be seen in figure 5-34 where we see a stereo model divided into bands parallel to the Y axis. A small slit, whose length is equal to the band width, is run down the length of the band and kept constantly at the elevation of the center line of the band. Thus, in figure 5-34 all points falling on the shaded area in the slit would be exposed at the elevation of point P. The Zeiss company has recently developed an instrument which produces orthophotos by this method. There are other more sophisticated solutions as well that virtually expose all images at their correct height. In either case the end result is a photograph whose scale is constant and hence does not contain the effect of relief and tilt displacement.

5.2.2.7. Photogrammetric Instruments

5.2.2.7.1. Principles of Stereoscopy. Stereoscopic vision is not only an important part of our daily life, it is of fundamental importance in photogrammetry. For that reason some of the basic principles of the subject will now be considered. The logical place to start in a study of stereoscopic vision is how we ourselves see stereoscopically, often called natural stereoscopic vision. Stereoscopic vision is so commonplace with us that we tend to take it for granted; a simple test will illustrate this point. The reader should look at some objects at a distance of 10-20 meters first with one eye closed. It will be noted that one does not see in three dimensions, although previous knowledge and other factors such as apparent size, shadows, etc. may give some information as to the relative position of objects. If one then opens the other eye the relief will appear to jump out and depth perception should be quite apparent. We shall now try to explain this phenomenon and try to see how we see stereoscopically. The human eye is very similar to a camera. It has a lens, a diaphragm, and a surface where the image is formed; it is furthermore a central projection. When we observe an object with both eyes the line of sight of each eye is directed toward the object (convergence) and the eyes focus for the distance to the object (accommodation). The two sets of images, one for each eye, are formed

in the back of the eye and mentally fused to form the picture. Because the eyes are separated what we actually get is two slightly different pictures of the same object. If there is a difference in the distance to two or more objects there is a dissimilarity in the views received by each eye (see fig. 5-35). These dissimilarities, or parallaxes as they are often called, are coordinated by the brain and transformed into depth perception. If there is only monocular vision, of course, there will not be any dissimilar images of the objects and hence there will not be any stereoscopic vision. This is much the same situation as we encountered with one projected aerial photograph. With one photograph we cannot determine spatial position, only directions; however, with two successive overlapping photographs we can project and determine the spatial position of points through intersection. The same is true for stereo vision; when we use two eyes the convergence is a kind of intersection and allows us to see in three dimensions. There is a limit to the size of the difference of convergence that the eye can determine. Furthermore, it can be seen that as the distance to the objects increases, the angle of convergence (θ) decreases. It is for this reason that for objects at a distance greater than about 700 meters we do not have appreciable depth perception.

We can recreate stereo vision, often called artificial stereoscopic vision if we can establish conditions very similar to those which govern natural stereoscopic vision. To do so we must present to each eye a slightly different picture of the same area taken under certain conditions. The first condition is that the two photographs must be taken from two slightly different positions and in such a way that the optical axis of the camera in both positions lies approximately in the same plane. There must furthermore exist a reasonable ratio between the distance between successive exposure stations (the base) and the distance to the objects (height). This is often called the base/height ratio and should be in the neighborhood of 0.3 for stereoscopic vision. Finally the two photographs should be about the same scale. It will come as no surprise that vertical aerial photographs can be taken in such a way that they meet these conditions and can hence be viewed stereoscopically. The accomplishment of the 60% overlap between successive photographs means in fact that we photograph the same area from two slightly different positions. The way in which the vertical aerial photography is taken assures the proper base/height ratio and approximately the same scale. The problem then that exists is how to direct each eye to the appropriate photograph; there are a number of ways of accomplishing this. One way is to optically constrain, or direct, each eye to the proper photograph. This can be done using a stereoscope or binoculars which direct each eye to the desired area on the appropriate photograph. It can also be done using filters, either colored or of the polaroid type. In using the colored filters those of a red and green nature are most common. These colors are selected in such a way that the rays emitted from one will not pass through the other.

One photograph, of a model, is projected through a red filter and the other through a green filter. Similar glasses are then used to observe the images; the result is that the eye viewing through the red lens sees only the images projected through the red filter and the other eye only those of the green. The result is a three-dimensional black and white picture of the model. This system is used in a number of photogrammetric instruments. A system of polaroid filters and glasses can also be used, but is of a nature that it has not as yet been applied in photogrammetry. The reader who, a few years back, went to the three-dimensional movies will most probably have viewed the screen using polaroid glasses.

5.2.2.7.2. Floating Marks. We will shortly be engaged in developing instruments to enable us to produce topographic maps and carry out other surveys. Of prime importance in photogrammetric measurements are the measuring marks, often called floating marks. In figure 5-36 we see a diagram depicting the photographing of a steel needle. We see of course a very close resemblance to our situation of natural stereoscopic vision. We see we have central projection, with dissimilar images caused by a difference in distance (in this case height) to the two objects in question. Were we to view these two photographs with a stereoscope we would then get the impression of the needle AB rising above the flat ground. It can also be seen that the effect of height differences can be explained in another way, namely that they cause an image displacement in the photograph which we recognize as relief displacement. Since these displacements are, as we well know, related to height differences, by measuring them we can determine the height differences in question. This is, in fact, the basis for the measurements of the third dimension in photogrammetry. Thus, the relief displacement that plagues us in the single photograph actually is a very valuable thing. When we deal with displacements in a single photograph we talk in terms of relief displacement, however, when one is dealing with two photographs that form the model we use a displacement on two successive photographs and call it x parallax. By definition the x parallax of the point A is the x component of the distance n'a' plus the x component of the distance n''a'', where n is the photo nadir point and a' and a'' are the photo image points. Similarly the x parallax of point B is n'b' = n''b''. What is perhaps even more important is the difference in parallax between point A and point B is a'b' + a''b'', or the sum of the relief displacement of the height difference A-B on the two photographs. Since the relief displacement is a function of the height difference, it stands to reason that x parallaxes are equally a function and it is therefore x parallax differences that we measure in photogrammetry to determine heights. There is, of course, a y component to this parallax, but differences in y parallax are not caused by height differences but rather, as we have seen, by differences in relative orientation. The concept of the floating mark is directly

related to x parallaxes; in fact such marks are used to measure parallax differences. Referring once again to figure 5-36, if we were to take two small black measuring marks, place one at b' and the other at b'', and then view the photos stereoscopically, we would see them as one mark superimposed over point B in the three-dimensional space. This phenomenon can be explained in two ways: (1) the marks logically occupy the position of both images of point B in the two photographs or (2) there is no difference in x parallax between the two marks and the images of point B. If the marks are now moved in x direction to the positions m' and m'', and if we maintain our stereoscopic vision, it can be seen that the spatial position of the single fused mark will have changed and is now located at M. The explanation, of course, is simple. The marks at m' and m'' have a different value of x parallax then either point A or point B. If in our stereo model we observe the formed point M with respect to point A we see that M will appear to be floating above point A, hence the name floating marks. Now that the concept of x parallaxes and the floating marks has been established the concept of parallax measurements comes easily. If the marks were to be attached to a bar we could determine the total distance b'b'' when the marks have been placed at b' and b''. Furthermore, these marks can be placed stereoscopically at point B in the model rather than monocularly placing each image. Next the marks can be stereoscopically placed on the model point A by moving them in x direction. If again the bar is read, the total distance a'a'' can be read. Obviously b'b'' - a'a'' = the parallax difference between point B and A, and can be used to compute the height difference. It will be seen shortly that many instruments have mechanical devices to automatically convert x parallax differences to height differences.

5.2.2.7.3. General Principles of Instruments. If we now revert back to section 5.2.2.5.2 the reader will remember that the theoretical possibilities for the use of photogrammetry for three-dimensional measurements were developed. However, we lack the instrument to be able to carry out this theory. It is logical to first consider what we must expect this instrument to do. The first thing, of course, is that we must project, in some way or another, the respective images and then locate the intersection. Secondly it will be necessary to observe the model stereoscopically, and finally it will be necessary to measure and plot the position of the spatial intersection.

The consideration of the requirements of these three systems—the projection system, the observation system, and the measuring and plotting system—is the first logical step in the development of instrumentation. Under the projection system the basic task is, as has been mentioned, that of projecting images from corresponding photographs and effecting their intersection. It has previously been established that the orientation of the photographs in the sense

of the interior, relative, and absolute orientation must be established. This immediately places certain requirements on our instrument. For the interior orientation we must provide facilities for centering the photograph and for establishing the correct principal distance. For the relative orientation facilities must in one way or another be provided to change one or both projectors in accordance with the movements necessary to carry out the relative orientation. One such possibility would be to provide one projector with the possibility to rotate around three mutually perpendicular axes (phi, omega, and kappa) to displace it in Y, and Z. For the absolute orientation one projector must have the facility to be shifted in X (bx) and we will require means to change both projectors in phi (common phi) and omega (common omega). The projection system in general can be carried out by optical means, by mechanical means, or by a combination of the two usually referred to as optical-mechanical projection.

The task of the observation system is to provide means for viewing the model and the floating marks stereoscopically. One of the most important problems in development of the observation system is that of obtaining a sharp focus of the projected images. As with the projection system there are a number of general solutions to the observation system. Two solutions are observation of direct optical projection and direct observation of the photograph with a binocular microscope.

The basic task of the measuring and plotting system is to introduce the floating mark into the model, to provide a means for the orthogonal plotting of the X and Y position of the intersection, and to measure the Z coordinate. It may be further required that the X and Y coordinates be determined as well as plotted. There are not any general solutions to the problems of the measuring and plotting system as found in the projection and observation systems.

5.2.2.7.4. Types of Instruments. In practice the projection system and the observation system go hand in hand and fundamentally provide a system of classification by construction. This results in instruments for direct optical projection, mechanical projection, and optical-mechanical projection. This classification will be employed in the discussion of some representative instruments. In order to intelligently discuss these instruments, however, one must also consider the use to which they will be put. The first use is that of our main aim, namely the production of planimetric and topographic maps. These maps, of course, are produced in a wide range of scales, with the precision requirements being a function of the map scale. In this context the following definitions are made: large scale 0-24,999; medium scale 25,000-74,999; small scale 75,000 and smaller. It so happens that the larger the scale of map the greater the precision requirements. This is a tendency that reflects in the requirements of

the instruments as well and it will later be found that certain instruments are designed for large scale map compilation and others for medium scale and still others for small scale. There is yet another use for photogrammetric instruments and that is for the determination of the control points needed for the absolute orientation. This process is called aerial triangulation and will be discussed in section 5.2.2.8. In like manner the compilation of planimetric and topographic maps will be discussed in section 5.2.2.9; however, in this discussion of the types of the instruments brief mention will be made as to the normal use of the various classes of instruments.

5.2.2.7.4.1. Direct Optical Projection Instruments. In direct optical projection, the fundamental concept is just what the name implies. The photographs, usually printed on glass plates (diapositives), are placed in projectors and optically projected. A screen (tracing table) which is moveable in X, Y, and Z is provided to catch the projected images. The intersection is effected by placing the tracing table in an X, Y, and Z position such that the two projected images, one from each projector, fall one on top of the other. The observation is effected usually by the anaglyph system of using the colored filters. The focusing problem is solved by using a lens whose focal length permits the formation of the image at a finite distance, usually between 360 and 760 mm. The measuring of the height of the established intersection is done on the tracing table which contains the measuring marks and the plotting in X and Y is also accomplished by means of the tracing table. Normally with the direct optical projection plotters there are no facilities to measure X and Y coordinates.

One of the well-known examples of the direct optical projection plotters is the Multiplex. This instrument was originally developed by Zeiss Company of Germany and is now produced by a number of companies including the Bausch and Lomb Company of Rochester, New York. For reasons pertaining to the illumination and to produce a compact instrument the diapositive size and principal distance are reduced five times. The projection distance is 360 mm, which results in an overall enlargement ratio from the original photograph to the scale of the model of 2.5 times. Each projector has all five movements (phi, omega, kappa, by, and bz) for the relative orientation, and a bx for the scaling of the models. The common phi and common omega movements necessary to level the model are applied by moving the frame upon which the projectors are mounted. It will be recalled that this was discussed in section 5.2.2.5.2. The stereoscopic observation of the model is effected by the anaglyph of using red and green filters. The heights are read from a dial on the tracing table, the instrument does not have facilities to determine the X and Y coordinates of the intersection. There is, however, a pencil centered under the measuring mark in order to orthogonally plot the planimetric position of

points. The Multiplex unit consisting of two projectors costs about $4,000, and is generally used for medium and small scale compilation. A picture of the projectors can be seen on page 565 of the ASP Manual.

Another instrument of the direct optical projection class is the Balplex. This instrument is very similar to the Multiplex. It differs mainly in that its diapositive is reduced only two times, it has an improved lighting, and uses a projection distance of 400 mm, 525 mm, or 760 mm depending on the model. Unlike the Multiplex, which can only be used with wide-angle photography (c = 152 with a few mm's leeway), the Balplex will also accept super wide-angle photography (c = 88 mm) as well as wide angle; however, two different sets of projectors will be required. The cost of the Balplex two-projector unit is about $5,000 for wide-angle projectors and $7,300 for the super wide-angle projectors. It too, is used principally for medium and small scale compilation. A view of the Balplex plotter can be seen on page 571 of the ASP Manual.

Still another direct projection type instrument is the Kelsh Plotter. This instrument uses the same basic concepts as the Multiplex and the Balplex but differs in that it employs full-size diapositives; thus there is no reduction of principal distance. The projection distance is 760 mm. The projection of images and the formation of the intersection are carried out with the Kelsh Plotter in the same manner as with Multiplex and Balplex, namely optical projection with the intersection being captured on a screen. The Kelsh Plotter normally gives a five times enlargement between the scale of the photograph and the model scale; a reduction between the model scale and the plotting scale is further afforded by the use of a pantograph. The measuring and plotting facilities are the same as for the other direct optical projection instruments which have been discussed, namely the Multiplex and the Balplex. The main advantage of the Kelsh Plotter over the Multiplex and the Balplex is the fact that it affords the use of the full-size diapositives, allowing greater enlargement between the photo scale and that of compilation and a greater precision, particularly in heights. For this reason the Kelsh Plotter is largely used for large-scale mapping. It is normally available only for wide-angle photography (c = 152 mm), costs $10,000-$15,000, and is manufactured by a number of companies, predominately in the United States. Page 573 of the ASP Manual contains a view of one of the models of the Kelsh Plotter.

5.2.2.7.4.2. Mechanical Instruments. In the mechanical instruments the projection and the formation of the intersection of the rays is carried out mechanically with the rays being replaced by steel rods. The interior orientation is carried out by having the steel rod pivot

around a center point which is the mechanical determination of the projection center. This can be seen in figure 5-37. The principal distance is also determined mechanically as can be seen in this same figure. The intersection of corresponding rays is carried out, physically, using the same rods. An example of this concept is found in figure 5-38. Each steel rod is directed toward the respective photo image (p' and p'') with the intersection of the two rods forming the spatial position P. The relative and absolute orientations are facilitated by giving the necessary movements to the mechanical projectors. The position of the intersection can then be plotted in X and Y and measured in Z, as is done in the direct optical projection type instruments, or it can be measured in all three coordinates. The observation is carried out directly in diapositives by a binocular viewing system. The images p' and p'', along with the surrounding areas, are picked up optically and directed to the respective eye. It is in this optical system that the measuring marks are introduced as well. The result is a stereo view of the part of the model in question with the right and left floating mark indicating the point to which each steel rod is pointing. Thus, in figure 5-38 one mark would be superimposed on p' and the other on p''. The stereo vision would then show one fused mark resting on the ground at point P.

Mechanical instruments are produced by a number of companies, including Wild Company and the Kern Company, both of Switzerland, and the Galileo Company of Italy. Some representative examples of these instruments will now be described. The first instrument to be described is the Wild A8 Autograph; its mechanical construction can be seen on page 707 of the ASP Manual. It can be seen that the positions of the photo images p' and p'' are transferred mechanically to the points L' and L''; the rods then form the intersection to locate the model point P. It can further be seen that there is a small microscope that picks up the photo images and transfers them to the eyes via a series of lenses, mirrors, and prisms. In as much as the interior orientation is mechanical, the principal distance can easily be varied. This points out one of the advantages of the mechanical instruments, namely that photography representing a wide range of principal distance can be employed. The Wild A8, for example, can accept photography whose principal distance is between 100 and 230 mm. In contrast it will be remembered that with the direct optical projection instruments a change in principal distance outside of a few millimeters requires a new set of projectors. The A8 has only three elements of relative orientation (kappa, phi, omega) for each projector as well as the usual bx for scaling. For leveling there is a common movement (common phi) for leveling in phi with the leveling in omega being carried out with the individual elements. The A8 has both facilities for measuring the X, Y, and Z coordinates and for plotting the X and Y coordinates, the latter being accomplished by means of a coordinatograph. The cost of this instrument, which is designed for large-scale compilation, is about $35,000.

The Wild Company has also developed the model B8 which is designed for medium- and small-scale mapping. This instrument will accept photography whose principal distance is 88 mm (super wide angle), 115 mm, or 152 (wide angle). Its construction is similar to the A8, though much smaller and less complicated. It does not, for example, have facilities for reading X and Y coordinates, but can only plot them by means of a pantograph. It, of course, has the facilities to measure heights. The cost of this instrument is about $14,000. Page 716 of the ASP Manuel shows a view of this instrument.

The Wild Company further produces the A7 Autograph which is specifically designed for aerial triangulation. Its basic construction, of course, employs the mechanical projection and in many ways it is similar to the A8. The difference, with respect to the orientation, between the A7 and the A8 is that the A7 has all five elements of relative orientation (phi, kappa, omega, by, and bz) and uses these same elements for the absolute orientation. The basic difference between the Wild A7 and other similar instruments designed for aerial triangulation is that it has facilities to do what is called "change the base." As will be seen in section 5.2.2.9, aerial triangulation, when carried out on an instrument that has only two projectors, normally requires the facilities to use the outer halves, which is of course the normal case and can be seen in figure 5-39a. The use of the inner halves, called base out, requires a different position of the space rods as can be seen in figure 5-39b. The observation system as well as the measuring and plotting system of the A7 are very similar to those of the Wild A8. While the instrument is specifically designed for aerial triangulation, it can also be used for large-scale compilation. Its cost is around $65,000-$70,000. A picture of the instrument is found on page 705, ASP Manual.

The Kern PG-2 is also a mechanical instrument which is designed for the compilation of medium and small scale topographic maps. Its inner orientation is determined in a manner very similar to the Wild instruments and the PG-2 will accommodate vertical photography whose principal distance is within the range of 85 to 172 mm, thus wide angle and super wide angle photography. For the relative orientation only a total of five elements are available for both projectors with the leveling being done by changing the reference plane instead of the projectors. The normal model of the PG-2 has facilities only for measuring the Z coordinate with X and Y being determined by plotting the planimetric position by means of a pantograph. The observation system is that of binocular viewing of the diapositives. The instrument costs about $12,000. A picture of it is on page 720 of the ASP Manual. There are, of course, other mechanical instruments, but these will not be treated due to the limited scope of this course.

5.2.2.7.4.3. Optical-Mechanical Instruments. Instruments of the third class employ devices which are partly optical and partly mechanical in nature, hence the name optical mechanical. The basic concept is to use projectors which have the same characteristics as the taking camera. This not only includes the same principal distance but a projecting lens which is virtually identical to that of the aerial camera. Using these projectors, the diapositives are optically projected, and, since the projecting lens has the same focal length as the aerial camera, unless special means are taken the images would be formed (in focus) at a projection distance equal to the flying height. This problem is solved by placing an auxiliary lens system in the path of the bundle of rays, after they leave the projecting lens, which accomplishes the focus on a finite distance. The focused images are then carried to the eyes by means of a series of lens, mirrors, and prisms. The optical projection establishes the proper direction of the rays, assuming, of course, proper interior and exterior orientation. Then the direction of these rays is recreated and the intersection carried out mechanically.

In the true sense of the word the optical-mechanical instruments are in the present day, and particularly in Latin America, not too common. However, the Stereoplanigraph C8 produced by the Zeiss Company, while in the true sense of the word a direct optical projection type, will be treated in this section in order to differentiate it from anaglyphic instruments such as the Multiplex, Balplex, and Kelsh. On page 650 of the ASP Manual the basic scheme of the C8 can be seen. The auxiliary focusing system consisting of two small lenses can be seen just below the round projection lens. Further below can be seen a mirror, upon which the images are focused and which also directs the images to the eyes via an optical train. These two mirrors also serve to form the intersection (they are separated only for constructional reasons) and it is for this reason that the Stereoplanigraph C8 is actually of the direct optical projection type. The C8, like the Wild A7, is specifically designed for aerial triangulation and hence has the facilities to change the base. It further has the capability to determine X, Y, and Z coordinates and to plot X and Y positions by means of the coordinatograph. The instrument, of course, can be used for compilation as well as aerial triangulation; its cost is about $75,000. Page 652, ASP Manual shows a view of the instrument.

5.2.2.7.4.4. Approximate Instruments. There is yet another class of instruments which has not been mentioned, mainly because their construction does not recreate the bundle of rays as in the three types just discussed. This class of instruments, often called approximate instruments, uses the pair of aerial photographs positioned on a flat surface and from this position produces the topographic map by tracing the planimetry and determining the altimetry by measuring x parallaxes. Of course if one does this straight away the errors in the

map would be considerable. In fact, it will be remembered that back in section 5.2.2.3 we arrived at the conclusion that direct tracing from a single aerial photograph was not recommendable. In planimetry this would introduce errors due to relief displacement, tilt displacement, and variance of scale. The lack of relative orientation introduces further x parallaxes which cause errors in the height measurement. This phenomenon is often called model deformation. The approximate instruments have mechanical devices which in one form or another correct for some or all of the errors introduced by the lack of interior and exterior orientation. One instrument of this class is the Zeiss Stereotope (see page 549, ASP Manual) which has been developed for small-scale mapping and low-precision engineering projects. The only orientation carried out in this instrument is with respect to kappa. The planimetry is traced from one of the two photographs. The height measurements and resultant contours are determined from x parallax measurements. A series of mechanical devices correct for errors due to relief and tilt displacement and model deformation. The instrument costs about $3,500.

5.2.2.7.5. Analyses of Instruments by Use. Now that the types of instruments and representative instruments have been presented using the classification by construction it will be helpful to reclassify said instruments by use and to make some comparisons and analysis of the various instruments. If instruments are classified by the use for which they are principally designed, they may be grouped into the following classifications: third order instruments designed for small-scale mapping and revision, topographic plotters designed for medium- and small-scale mapping, precision plotters designed for large-scale mapping, and universal instruments designed principally for aerial triangulation.

In the third order instruments we find among others the Zeiss Stereotope and the Santoni Stereomicrometer. These instruments offer the advantages that they are inexpensive and easy to operate. The disadvantages are a low precision and their lack of flexibility for various uses. The main use of the third order instruments is in map revision work, small-scale, low-precision engineering projects, and very small-scale topographic mapping (i.e., smaller than 1/100.00).

The topographic plotters include the Multiplex, Balplex, the Wild B-8, the Wild B-9, the Kern PG-2, and others. This class of instrument is designed for medium- and small-scale mapping (i.e., 1/25,000 to 1/250,000). They offer the advantages of fairly low price, reasonable precision, and good production.

The precision plotters are those designed for large-scale mapping and high-precision engineering projects. The necessary high precision of these instruments results in a high cost. Instruments included in this class are the Wild A8 and the Santoni Stereosimplex III. The Kelsh Plotter is often included in this class but it actually falls somewhere in between the

topographic plotters and the precision plotters. The Kelsh Plotter is normally used both in large-scale mapping and medium-scale mapping.

The universal instruments by comparison can do the same task as the precision plotters plus aerial triangulation. They necessarily will cost considerably more than the precision plotters and except in special cases are used most generally for aerial triangulation. With the exclusion of the Kelsh Plotter, the precision of the precision plotters and the universal instruments are the same. To demonstrate their precision, they will measure X and Y coordinates to a precision of about 0.02 mm, which in a photograph of scale 1/10,000 amounts to 0.20 meters in the ground. The universal instruments include the Wild A-7, Zeiss C-8, Santoni IV and V, and others.

We have now completed the development of the instruments that we need for both our compilation and aerial triangulation. It was first shown that these instruments generally contain a projection system, a measuring and plotting system, and an observation system. If the instruments are classified by construction one arrives at the following types: direct optical projection, mechanical, optical-mechanical, and approximate. If they are classified by use one finds third order instruments, topographic plotters, precision plotters, and universal instruments.

5.2.2.8. Determination of Control

5.2.2.8.1. Introduction. The theory and the instruments for the production of our maps and surveys have been developed, the only thing we lack now is the control needed for absolute orientation. It will be remembered that at least two points of known horizontal position and three points of known vertical position are required for each model. Furthermore, it should be obvious that these points will have to appear in the photograph, in other words they will have to be photo-identified. Such control points, of course, can be obtained by field survey methods; however, there is also the possibility of using a combination of field survey methods and photogrammetry.

5.2.2.8.2. Field Survey Methods. There are a number of ways to establish horizontal and vertical control by field survey methods. In horizontal control the initial points are established by a combination of astronomic observations and distance measurements. The control is then extended by obtaining the distance and azimuth to new points. These new points may then be used to further extend the control. The obtaining of the azimuth and distance may be done by choosing the new points in such a way that they form a triangle or quadrilateral with other known points. The angles are then measured, and using spherical trigonometry the

required distances may then be computed. This process is called triangulation. Another way is to measure the distance, either by tape or by electronic means, as well as the bearing or azimuth. This process is called traversing. The vertical control is obtained by starting at sea level and measuring the height differences with an instrument called a level, this process is called leveling. The height difference can also be determined by measuring the vertical angle and the distance to an object, by barometers, and by other means. The choice of which of these various methods to use, both in the establishment of horizontal and vertical control, depends on various factors including the required precision, type of terrain, available resources, etc. However, in many cases field survey methods are long and arduous. Furthermore, we have found that photogrammetry can be used effectively in making the necessary measurements for the production of topographic maps; thus there is a strong possibility that it may be suitable for use in the determination of our control points.

5.2.2.8.3. Photogrammetric Methods of Determining Control. First let us consider the possible advantages of the use of photogrammetry for the determination of the control necessary for absolute orientation. One of the first that comes to mind is that of ease of operation. It is certainly far easier to do our surveying in the office than to struggle through mountainous country under adverse weather conditions. It may, furthermore, be a lot faster to use photogrammetry and it could be more economic. As an example of the advantages of the use of photogrammetry let us take a flight of photography consisting of 30 models. This photography will require 62 control points if four horizontal and vertical control points are established in each model. This is more than the minimum number necessary, but is advisable in order to have checks upon the work. Sixty-two control points by field survey methods could take as long as several months to establish and in any case represent a considerable amount of work.

Let us now see if we can establish all or part of the 62 control points through the use of photogrammetry. In figure 5-40 a portion of our flight of 30 models can be seen. Let us assume that sufficient control exists to carry out the absolute orientation of model A. Upon completion of the absolute orientation of this model we will have succeeded in establishing the proper exterior orientation of both the first and the second photograph. Further analysis of the situation, however, shows that the second photograph plays a role in the formation of model B. Consider, for example, the effect of adding a projector containing the third photograph and carrying out a relative orientation of model B using only elements of this third projector. The result of this operation would be the establishment of the proper exterior orientation of the third photograph. This must be true because the relative orientation establishes the correct relative position between photograph 2 and photograph 3. Furthermore, since photograph 2 is

already in the correct position this means that photograph 3 is now in the correct position. It is of course very important that during this operation the projector containing photograph 2 not be touched. It is for this reason that instruments designed for aerial triangulation provide all five elements of relative orientation for each projector. Once the relative orientation of the photograph of the third projector is complete it may be necessary to adjust the scale of model B by altering the bx, so that it is the same as the scale of model A. However, this can be done using points common to the two models. It should now be obvious that we have carried out the absolute orientation of model B without any additional control points. We can, furthermore, continue and add model C in the same manner. Thus, it might seem that the entire strip of 30 models could be oriented using control only in the first model. It is not quite that simple, as there is an accumulation of errors throughout the strip as well as the effect of earth curvature and other factors. The result is that a few control points scattered throughout the strip are necessary. However, even if we need 10 or 12 field survey points in order to establish the absolute orientation of 30 models by aerial triangulation, this will in many cases represent a large saving in time and money. The process of extending the absolute orientation by the process just described is often termed "bridging," because the absolute orientation of an entire strip may be determined and based on a few control points.

Now that the basic concept of aerial triangulation has been developed, we can see how it might be done in practice. One of the easiest ways would seem to be to have a long bar upon which one could keep adding projectors. This was in fact the purpose for which Multiplex was originally developed. The normal procedure is to have ground control in the beginning, middle, and end of the strip. The relative and absolute orientation of the first model is carried out. The second and succeeding models are added by means of the relative orientation and by scaling to the previous model. This process is continued until control points are reached. Then the entire strip is adjusted as a whole to fit not only the new control but the previous control as well. The work is continued in this manner until all the models have been added and the entire strip is fit to an aggregate of all the control points. Although as many as 20 projectors have been used, the normal is about six to eight. The Balplex may be used in the same manner as the Multiplex. Once the entire strip has been completed, there are two choices: (a) all the models may be compiled or (b) the position of four points per model may be pricked on the plotting sheet and the heights read; each model can then be oriented individually. The latter of these two possibilities is most commonly used as it allows more people to work simultaneously. Although aerial triangulation with Multiplex or Balplex is a step in the right direction, it still leaves something to be desired. In the first place it requires a lot of control and has a fairly

low precision. Furthermore it is largely a graphical process which is fairly tedious. For this reason aerial triangulation using Multiplex is not widely used at the present time.

The next step in this investigation should be to improve on the procedures that have been developed for use with Multiplex. The logical approach would be to develop an instrument which has the facilities to determine X, Y, and Z coordinates and which has a higher precision than Multiplex. This can be done provided that we are willing to limit ourselves to only two projectors. Thus, we are faced with the problem of carrying out aerial triangulation with an instrument having only two projectors and at the same time maintaining the concept of one projector remaining fixed as required by the bridging concept. We are aided in the solution by the fact that we can determine X, Y, and Z coordinates. It means, in fact, that once we have established a model we may read the coordinates of the supplementary control points, which are called pass points, as well as points for the connection of the models. Once this has been done we are free to destroy the model. Thus, once the first model is completed, we remove photograph 1 and insert photograph 3. This leads to a situation shown in figure 5-41. We have in the instrument the two diapositives which correspond to the second model, but due to the fact that photograph 2 cannot be touched we have a situation where the common overlap is inside. It is for this reason that the instrument must have the facility to change the position of the space rods, which we now know as changing the base. Therefore by changing the instrument so that the inner portion of the two diapositives can be used, the model can be formed and the orientation can be carried out using the lefthand projector, leaving the righthand projector which contains photograph 2 fixed. Of course, in doing so we destroy model 1, but if we have read the coordinates of the pass points and the connection points this will be of no great concern. When the second model is completed the fourth photograph is placed in the righthand projector and the space rods are placed in the normal (base in) position. This then is the manner in which we may carry out aerial triangulation in an instrument that has only two projectors.

We may now take a look at the actual procedure used. The first model is oriented relatively and absolutely. When this is finished the coordinates of the pass points are read, followed by the coordinates of points which are used to connect the following model. The lefthand diapositive is then removed and the third diapositive inserted. The base is changed to position "base out" and the relative orientation of the second model carried out using, of course, the elements of the lefthand projector which contains photograph 3. Next the bx is altered until the two models have the same scale and finally the two models are connected so they have the same coordinate system. In the same manner the entire flight of photographs is triangulated. The adjustment of the strip to the control points cannot be done in the same manner as with the Multiplex because as soon as a model is finished it is destroyed; the only record left is coordinates of the

pass points and connection points. The adjustment is therefore done by numerical means. In as much as the universal instrument will produce the strip coordinates expressed in millimeters in an arbitrary coordinate system they must be converted to the ground coordinate system. This transformation is combined with the removal of the accumulated errors, or adjustment as it is called. These strips may be adjusted independently to the control points, which is called strip adjustment, or they may be adjusted simultaneously to control points and to adjacent strips called "block adjustment."

Thus it is that a method of control determination by means of photogrammetry has been developed. It will still require a small amount of control determined by field survey methods but in most cases offers a marked advantage over the exclusive use of field survey methods. For example, if we were to aerially triangulate the strip of 30 models with a universal instrument and strip adjustment at a precision suitable for a map of 1/50,000, about eight ground control points would be required. These eight points can be contrasted with the 62 points that would be needed to establish the control without the benefit of photogrammetry. Assuming aerial photography of 1/60,000, the coordinates of the pass points will be determined to a precision of about 3 meters.

In aerial triangulation we have been concerned with three dimensions; however, there are many cases where only the X and Y coordinates for the pass points are needed. One example of this would be the control needed for rectification, while another case would be the situation in which it is feasible to obtain the vertical control by field survey means and the horizontal control by photogrammetric means. Since we are now concerned with just two dimensions we may return to the concept of the single photograph and the errors it contains. Let us take the relief displacement and see if there is something more that can be done other than what we have considered. In figure 5-42 we see a diagram depicting a height difference in the ground and its subsequent projection onto the aerial negative and the aerial positive or photograph. Below we also see a frontal view of the photographs with the relief displacement caused by the height difference A-B. It can be seen that b is displaced along a line from the photo nadir point passing through a. This means that the direction of the line passing through the nadir points and the photo images "a" and "b" is unaffected by this relief displacement. The same is true for any line which passes through the photo nadir point. Thus by using the line connecting the two principal points we can determine the correct position of A by intersection. This is shown in figure 5-43. This same concept can be applied to the tilt displacement as well, in such case, however, the displacement is not radial from the nadir point but from the isocenter. However, for vertical photography the isocenter, the nadir point, and the principal point fall

approximately in the same place. The result is that this process of determining points by intersection, called radial triangulation, normally uses the principal point as the radial center. The process can be performed graphically but is normally done by mechanical means. In such case the radial lines are replaced by slots, hence the name slotted templates. One template is made for each photograph with studs being placed at the intersection of the slots, in the layout. This layout includes a few control points and is carried out in such a way that the assembly can be adjusted to fit to the control. When the assembly is complete the positions of the pass points are pricked on the base sheet and the subsequent coordinates scaled off.

Thus it can be seen that photogrammetry can be used to a large extent to determine its own control points. If the X, Y, Z coordinates are required, then aerial triangulation can be used. In such a case the desired pass points are marked in the photographs and the aerial triangulation carried out. This aerial triangulation will produce coordinates for these pass points expressed in millimeters in an arbitrary coordinate system. A transformation and an adjustment is then carried out which transforms the coordinates into the ground coordinate system and removes most of the accumulated errors. If only the X and Y coordinates are required at a lower precision, then radial triangulation may be used. In such a case one template is made for each photograph and slots cut along the radial lines to the pass points, control points, and principal points of other photographs. These templates are then laid in an assembly, including control points, and the desired pass points are obtained by intersection. The photogrammetric determination of control points offers the advantages of speed, ease of operation, and in many cases economy.

5.2.2.9. Map Production

5.2.2.9.1. Introduction. The investigation which we have been carrying out has now produced all the elements essential to the photogrammetric production of maps and other surveys. What we now need is to assemble this information, which includes basic theory, instrumentation, and the determination of control, in order to establish the overall procedure to be followed. One of the foundations of any such system is the specifications that govern the final published map. For example, in the United States the following specifications apply to maps whose scale is smaller than 1/20,000: 90% of all well-defined map features may not have an error larger than 0.5 mm at map scale and ninety percent of all elevations interpolated from contours must be within one-half the contour interval. Similar specifications govern maps whose scale is larger than 1/20,000. All of the procedures, materials, and equipment used must be in accord with, and insure compliance to, these specifications.

5.2.2.9.2. Production of Planimetric and Topographic Maps. As was mentioned in the preceding paragraph, we will now proceed to discuss the step-by-step guidelines to be followed in the photogrammetric production of topographic and planimetric maps. Such steps will include a consideration of basic control, aerial photography, field classification, supplemental control, determination of photogrammetric control, restitution, field checks, and reproduction.

The establishment of basic horizontal and vertical control is the first step in the production of maps. This is done by a variety of methods, some of which were described in section 5.2.2.8.2. The result of this phase will be a series of monumented points whose X, Y, and Z coordinates have been determined to a very high degree of accuracy. It is unlikely that the photogrammetrist will have a great deal to do with this phase, though he should certainly have a broad knowledge of it.

Once the basic control has been established the aerial photography can be planned and attained. In its planning one must take into consideration the type of photography (vertical or oblique), the type of camera (normal angle, wide angle, or super wide angle), the desired photo scale, and the flying height. Once these factors are determined the flights are planned and the photography taken. This resultant photography is then evaluated and accepted or rejected. Any rejected photography, of course, must be reflown.

The next step is a field phase in which the field classification and the supplemental control are carried out. The field classification consists of annotating on the aerial photographs all details that will appear on the final map. This includes roads, houses, streams, vegetation, names, and many other features. The supplemental control is the establishment of the additional field control needed. It may consist of determining the few points upon which the aerial triangulation or radial triangulation will be based, or it may consist of establishing field control for every model. The methods used are very similar to those used for basic control and of course include the photo-identification of each point established.

If photogrammetry is used in the determination of the control, then this operation will follow the field phase. The control needed for the aerial triangulation or radial triangulation will have been planned previous to the field phase, with the number of points and their approximate location being indicated on the photographs. Once the supplemental control is finished, then the pass points are chosen and marked in the photographs. In the case of aerial triangulation the instrumental work is carried out, followed by the adjustment phase; the final result is X, Y, and Z coordinates for all pass points. If radial triangulation is carried out, the templates are made, the control plotted, the assembly made, and the final X, and Y coordinates obtained.

One of the first steps in the restitution is the choice of plotting instrument. Here we will have to choose first between the use of third order instruments, topographic plotters, and precision plotters; and secondly choose the specific instrument within these groups. The first actual step is to plot the X and Y coordinates of the points; this is followed by the relative and the absolute orientation of the model. Once this is complete the planimetry and the altimetry can be compiled. When the operator feels the model is complete, it will be checked by a supervisor and if accepted he goes on to another model. The normal procedure is to further check the compilation through the use of the photographs and other aids. This procedure is often called compilation edit. Once the entire map sheet is compiled and checked the various sections are photographically reduced to the final publication scale and joined together to form the entire sheet.

A copy of the compilation manuscript, reduced to final publication scale, is then sent to the field for the field check. In this phase the completeness is checked and any missing details added. In many cases the precision of the map is checked as well.

The final phase of the production of the map is the reproduction. The first step is to photographically transfer the reduced compilation manuscript to a plastic medium and then scribe all the detail. This work then is checked and if found to be correct and complete, the printing plates are made and the final map printed.

Thus, the entire process in the production of topographic maps is complete. It should be emphasized that there are many variations to the procedures described here and these are merely a representative group. For example many map sheets are inked instead of scribed, and instead of printed copies ozalid copies may be used.

5.2.3. PRESENT STATE OF THE USE OF PHOTOGRAMMETRY IN THE CARTOGRAPHIC FIELD. It should be obvious that photogrammetry need not be restricted to the production of planimetric and topographic maps. It is widely used in highway surveys, cadastral surveys, and in many other measurement situations within the cartographic field. The fact is that the use of photogrammetry is being developed to such an extent that it is becoming a highly complex but extremely valuable tool. It is for this reason that advanced training and education within the field of photogrammetry is not only becoming very necessary but in many cases almost mandatory.

5.3. NON-CARTOGRAPHIC APPLICATIONS OF PHOTOGRAMMETRY

5.3.1. GENERAL PRINCIPLES. The use of photogrammetry for applications in the non-cartographic field, often called "close-up photogrammetry," is growing rapidly. The basic principles are very similar to those that we have been studying and in some aspects the problem is quite a bit simpler than that posed by the cartographic field. Once again two photographs are taken of the same area or subject from two different positions. Then by projecting these photographs, either optically or mechanically, under conditions including the construction of the exterior orientation, the position of image points is determined by intersection. Since one is now dealing with terrestrial photographs, it is often feasible to measure some of the elements of orientation. For example many of the bases are measured or determined by using two cameras mounted on a fixed base. Similarly other elements may be measured. The restitution is generally carried out in instruments specifically designed for the purpose. Examples of these are the Wild A-40 and the Zeiss Terragraph. Close-up photogrammetry facilitates difficult measurements, offers a permanent record and the means for repeated measurements, and further offers the advantage that the objects measured are not touched. The following are just a few of the many uses to which close-up photogrammetry has been put:

a. Posture studies of human beings.
b. Preparation of graphics depicting historical buildings.
c. Determination of measurements for the preparation of suits and dresses.
d. Measurements of deformation caused by the loading of beams and other objects.
e. Determination of the position of celestial bodies.
f. Many applications in the medical field including the use of photographs taken internally.
g. The study of surfaces by taking microscopic photographs.
h. Missile tracking through the use of the ballistic camera.
i. The determination of quantities in open pit mining, in coal piles, etc.
j. Traffic accident analysis.
k. Calibration of antennas.

5.4. SOURCES OF INFORMATION

It is quite likely that the reader may desire more detailed information on the field of photogrammetry and the various applications. Sources of information in this field pose somewhat of a problem principally because, though relatively plentiful, they are found widespread throughout books, technical manuals, technical magazines, and other periodicals. Publications covering all aspects of the field do not exist and therefore if one desires to have a good collection

of source material an extensive reference library is required. There may further be a problem of language, as most of the important publications are found in English, German, or French. I know of only one attempt to catalogue the various photogrammetric publications and that is the International Bibliography of Photogrammetry published by the International Training Center (ITC), Delft, Netherlands. The publications are briefly described on 3 × 5 inch (7 × 12 cm) cards and are grouped according to author and subject matter. The ITC is working continuously on this bibliography and currently has approximately 3000 publications catalogued. This bibliography can be purchased at a cost of about $10 per 500 titles. These cards are very helpful but unless one has access to a large number of publications the merit of a subscription would be doubtful.

A partial list of some of the more important sources of information will now be given. Unless otherwise noted the publication is available only in English and for the most part it will be limited to the field of photogrammetry.

BOOKS

1. Basic Metrical Photogrammetry by D. Lyon, P. O. Box 492, St. Louis, Missouri, $4.00.
2. Manual of Photogrammetry, Third Edition American Society of Photogrammetry, $19.00.
3. Manual of Photographic Interpretation, American Society of Photogrammetry, $12.00.
4. Outline of Photogrammetry by Schwidefsky, Pitman Co., New York, $9.00 Spanish Edition available.
5. Photogrammetry by Hallert, McGraw-Hill, New York, 1960, $11.00.
6. Survey Adjustments and Least Squares by Rainsford, Frederick Unger Publishing Company, New York, $9.50.
7. Text Book of Photogrammetry by Zeller, H. K. Lewis and Company, London, 1952.

There are of course many more books.

U.S. GOVERNMENT TECHNICAL MANUALS

There are a large variety of U S. Government publications. If the student is interested in obtaining these he should write to the U. S. Government Printing Office, Washington, D. C. and request their list of publications in the field of Surveying and Mapping.

REGULAR TECHNICAL MAGAZINES

1. Photogrammetric Engineering, published by the American Society of Photogrammetry. This publication is issued 12 times a year and covers a wide range of subjects in the field of photogrammetry and photo-interpretation. The cost is $15.00 per year, including membership in the society. Write to the American Society of Photogrammetry, 6269 Leesburg Pike, Falls Church, Virginia.

2. Photogrammetria, published by the International Society of Photogrammetry. This publication is rather technical in nature and is devoted more to the field of photogrammetry than to photo-interpretation. It is published six times a year; the majority of the articles are in English, however some appear in German or French. The publication can be obtained, at a cost of $6.75 per year, from Elsevier Publishing Company, P. O. Box 211, Amsterdam, Netherlands.

3. The Photogrammetric Record, published by the Photogrammetric Society, London, England. This publication is devoted mostly to the field of photogrammetry and is published twice a year. The price of the publication if $3.70 per year and it can be obtained from the Photogrammetric Society, 24 Bruton Street, London, VI, England.

4. Surveying and Mapping, published by the American Congress of Surveying and Mapping. This publication is primarily devoted to field surveys but also includes some articles on mapping. It is published four times a year, costs $8.00 and can be obtained by writing to the American Congress on Surveying and Mapping, Room 430, Woodward Bldg., 733, 15th Street, N.W., Washington, D. C., 20005.

5. The Canadian Surveyor, published by the Canadian Institute of Surveying. This publication is devoted to both field surveys and photogrammetry and is published five times a year. It can be obtained by writing to the Canadian Surveyor, P. O. Box 3151, Postal Station C., Ottawa, Canada.

There are a number of other publications in various languages as well.

IRREGULAR PUBLICATIONS

1. Publications of the International Training Center for Aerial Surveys (ITC). The ITC has a series of publications which are put out at irregular intervals. These publications are produced in two series; Series A devoted to Photogrammetry and Series B devoted to photo-interpretation. To date more than thirty publications have been produced in both series; however, some of the early publications are out of print. A subscription to either or both series can be obtained, at a cost of about $6.00 per twelve issues, by writing to the ITC at 3 KANAALWEG, Delft, Netherlands. Copies of the ITC lecture notes, both in photogrammetry, and photo-interpretation are also available in English and French at a small cost. Some of these lecture notes are in mimeograph form while a few have been prepared in a hardback edition.

2. International Archives of Photogrammetry. This is a series of reports and technical papers covering the International Congress of Photogrammetry held every four years. The

publication contains many technical papers, published in various languages, as well as the proceedings of the congress and reports on mapping in various countries of the world. This publication is usually produced in the city where the congress was held, however, many of them can be obtained by writing to N. V. Uitgeverij "ARGUS," the Hague Netherlands. The costs vary but are usually around $20-$25.

There are also a number of technical agencies which produce technical articles and reports from time to time. These publications are usually furnished free of charge, or at a nominal fee, and in some cases it is possible to be put on their mailing list. Some such organizations are:

The National Research Council of Canada
Ottawa, Canada

The United States Coast and Geodetic Survey
Washington 25, D. C.

The U. S. Geological Survey
Washington 25, D. C.

The U. S. Army Map Service
Washington, D. C.

The Pan American Institute of Geography and History

There are a number of photogrammetric societies throughout the world. Most of these societies put out a publication in the language of the country and this can also be a source of information. These societies also have activities other than publications; there are two societies whose activities merit further exploration. The American Society of Photogrammetry sponsors an annual Photogrammetric Congress held in Washington, D. C. in March of each year. The International Society of Photogrammetry sponsors a Photogrammetric Congress which is held in various cities every four years. For example, it was held in Washington, D. C. in 1952, in Stockholm in 1956, in London and 1960 and in Lisbon in 1964. These Congresses mainly consist of the presentation of technical papers and present equipment displays as well.

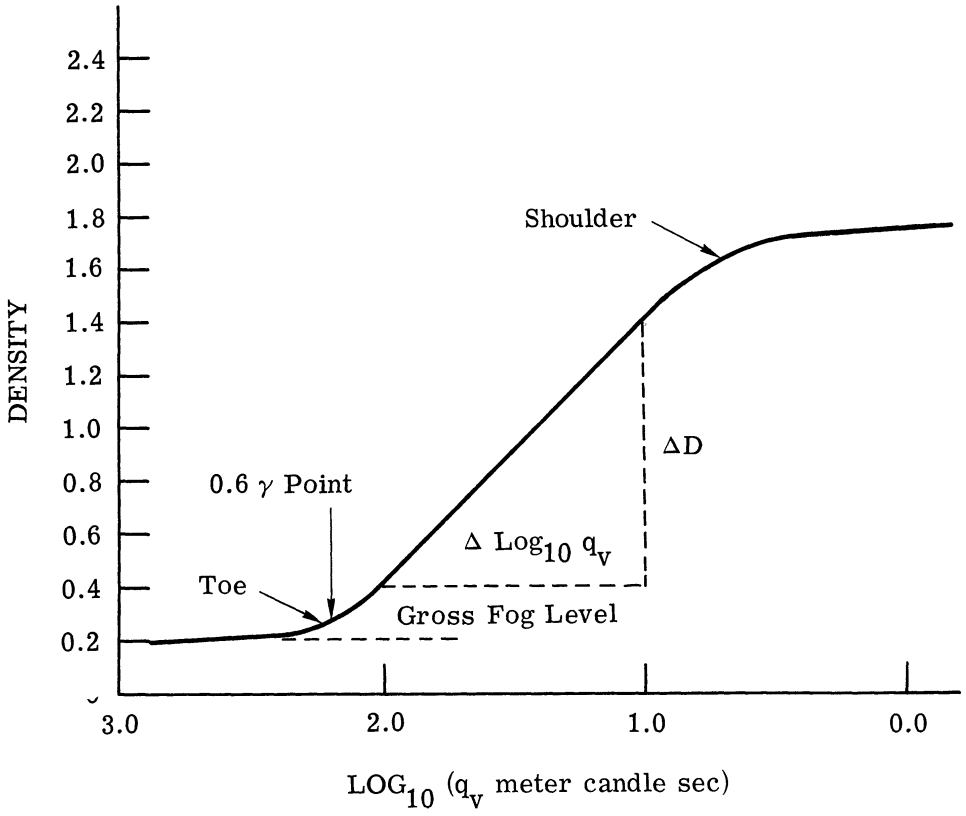


FIGURE 2-1. CHARACTERISTIC CURVE OF HYPOTHETICAL FILM

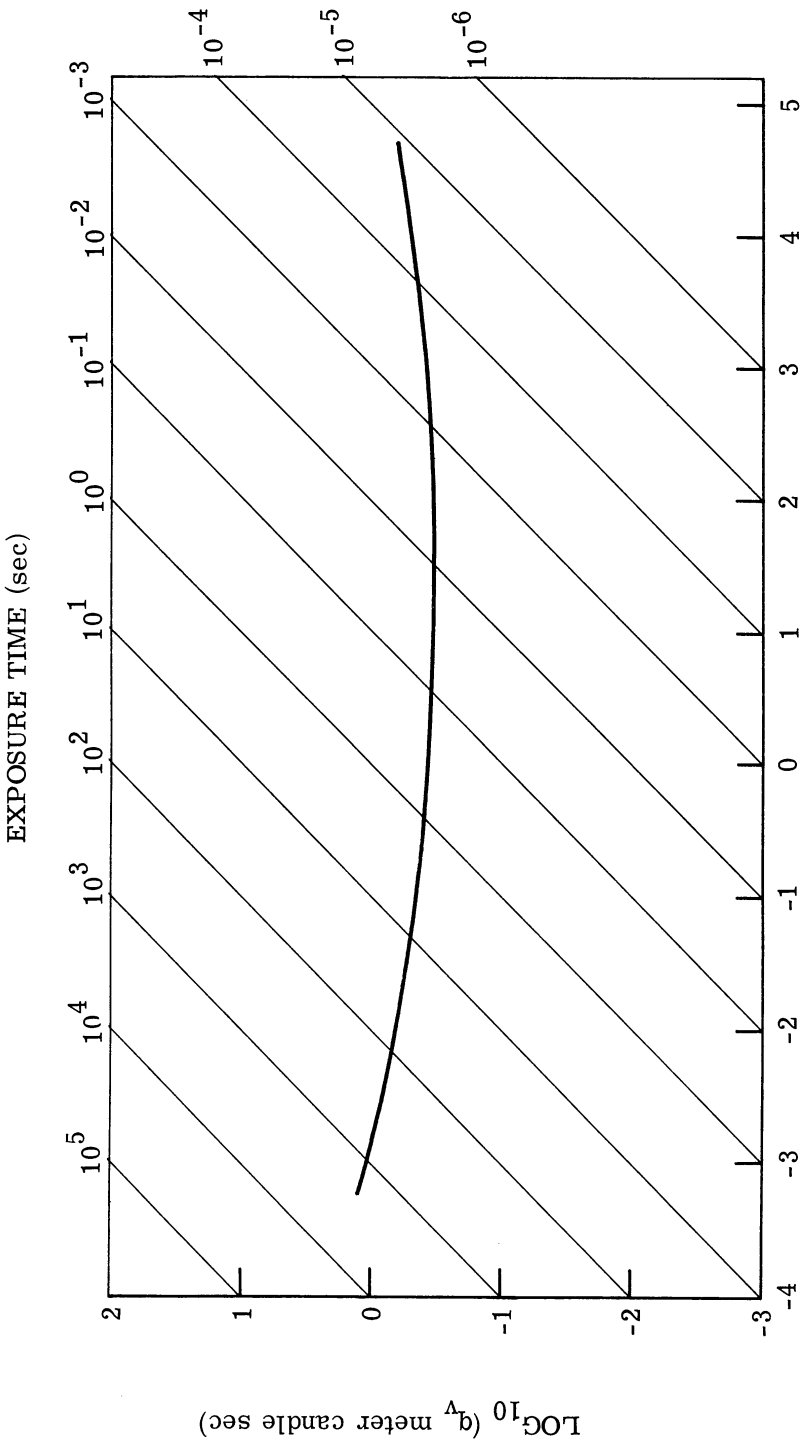


FIGURE 2-2. RECIPROCITY CHARACTERISTIC OF HYPOTHETICAL FILM

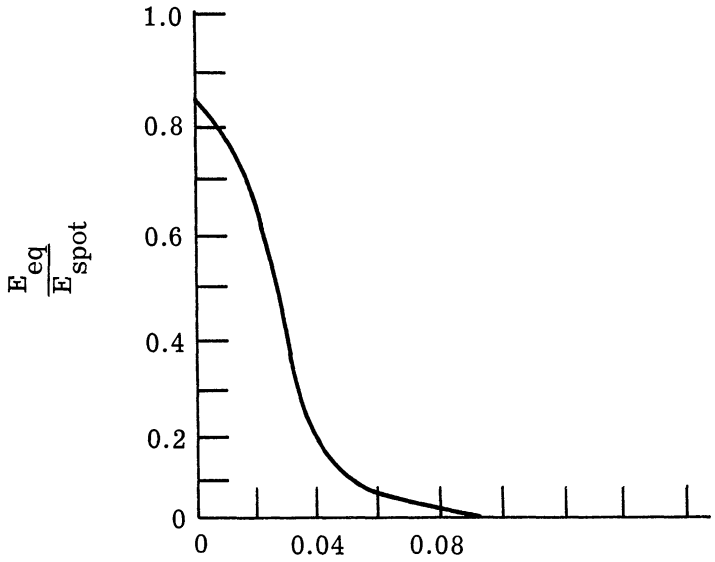
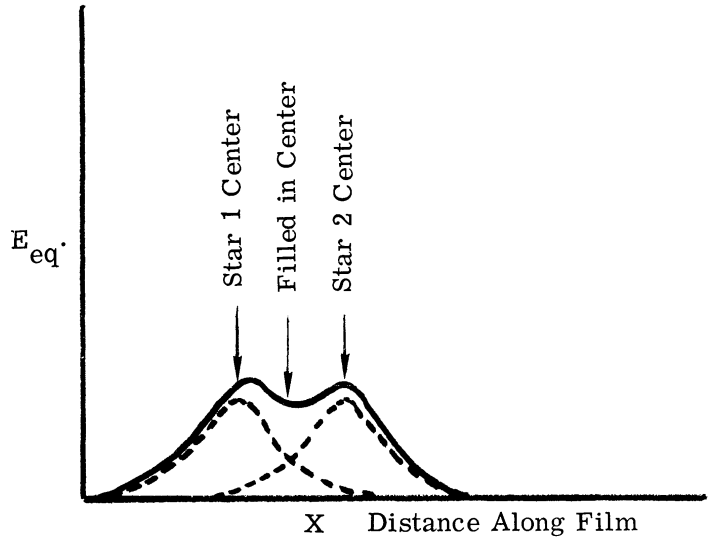


FIGURE 2-3. SPREAD FUNCTION



RESOLUTION OF TWO STARS

FIGURE 2-4. RESOLUTION OF TWO STARS

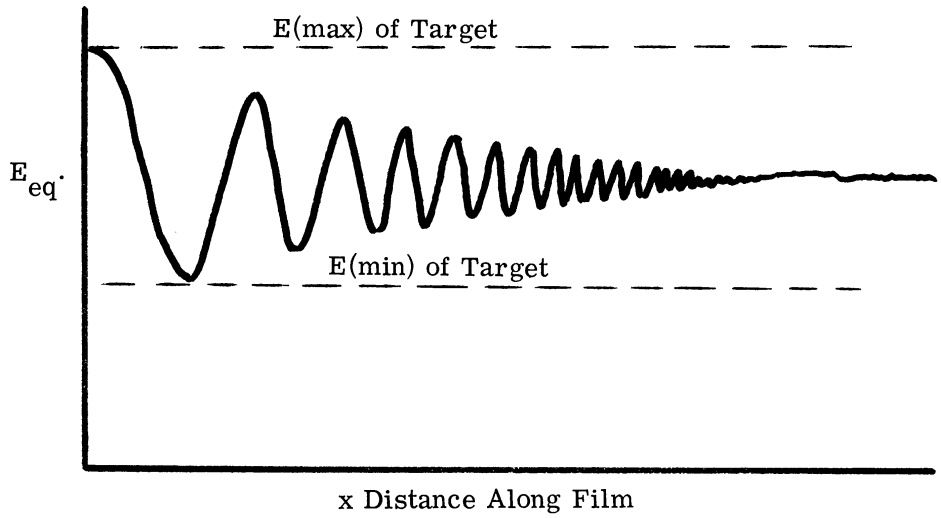
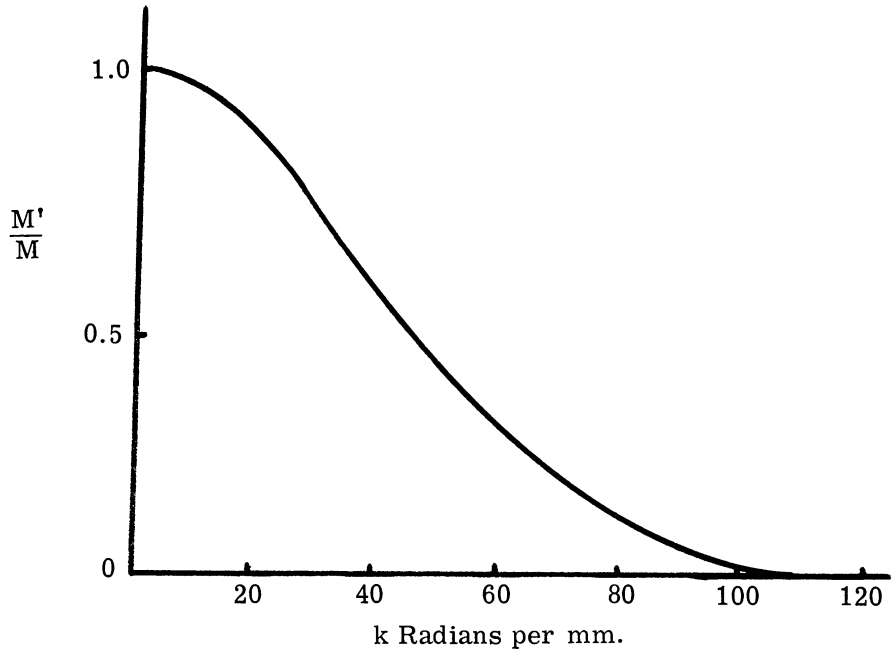
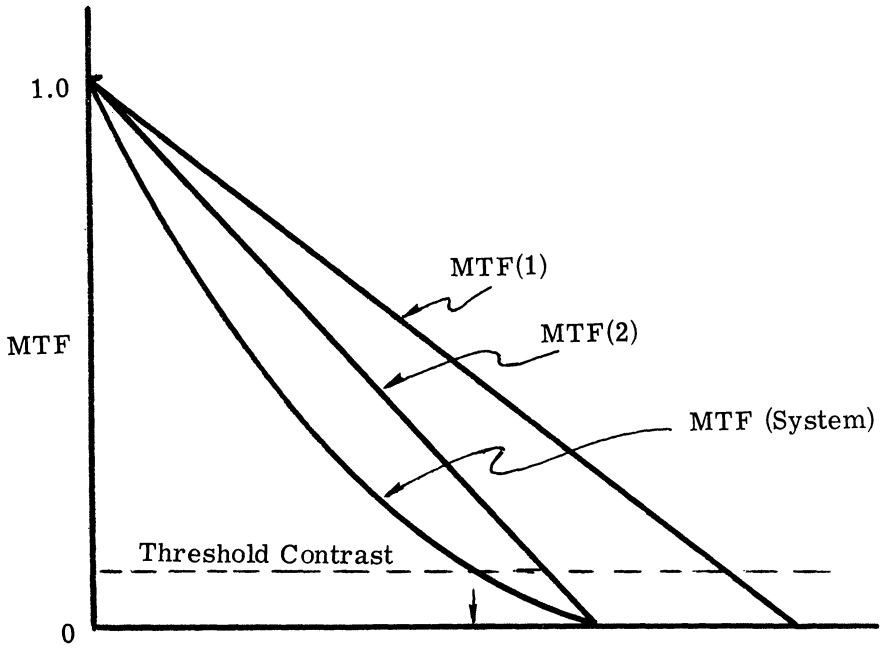


FIGURE 2-5. E_{eq} AS A FUNCTION OF DISTANCE ALONG THE FILM FOR A TARGET WITH INCREASING K VALUES ACROSS THE TARGET. Target modulation M is held constant.



MODULATION TRANSFER FUNCTION

FIGURE 2-6. MODULATION TRANSFER FUNCTION



SYSTEM MTF

FIGURE 2-7. SYSTEM MTF

LENS DATA INFORMATION SHEET
(Reference MIL–STD–150A for definition of terms and methods of measurement)

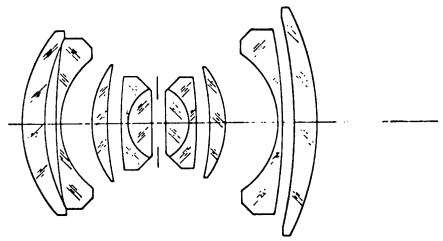
DATE: 15 Sept. 1967

TYPE (MIL–STD–150A)	NAME
Type II Aerial Mapping	Geocon I

FOCAL LENGTH	RELATIVE APERTURE	FORMAT	MANUFACTURER
6 in.	f/5.6	9 in. x 9 in.	J. G. Baker

STOCK NO.	SPECIFICATION NO.	STATUS

GENERAL DESCRIPTION



CHARACTERISTICS

EQUIVALENT FOCAL LENGTH: 6 in. (152.085 mm)	TRANSMITTANCE: 62% (Without filter)
BACK FOCAL DISTANCE: 2.92 in. (74.2 mm)	COLOR CORRECTION: 5461–6563 Å
FRONT VERTEX BACK FOCAL DISTANCE: 9.4 in.	DISTORTION *(Maximum)*: ± 7.5 Microns
FLANGE FOCAL DISTANCE: 3.95 in.	LENS CELL DIAMETER *(Maximum)*:
SEMI–FIELD ANGLE: 45°	OVERALL LENGTH: 6.5 in.
RELATIVE APERTURE: f/5.6	WEIGHT:

REMARKS:

Used on the following aircraft cameras: KC-4

FIGURE 3-1

FOCAL LENGTH	RELATIVE APERTURE	FORMAT	MANUFACTURER
6 in.	f/5.6	9 in. x 9 in.	J. G. Baker

Photographic Resolving Power in Lines /mm

FIELD ANGLE (°)	0	5	10	15	20	25	30	35	40	45
FILM Plus X, TARGET CONTRAST High, FILTER W-12										
RADIAL	42	45	41	46	54	49	47	37	26	29
TANGENTIAL	38	38	41	40	40	40	36	28	17	15
AWAR 40 L/mm										
FILM SO-213, TARGET CONTRAST High, FILTER										
RADIAL	70	74	78	71	94	71	61	51	40	37
TANGENTIAL	100	80	64	74	63	44	42	35	33	35
AWAR 54 L/mm										
FILM SO-213, TARGET CONTRAST Low, FILTER										
RADIAL	56	47	53	51	53	53	46	41	32	33
TANGENTIAL	40	47	46	49	40	33	35	23	22	24
AWAR 38 L/mm										
VISUAL, TARGET CONTRAST, FILTER										
RADIAL										
TANGENTIAL										

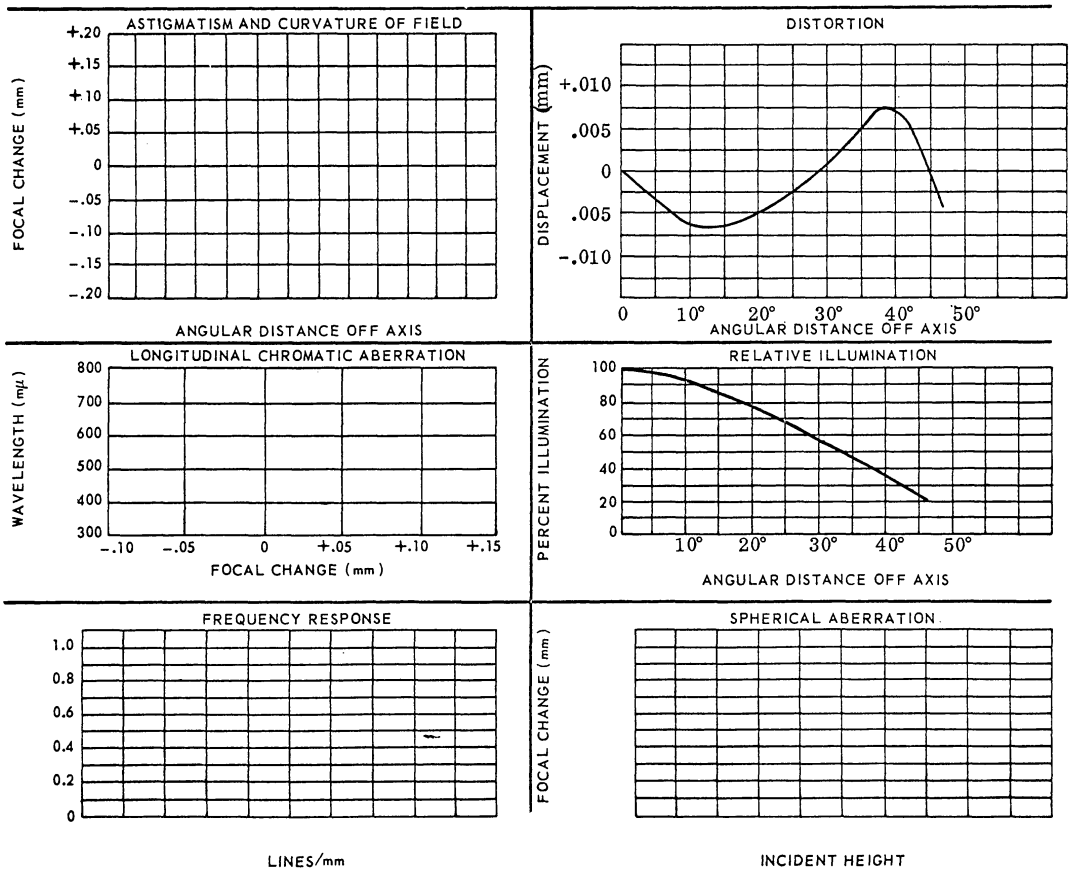


FIGURE 3-2

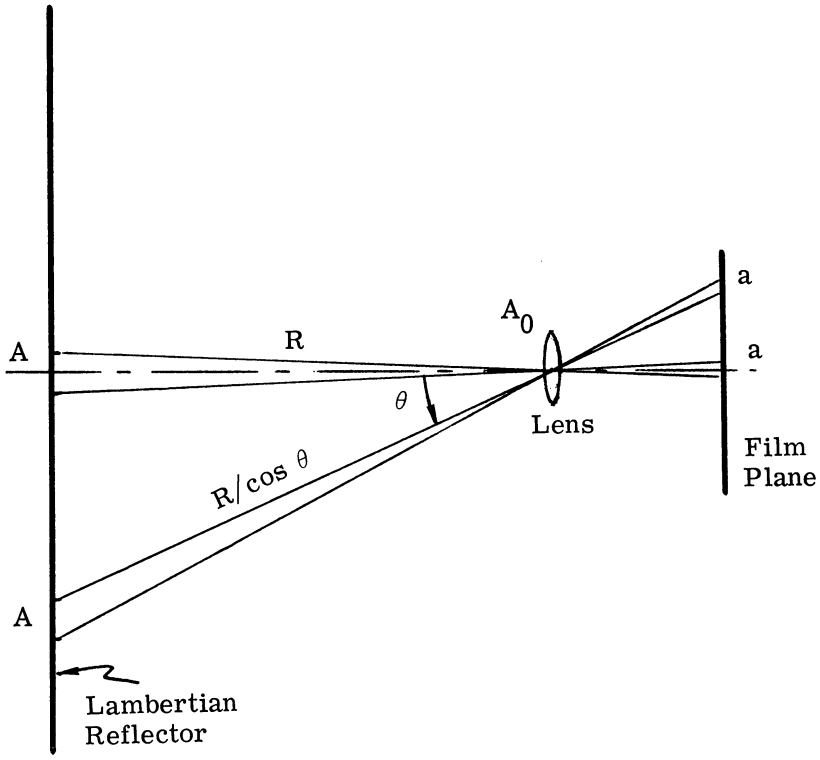


FIGURE 3-3. SCHEMATIC OF CAMERA ILLUMINATION

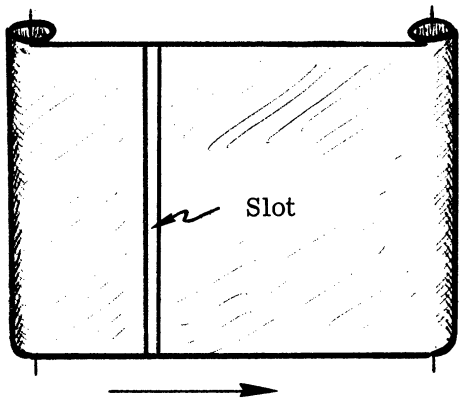


FIGURE 3-4. SCHEMATIC OF FOCAL PLANE SHUTTER

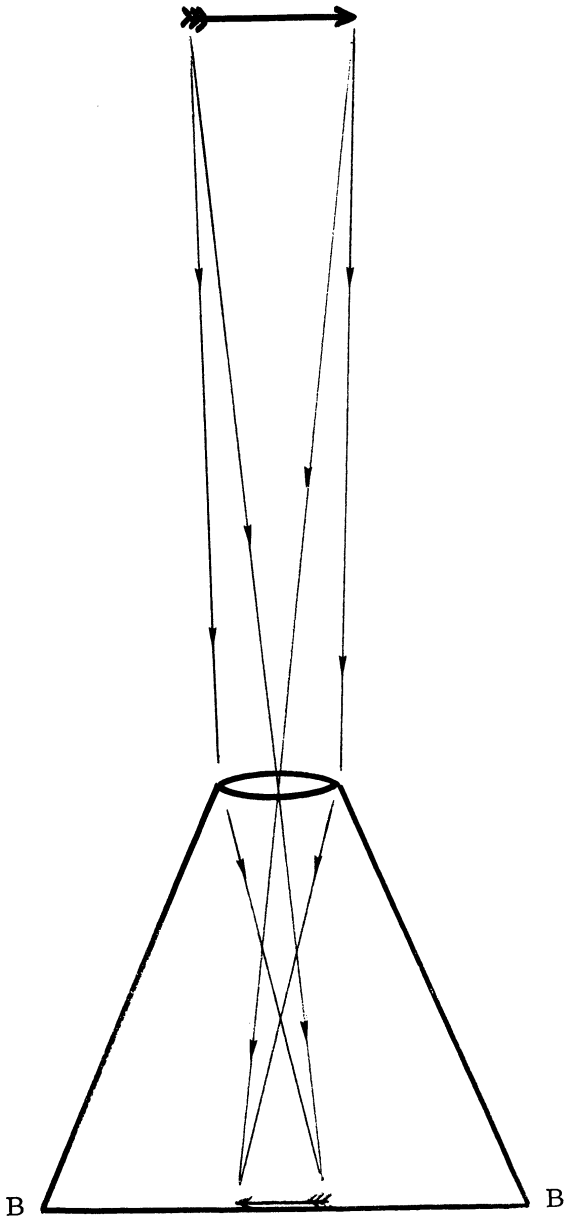


FIGURE 5-1

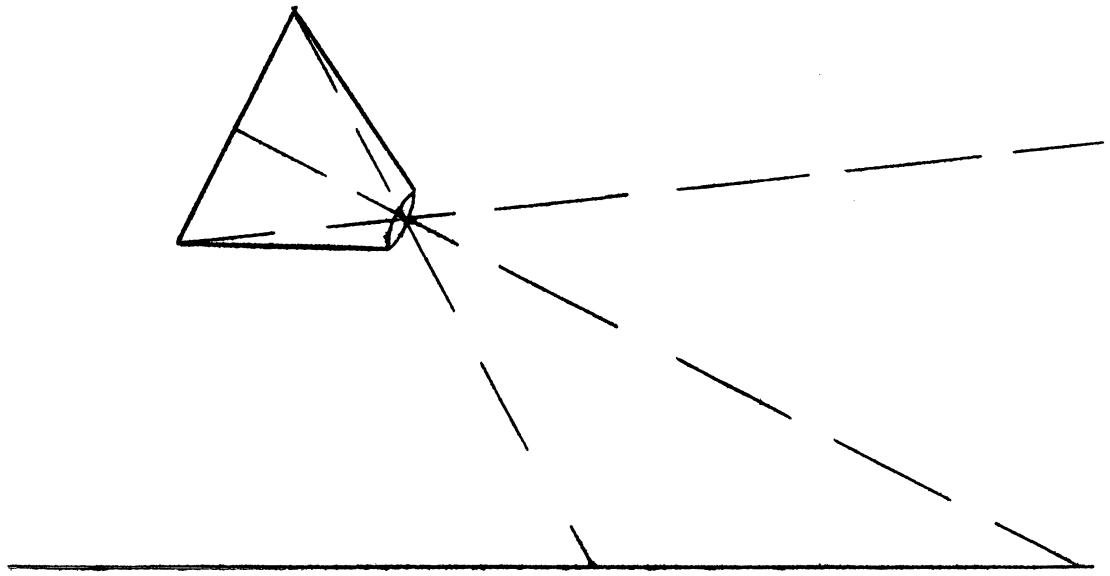


FIGURE 5-2

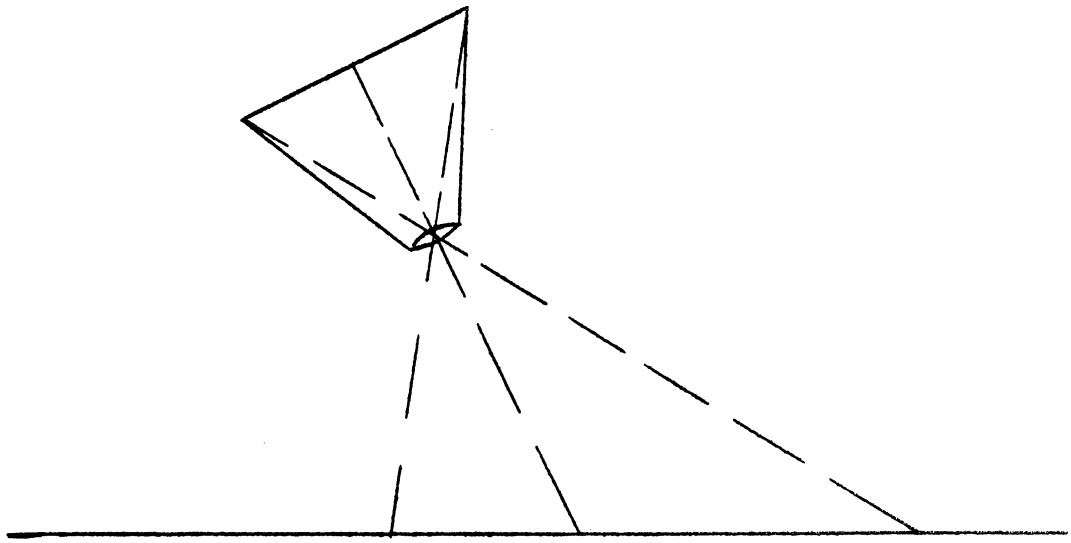


FIGURE 5-3

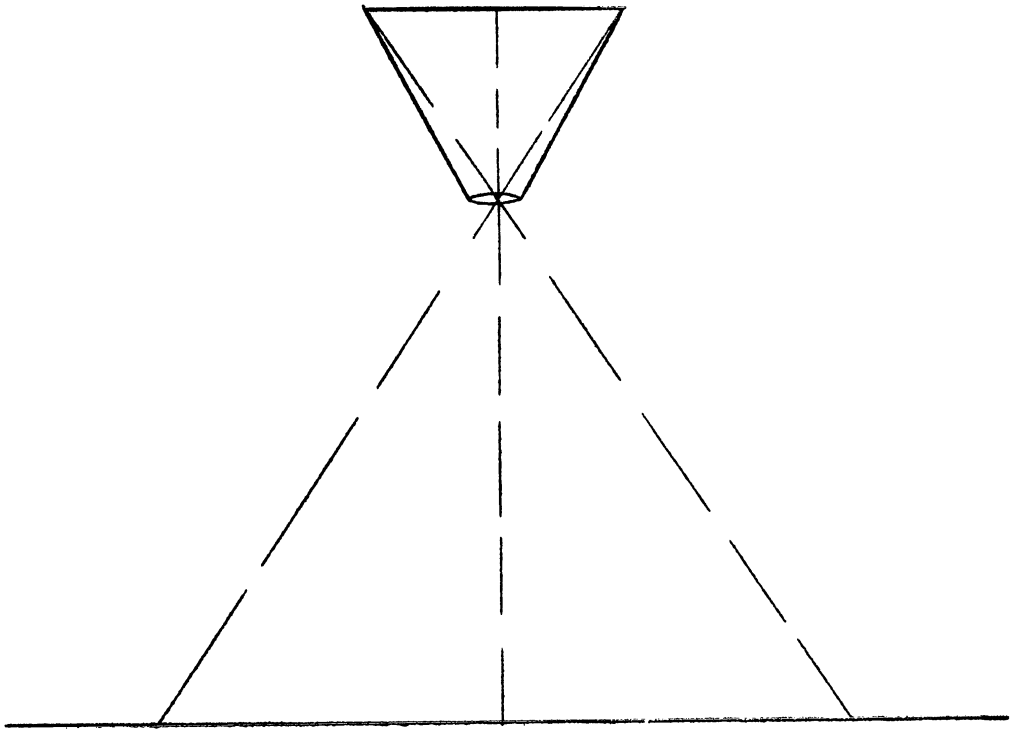


FIGURE 5-4

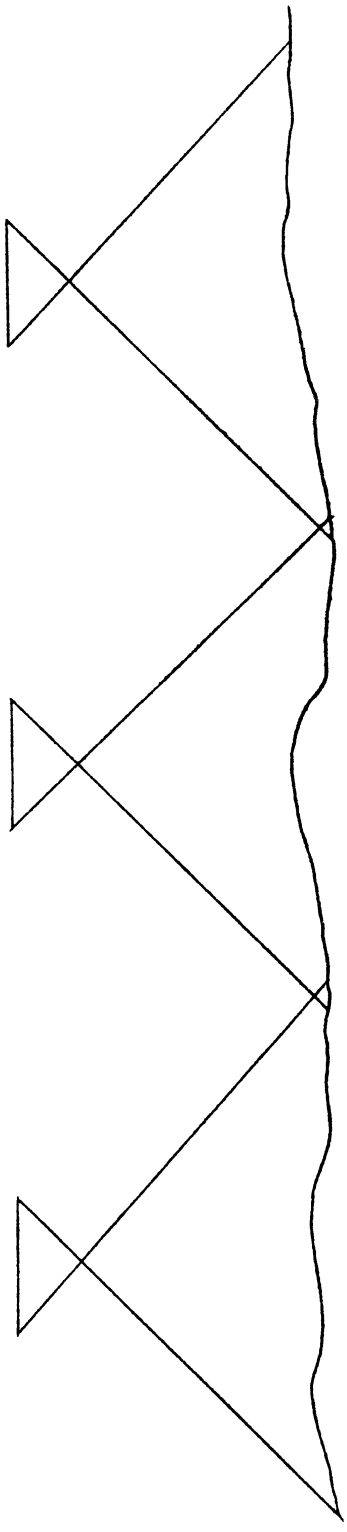


FIGURE 5-5

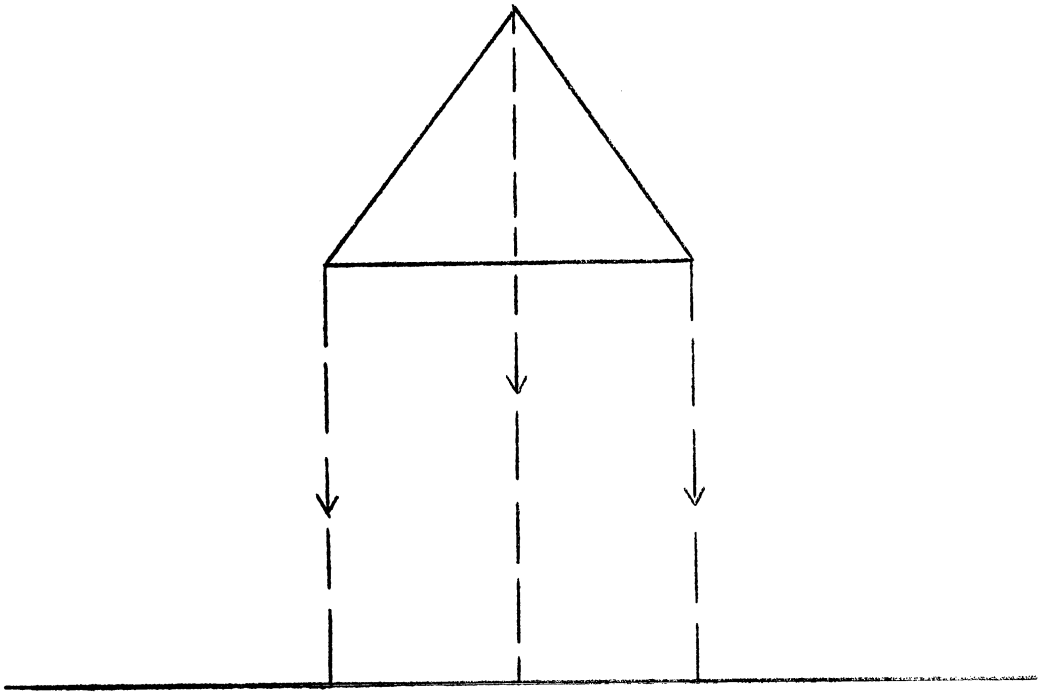


FIGURE 5-6

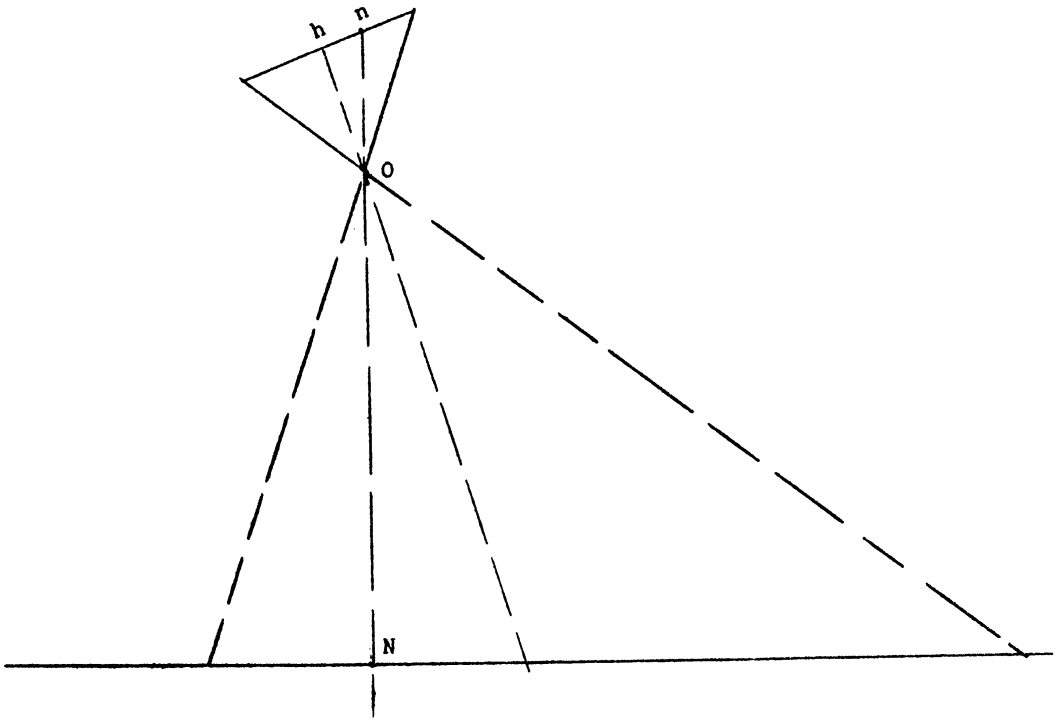


FIGURE 5-7

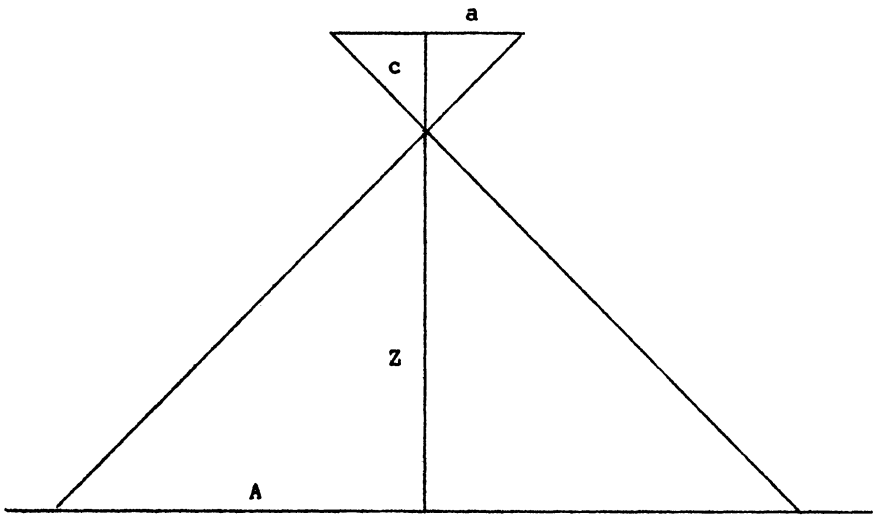


FIGURE 5-8

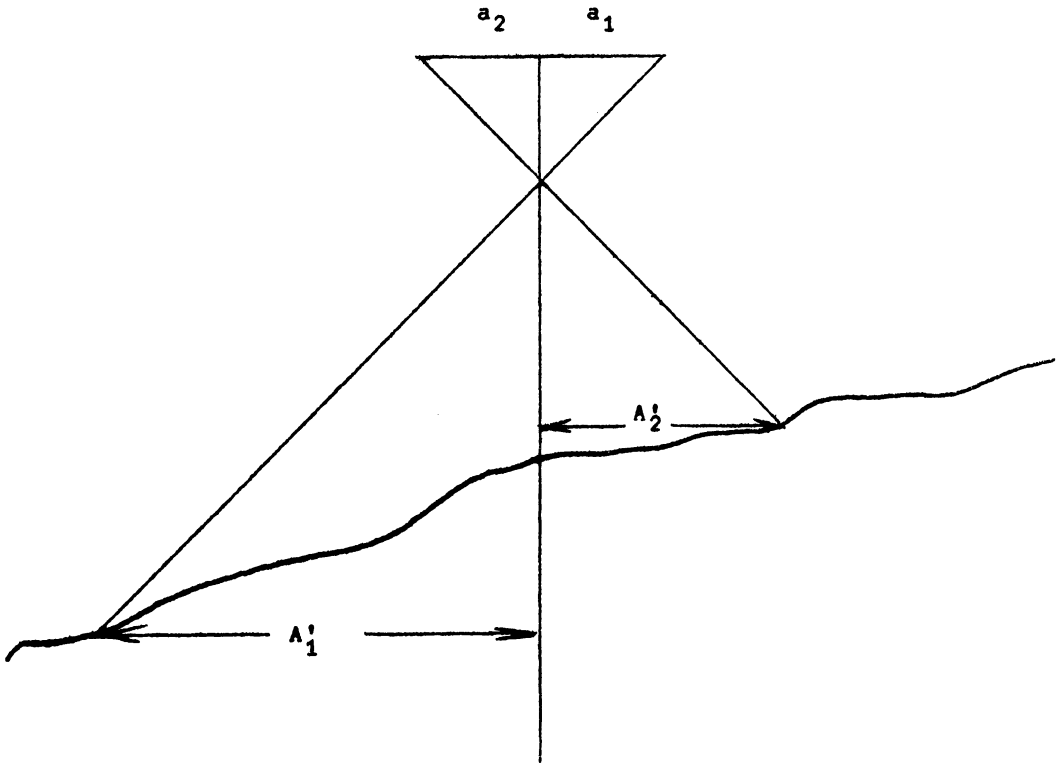


FIGURE 5-9

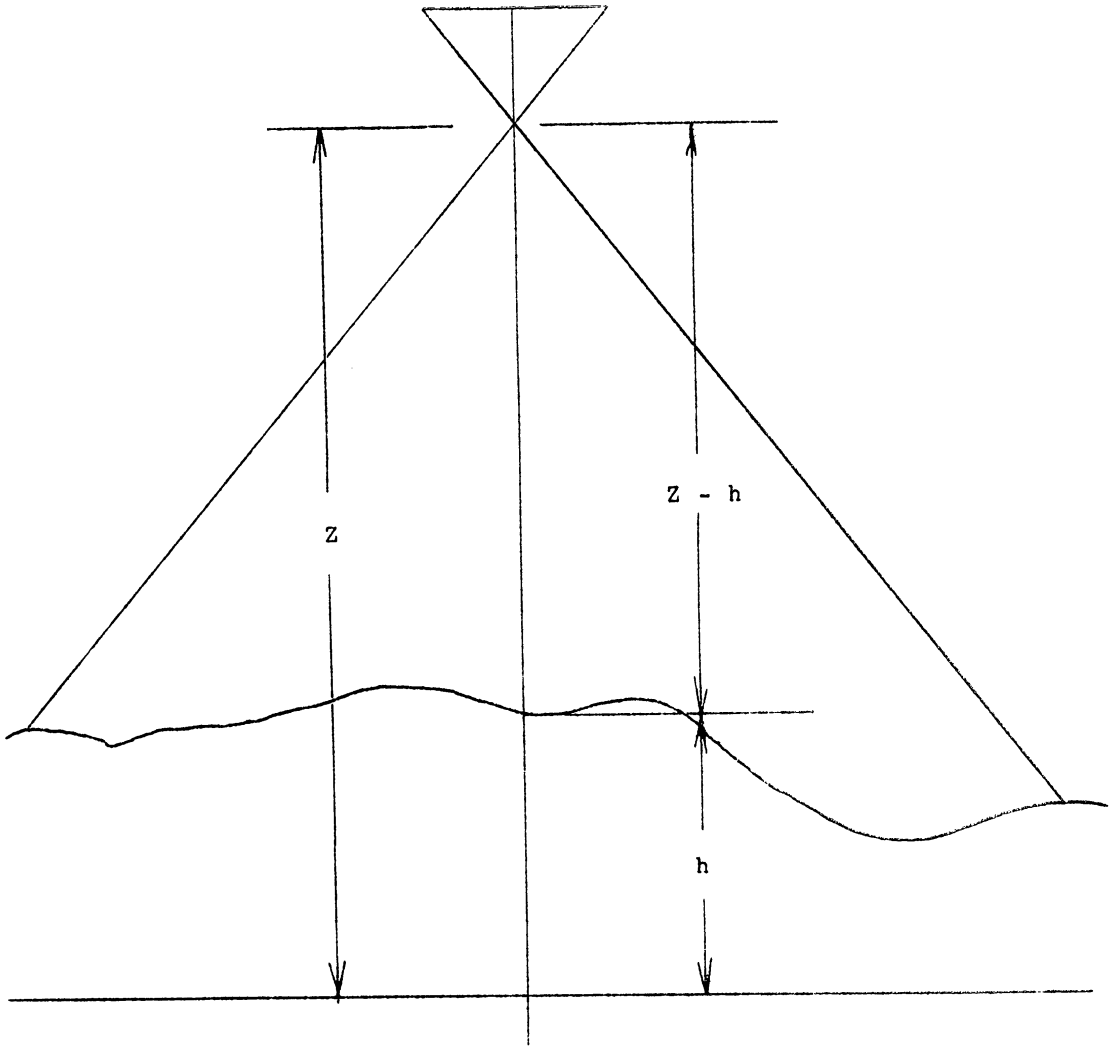


FIGURE 5-10

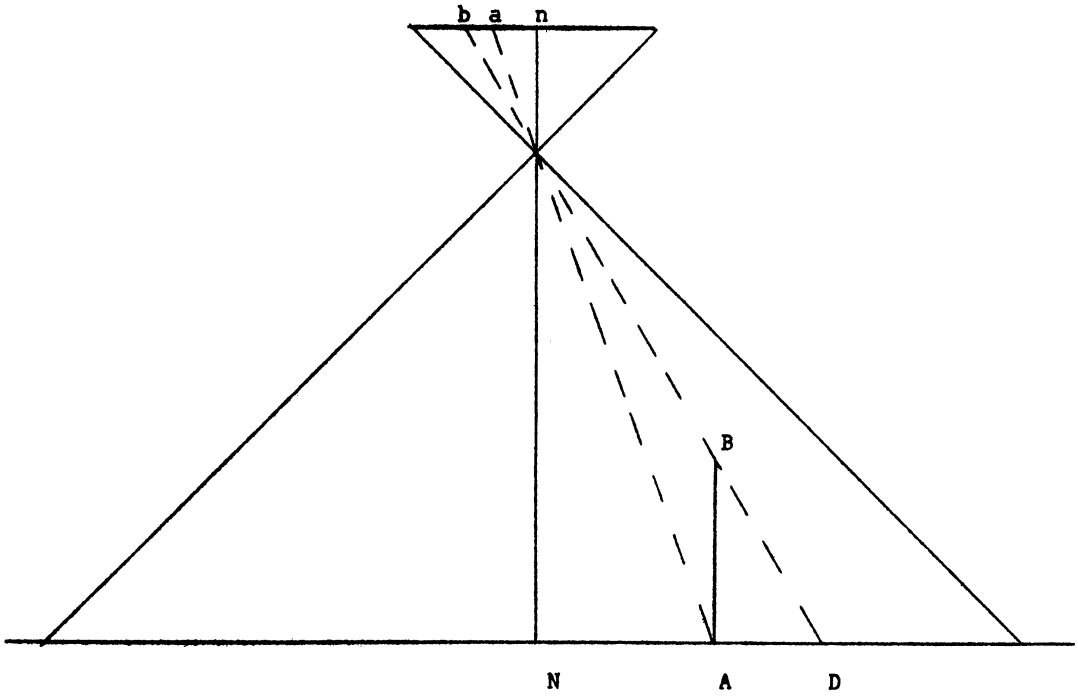


FIGURE 5-11

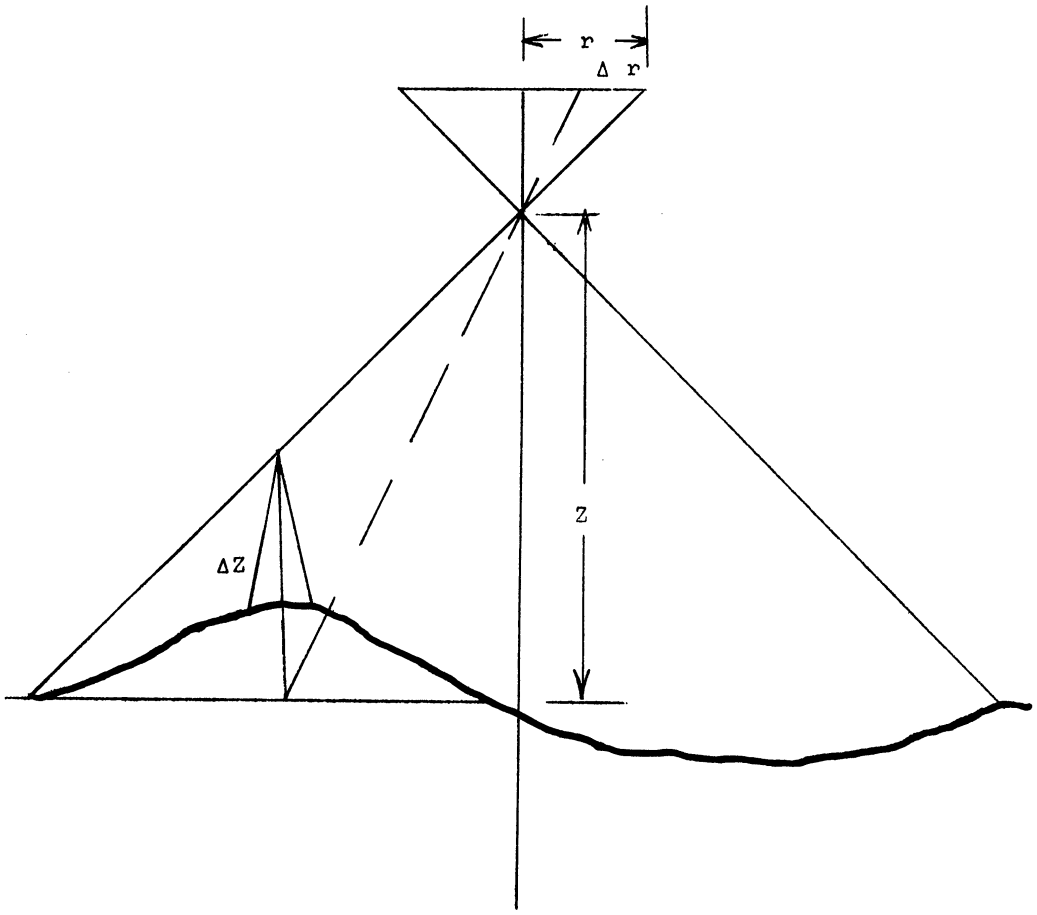


FIGURE 5-12

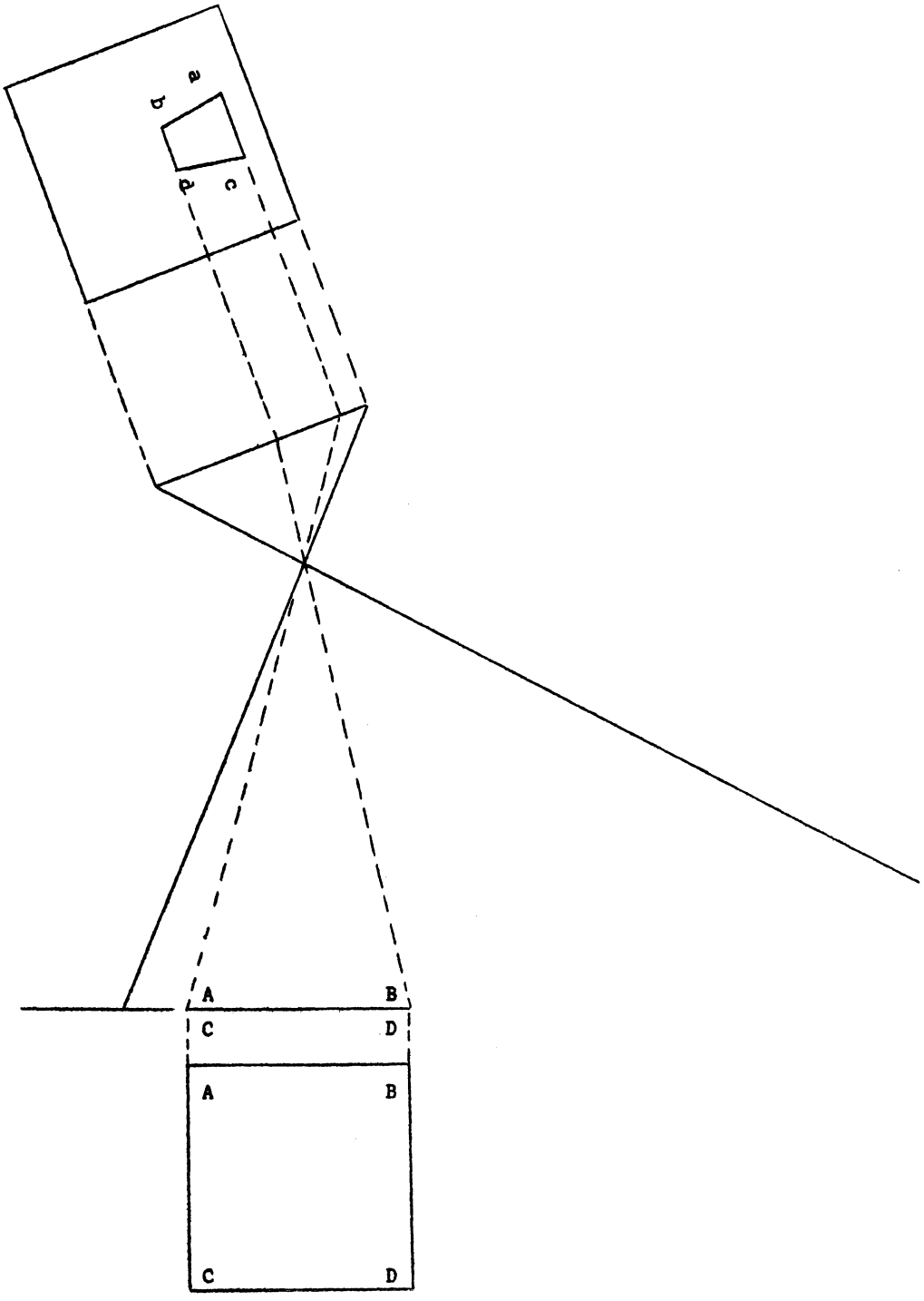


FIGURE 5-13

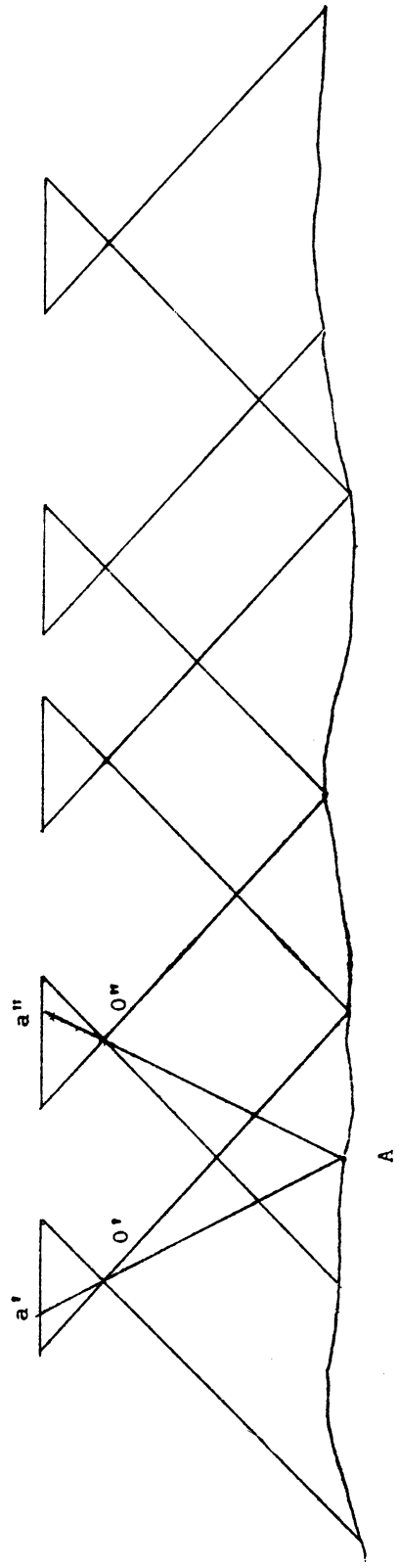


FIGURE 5-14

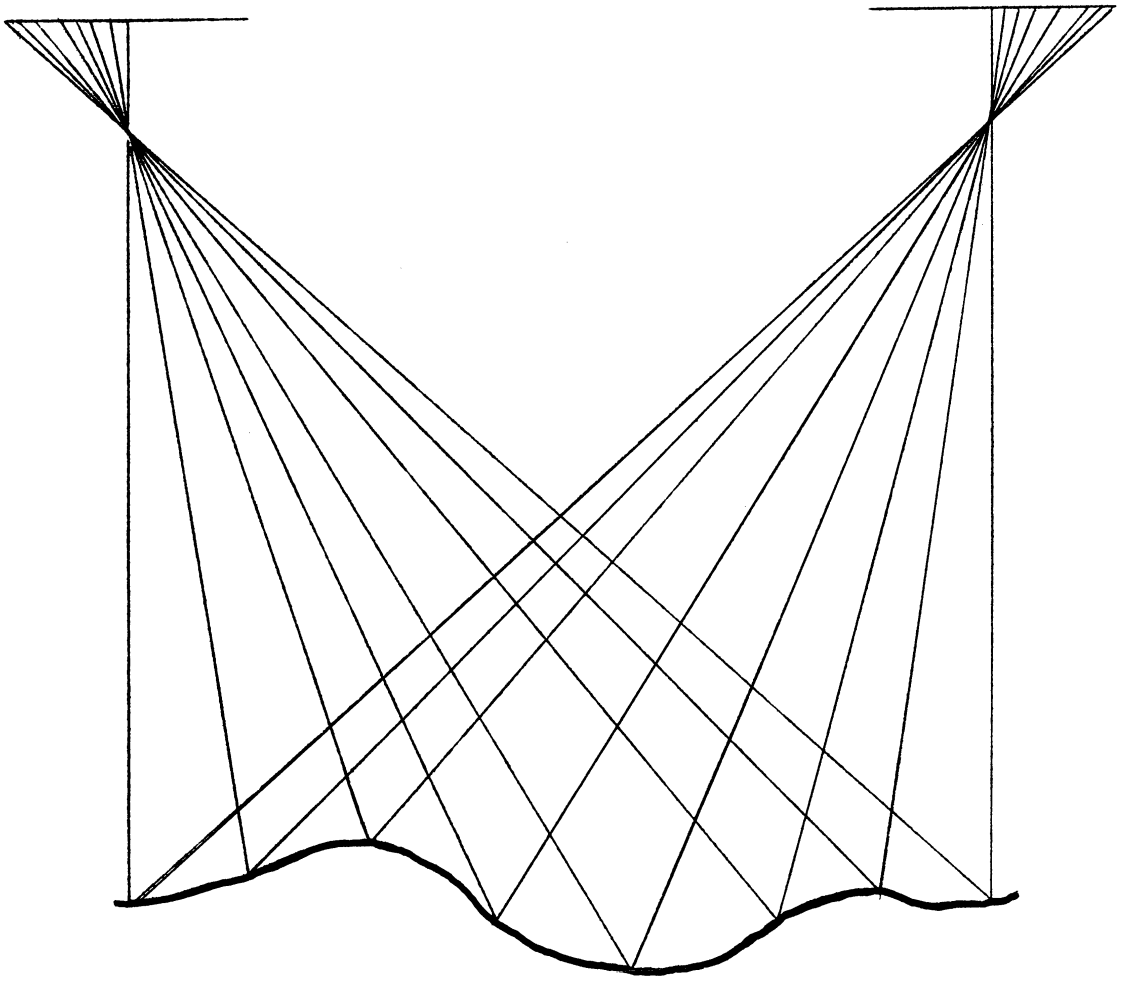


FIGURE 5-15

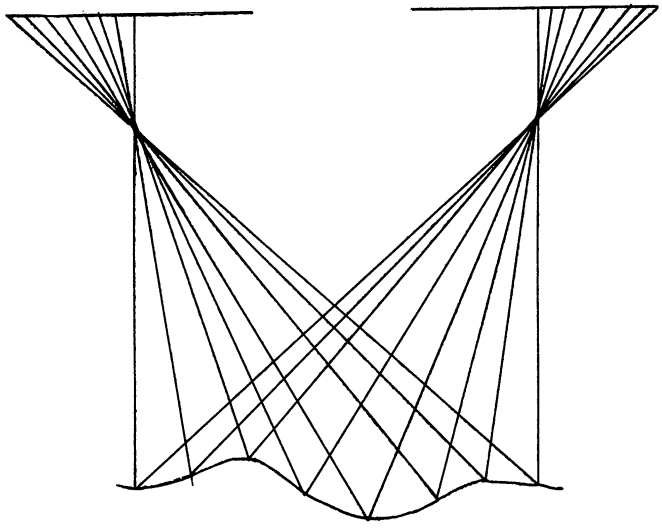


FIGURE 5-16

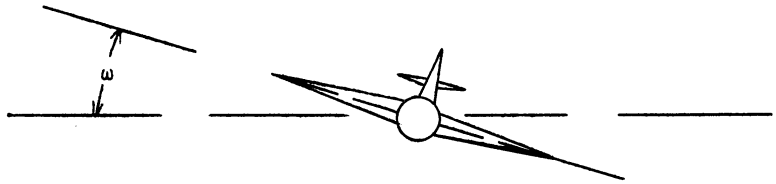


FIGURE 5-17

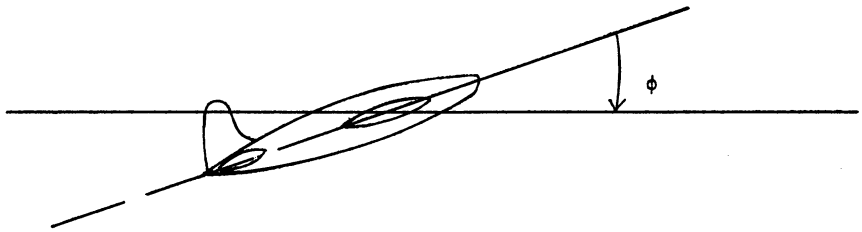


FIGURE 5-18

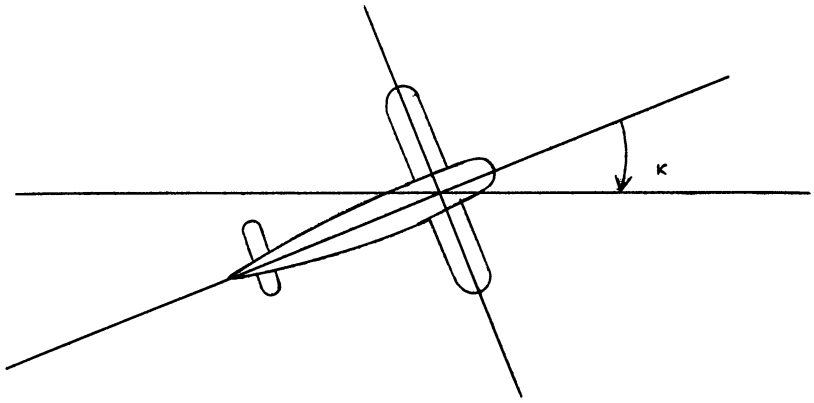


FIGURE 5-19

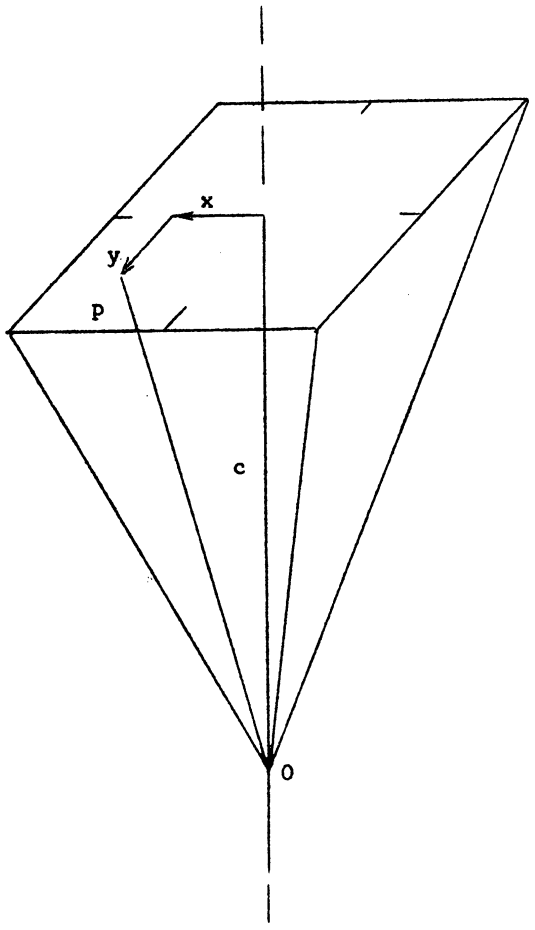


FIGURE 5-20

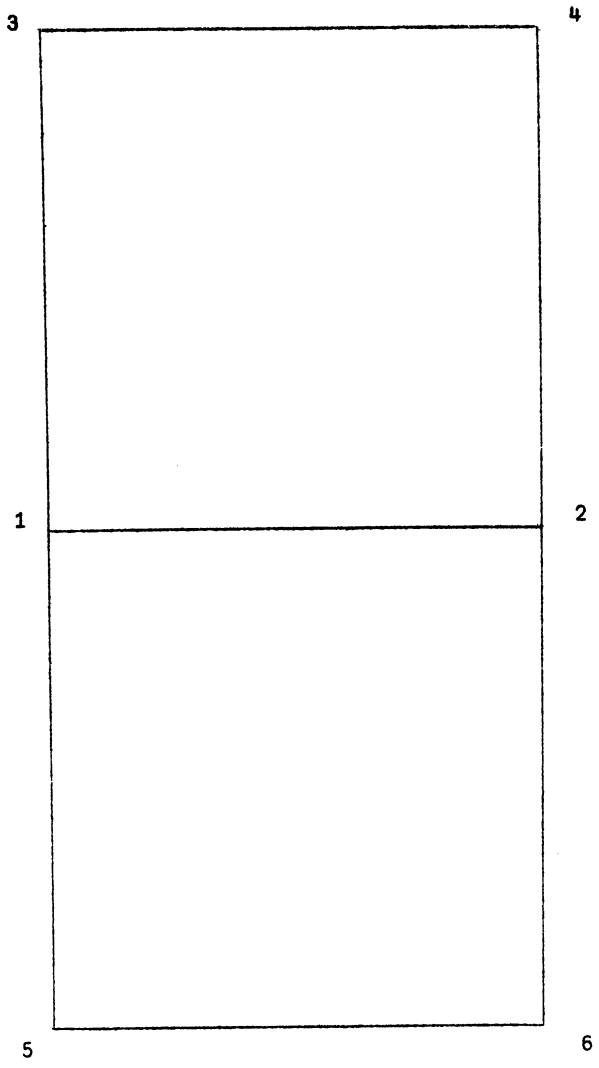


FIGURE 5-21

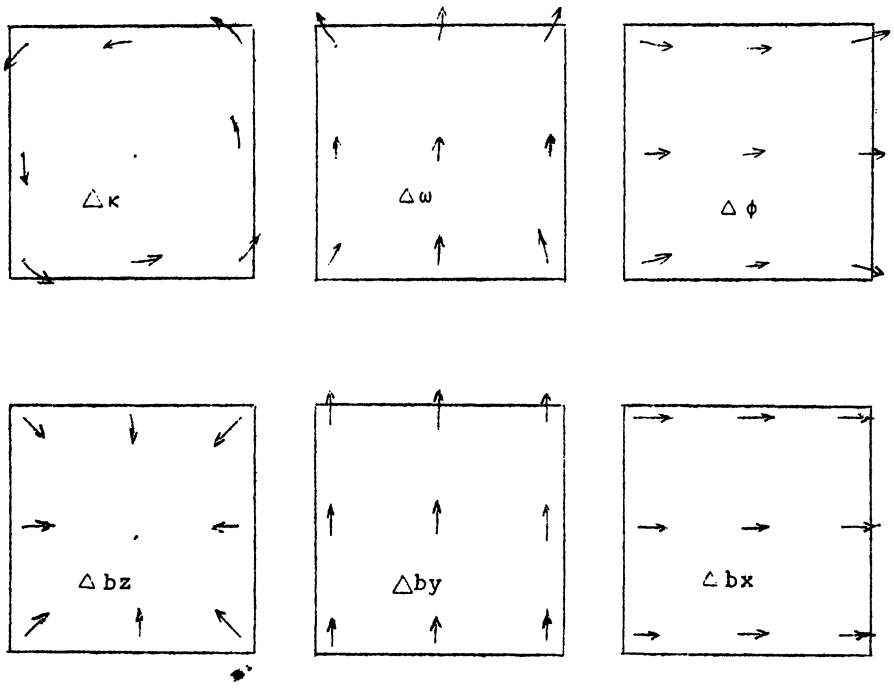


FIGURE 5-22

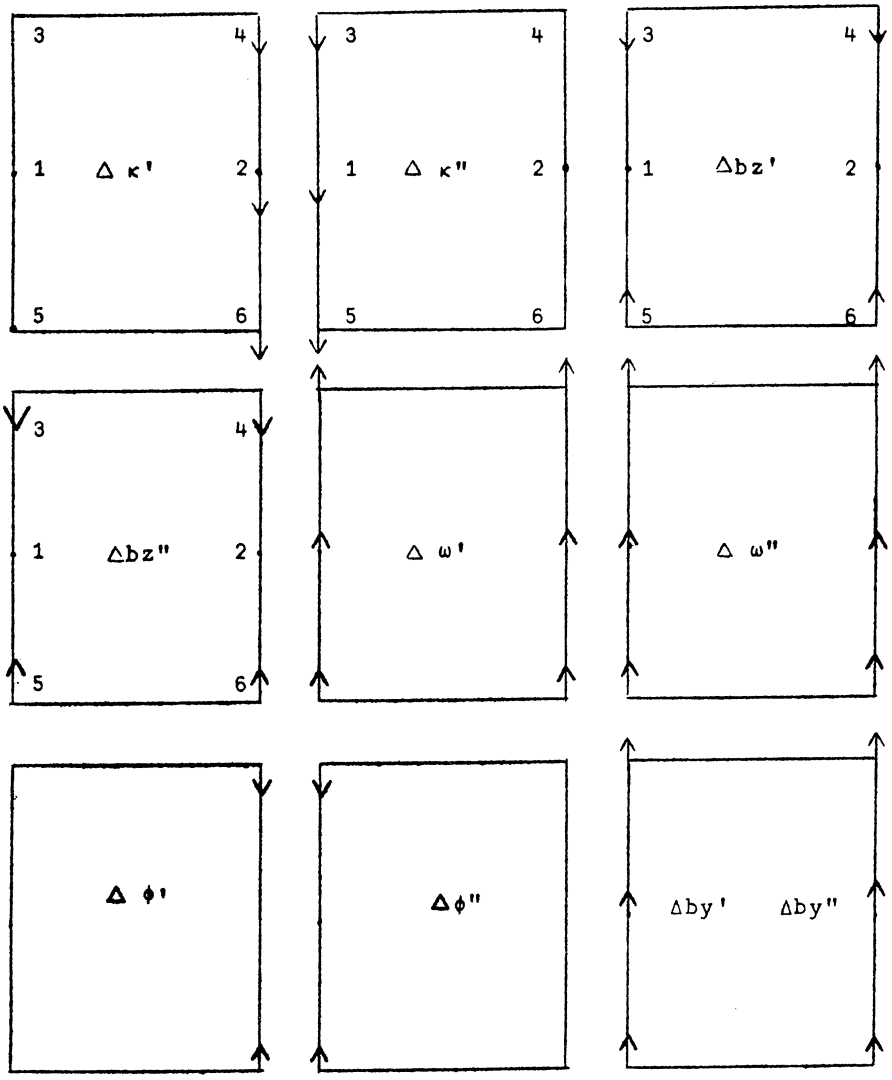
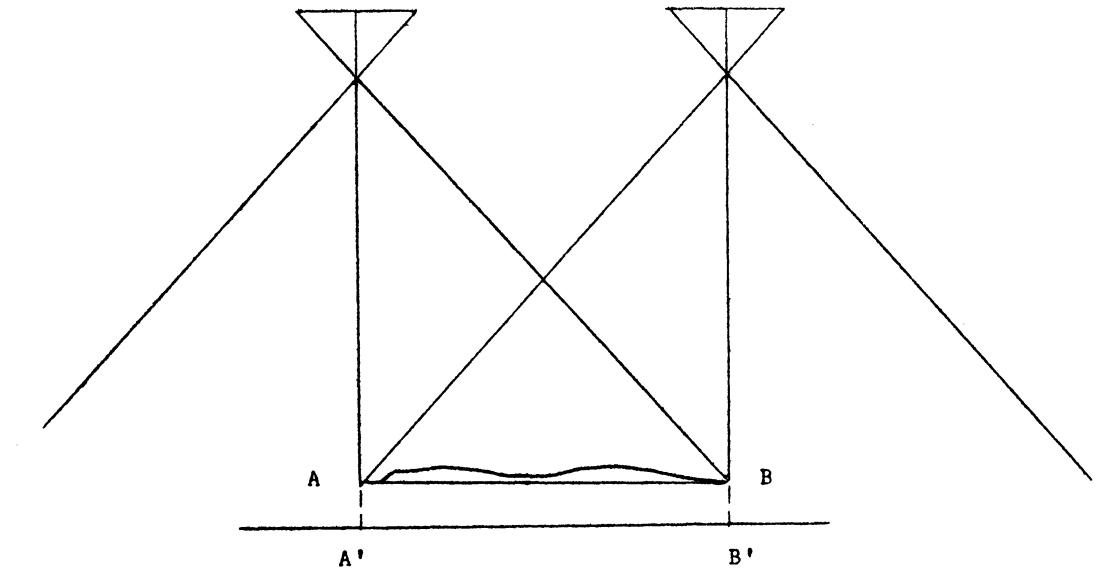
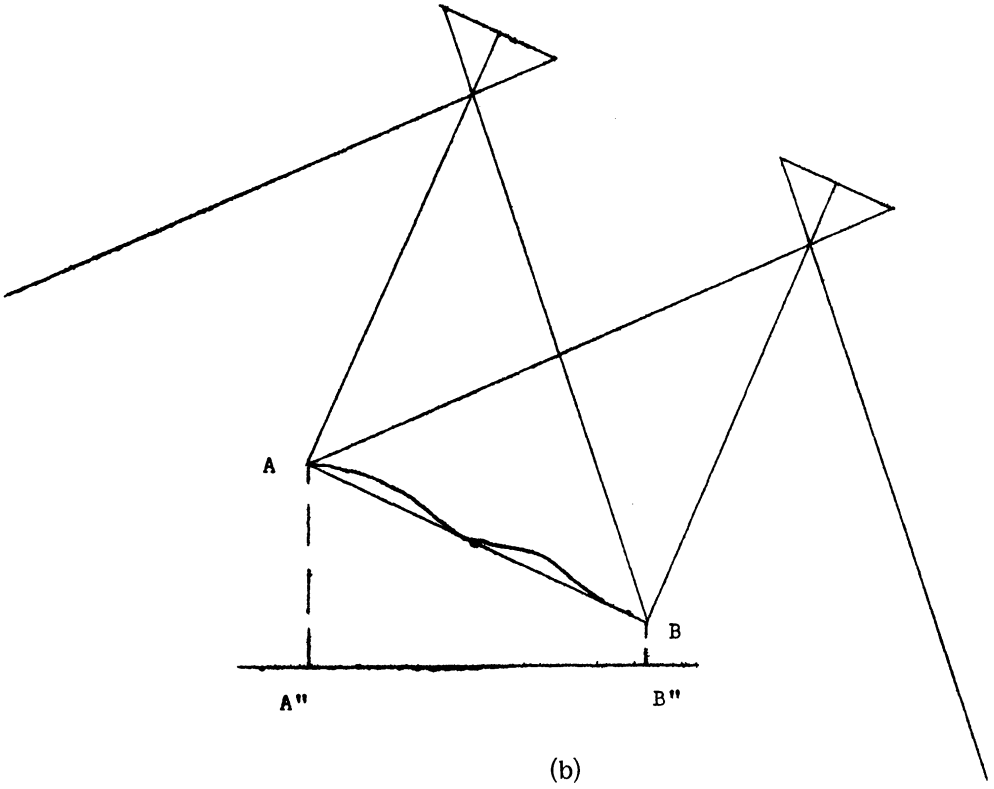


FIGURE 5-23



(a)



(b)

FIGURE 5-24

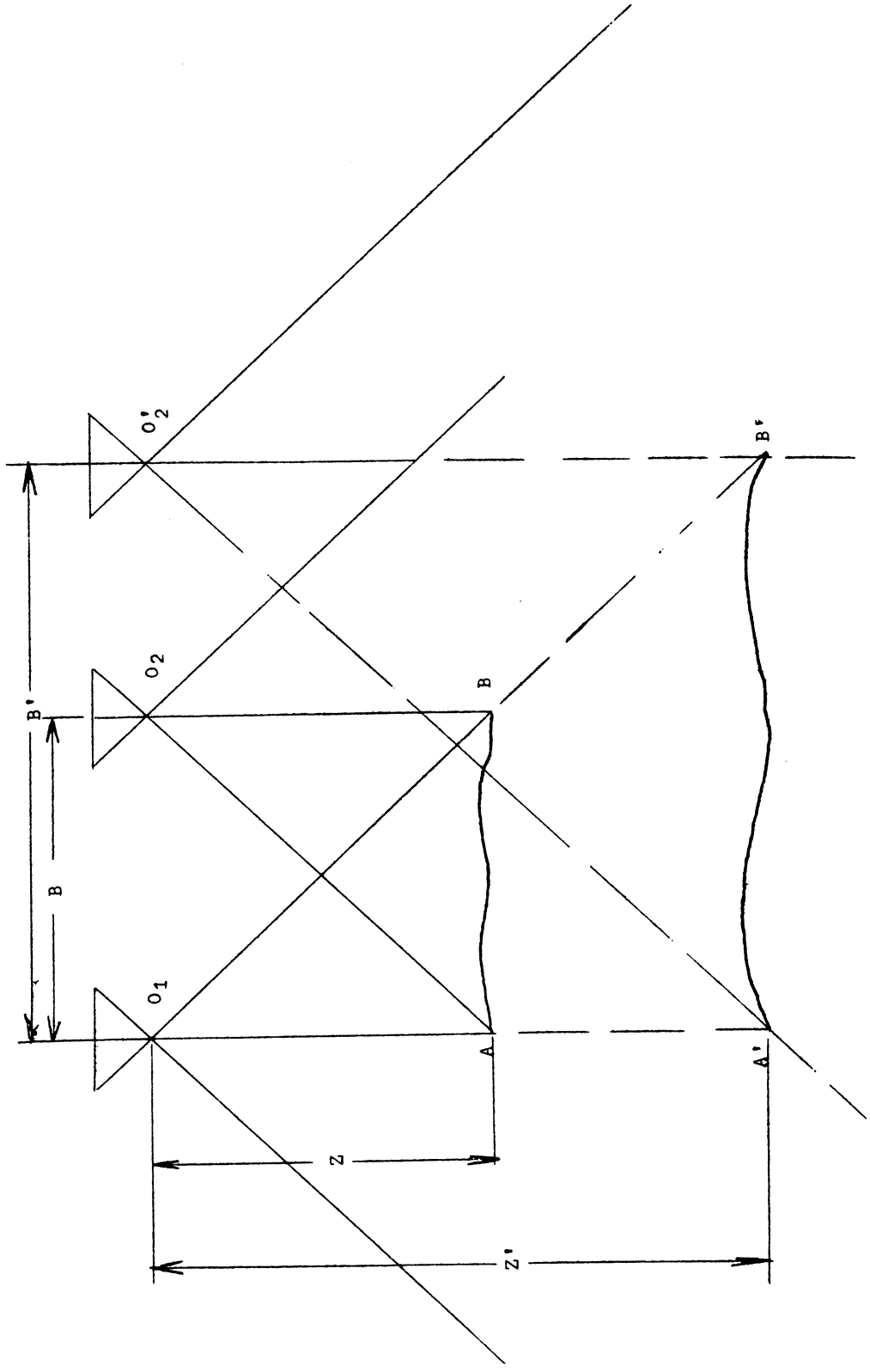


FIGURE 5-25

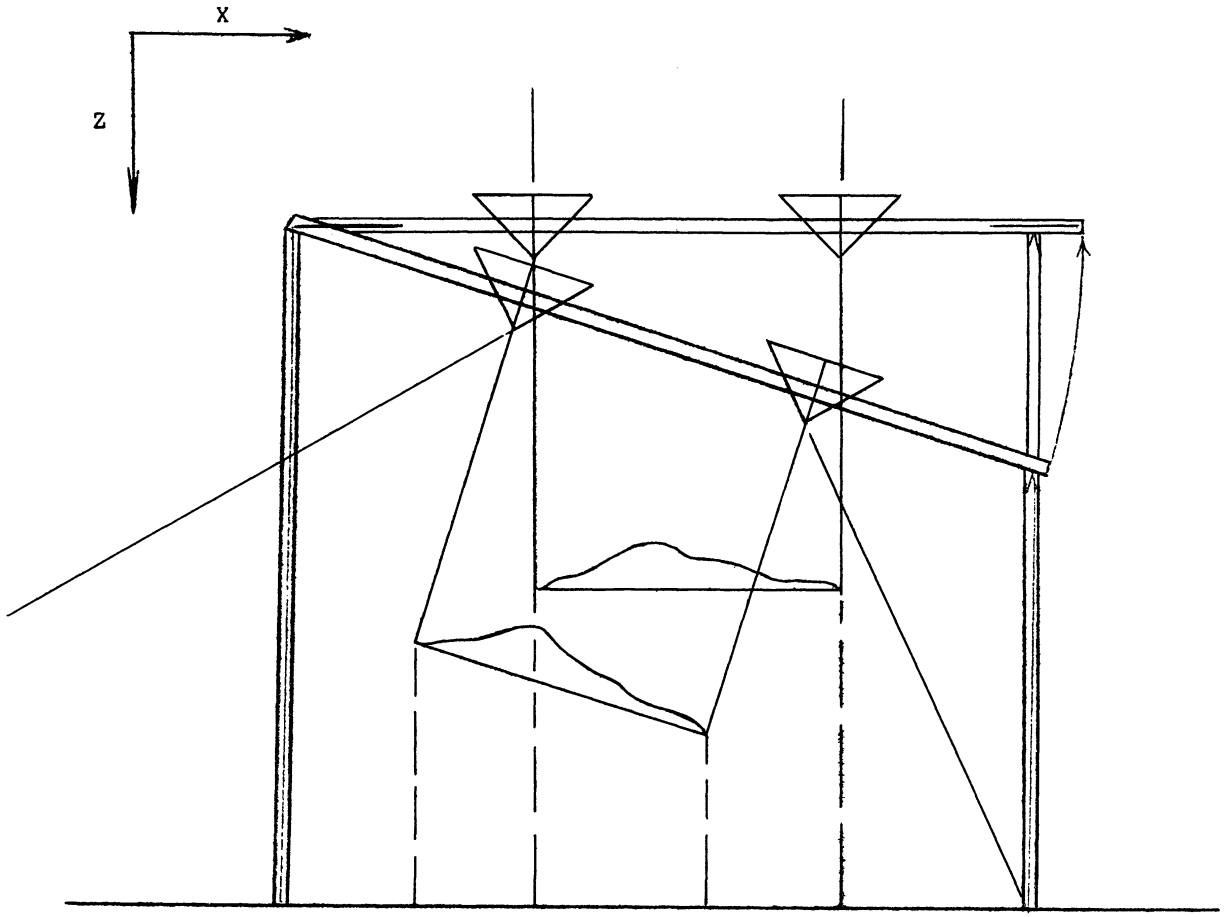


FIGURE 5-26

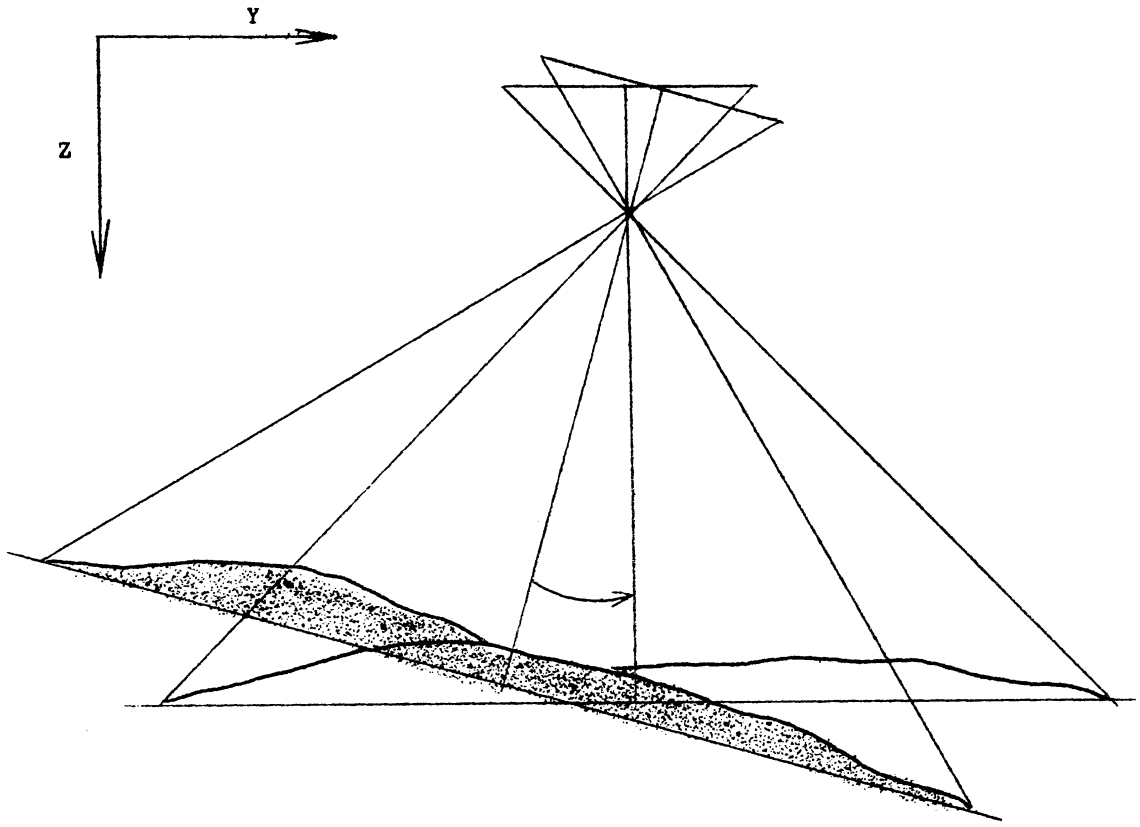
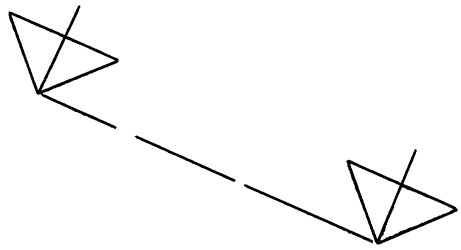
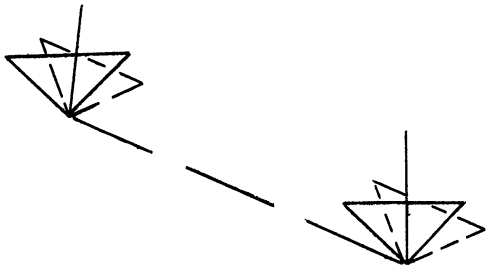


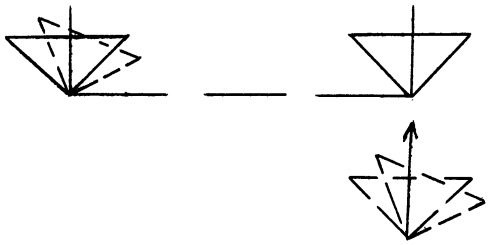
FIGURE 5-27



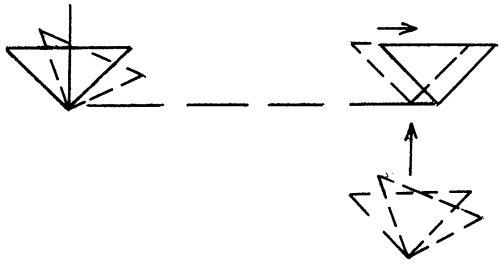
(a)



(b)



(c)



(d)

FIGURE 5-28

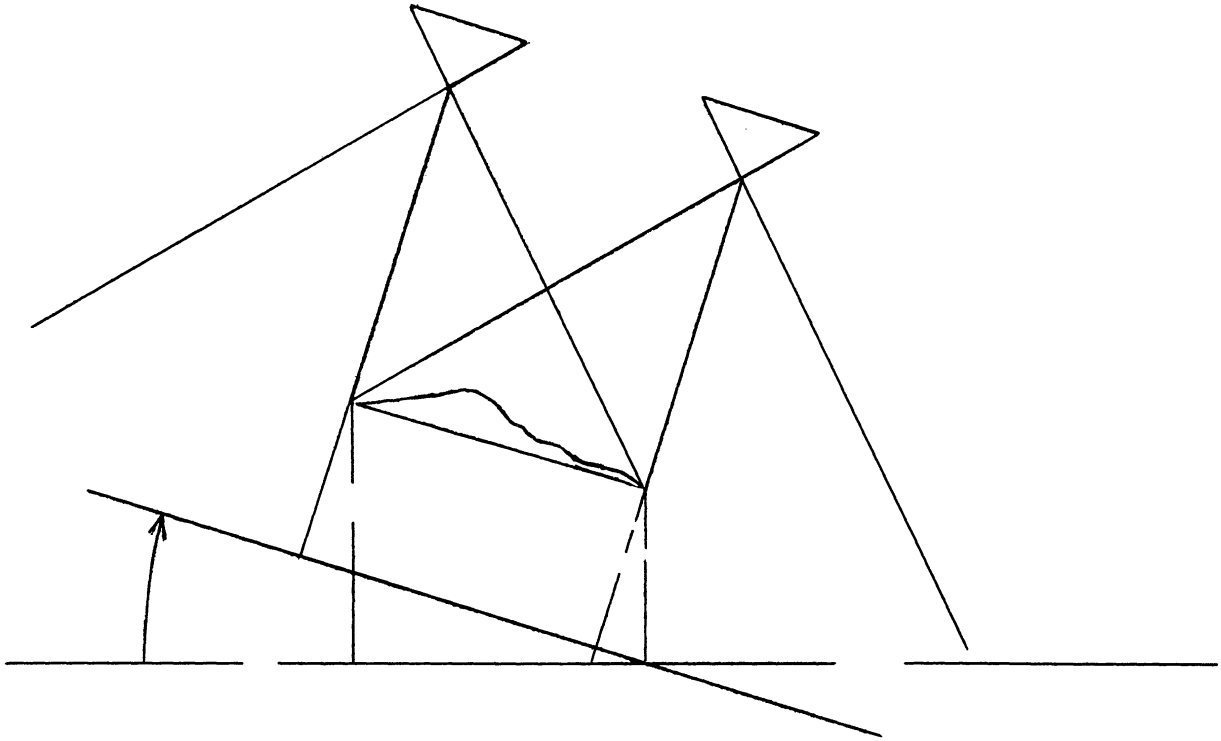


FIGURE 5-29

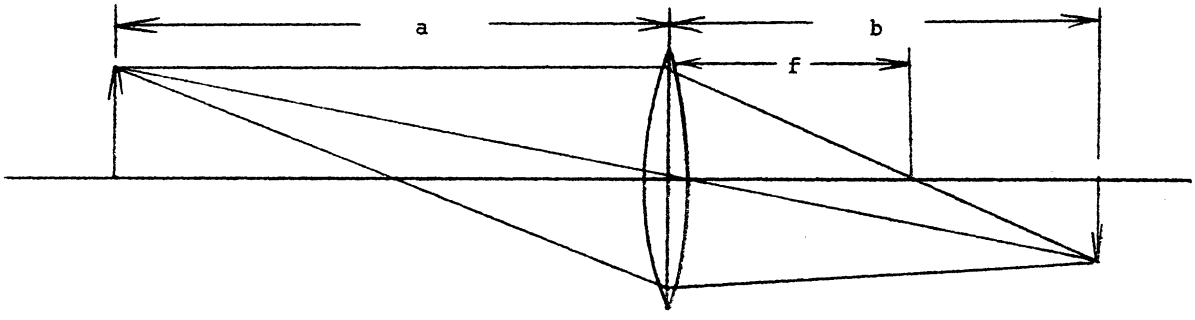


FIGURE 5-30

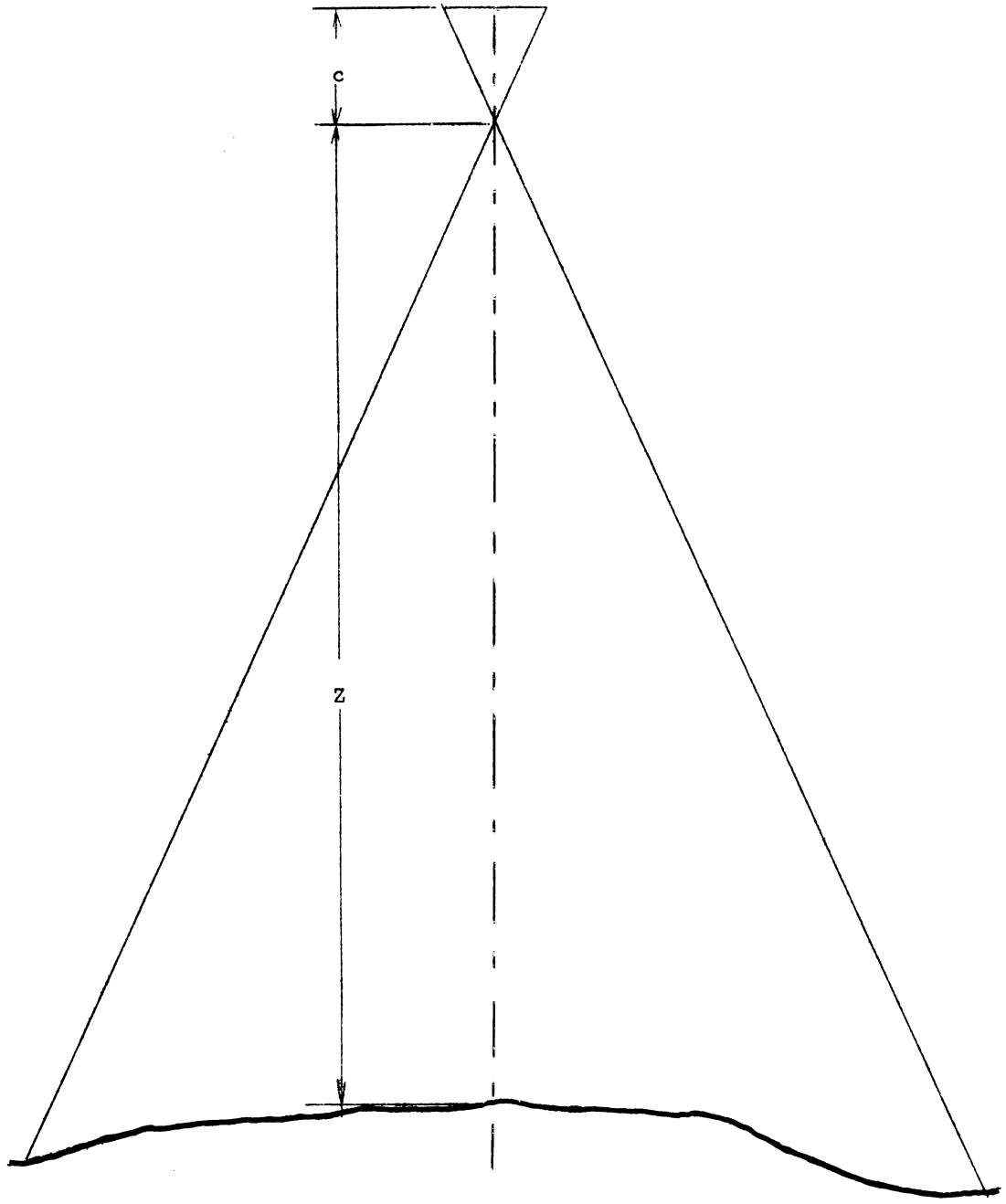


FIGURE 5-31

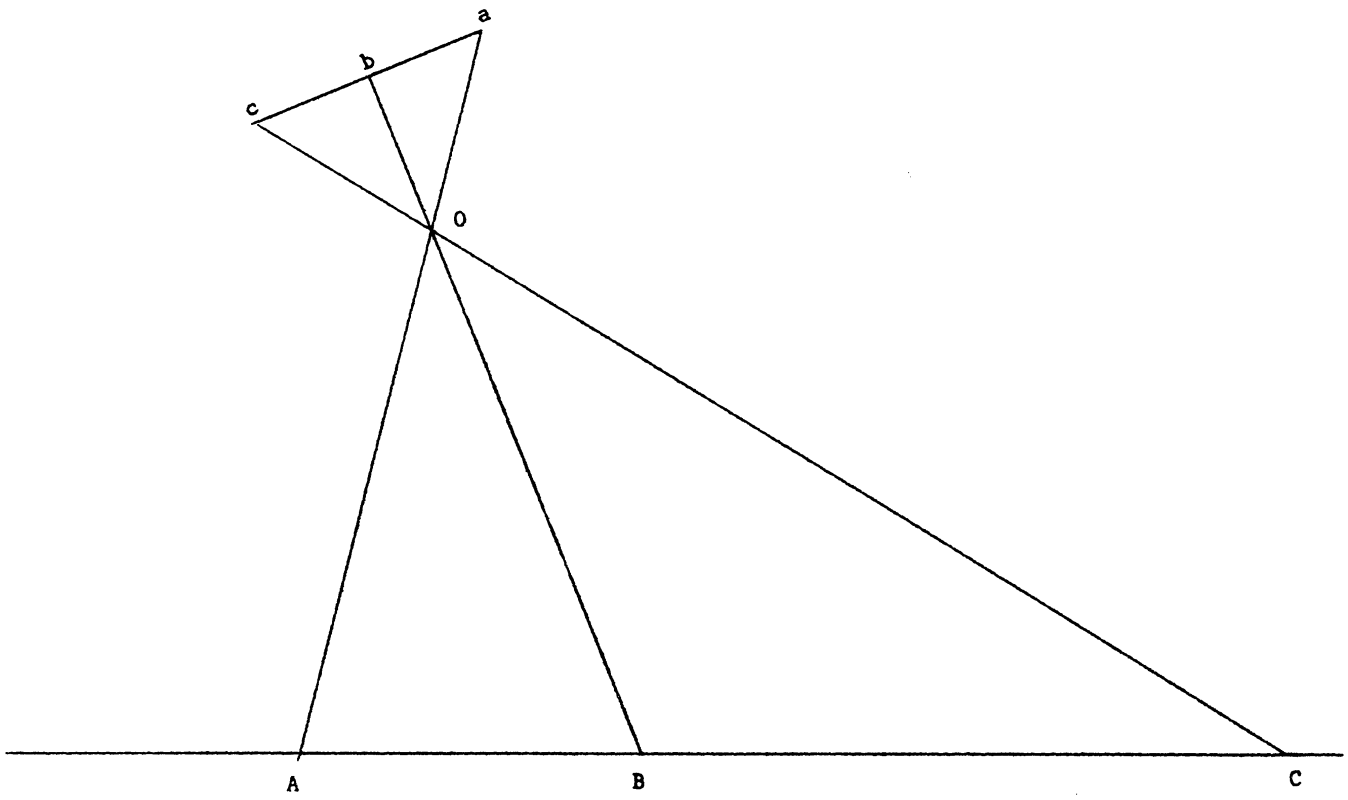


FIGURE 5-32

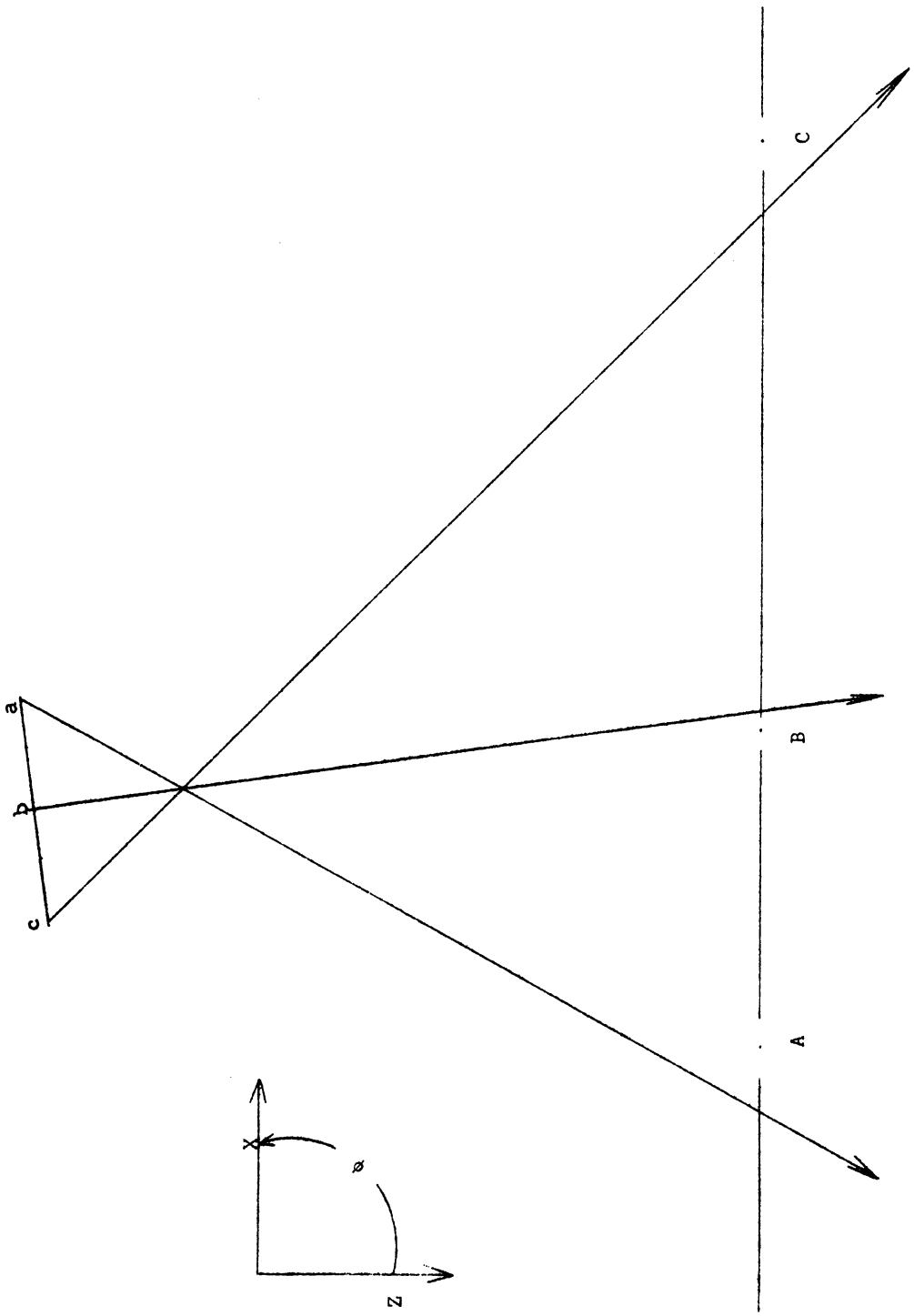


FIGURE 5-33

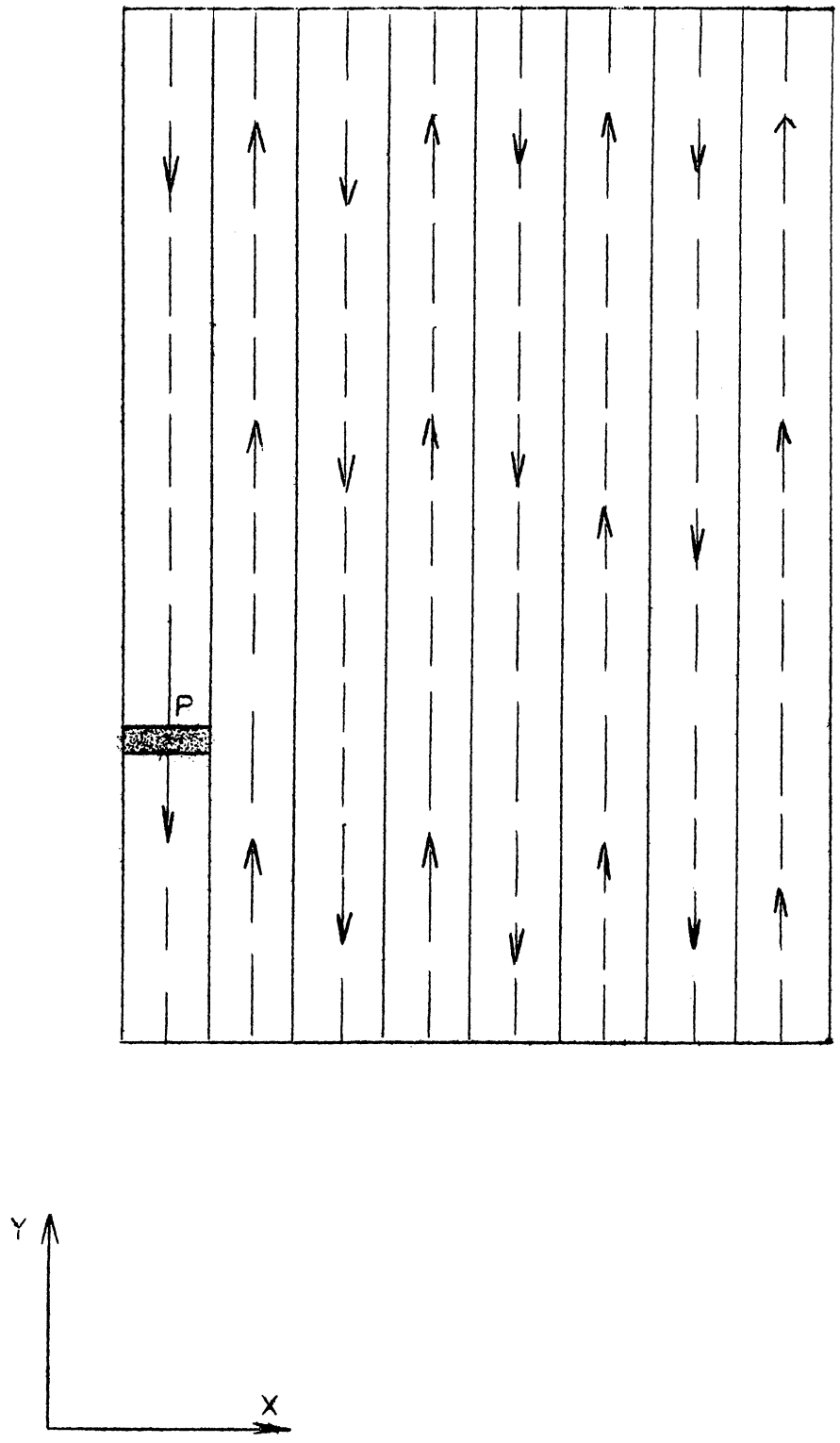


FIGURE 5-34

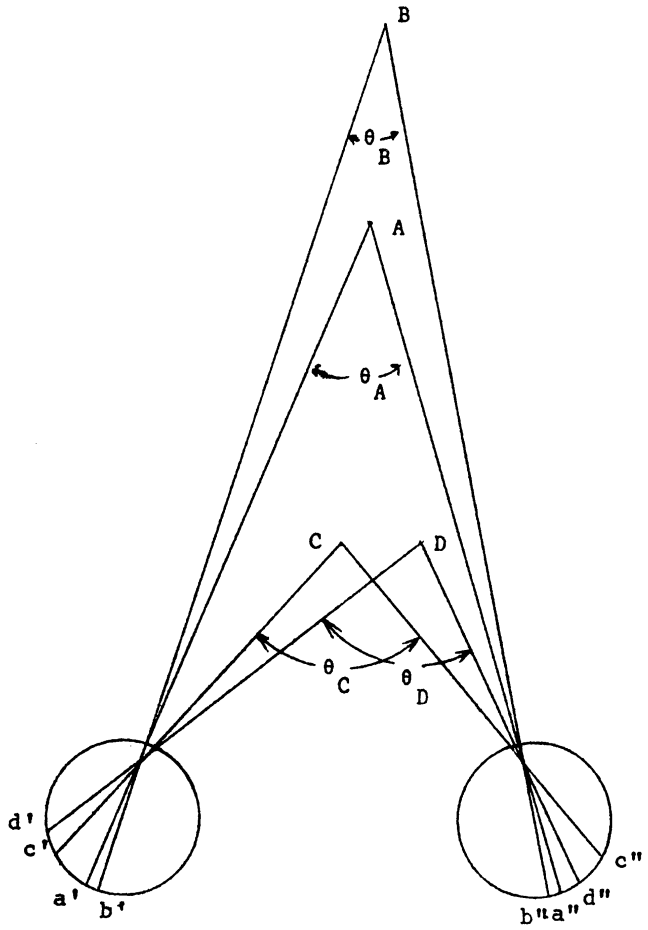


FIGURE 5-35

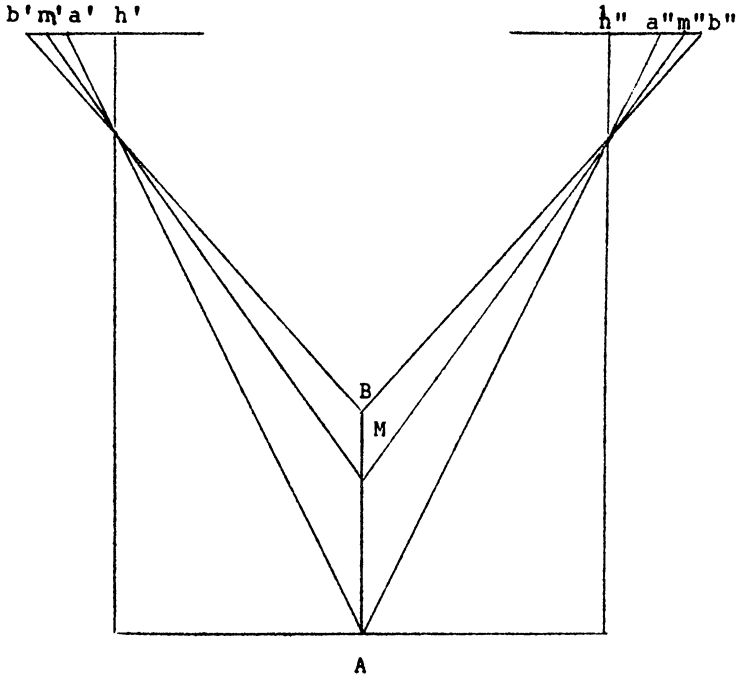


FIGURE 5-36

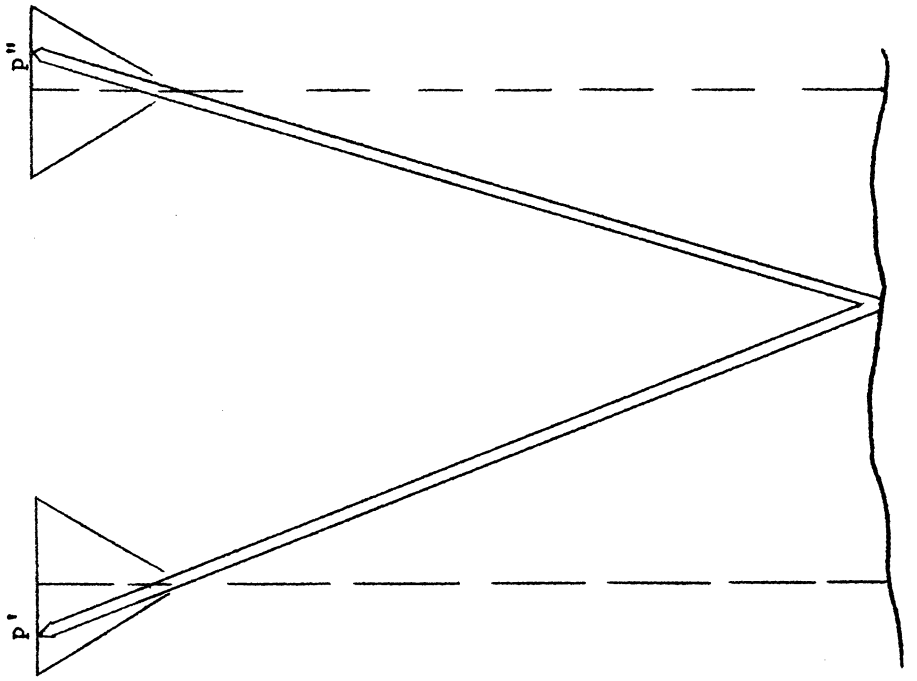


FIGURE 5-38

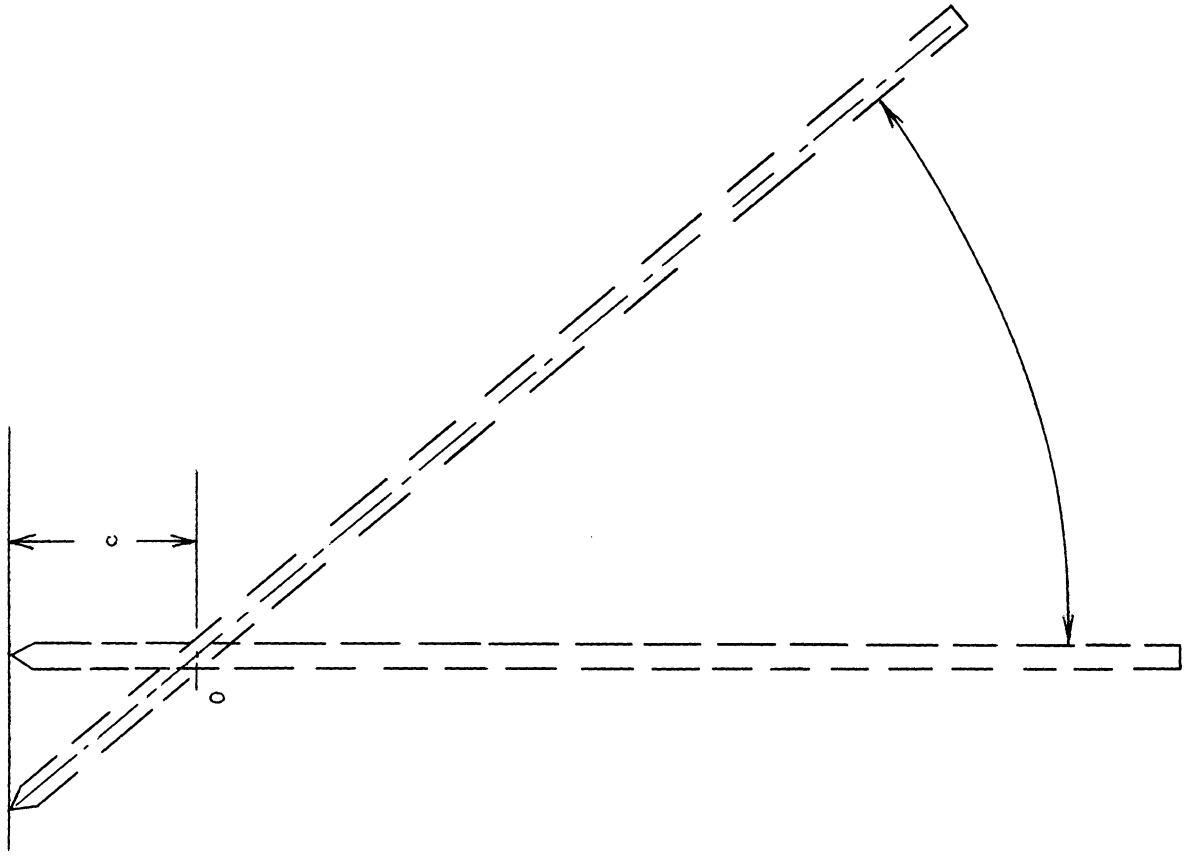
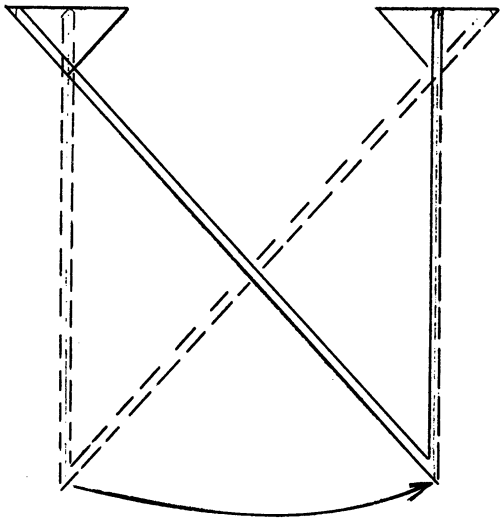
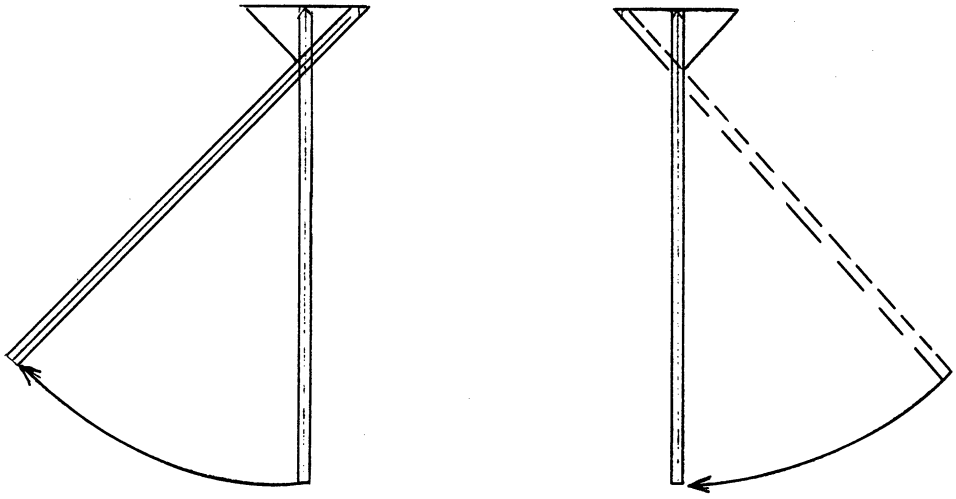


FIGURE 5-37



(a)



(b)

FIGURE 5-39

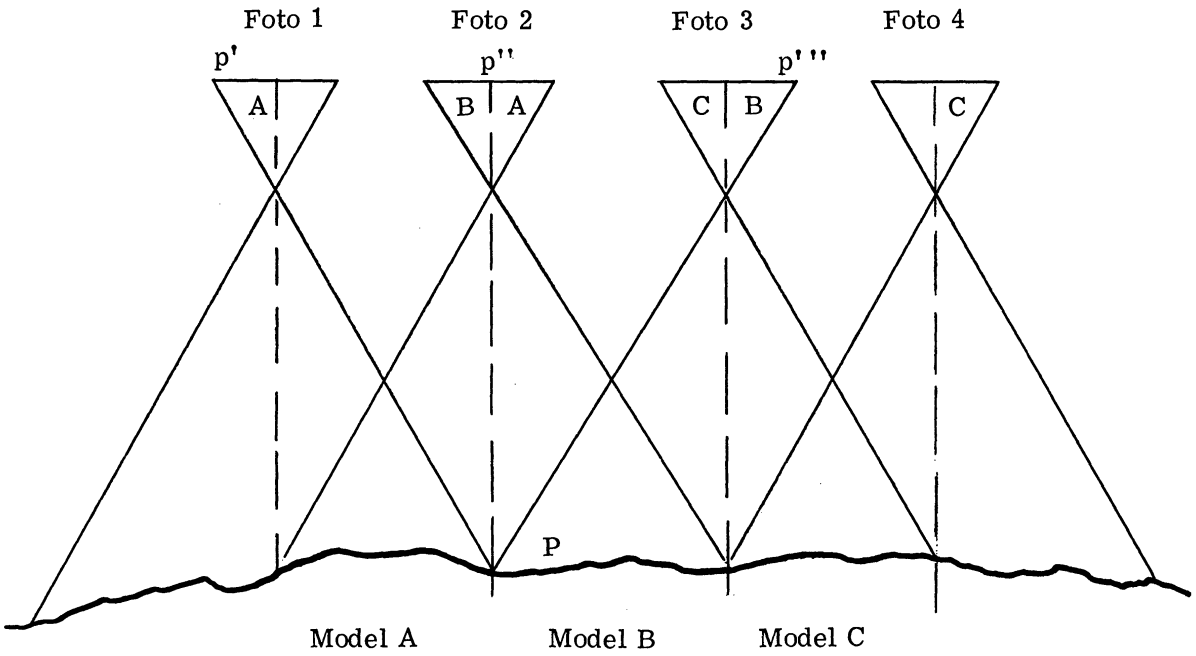


FIGURE 5-40

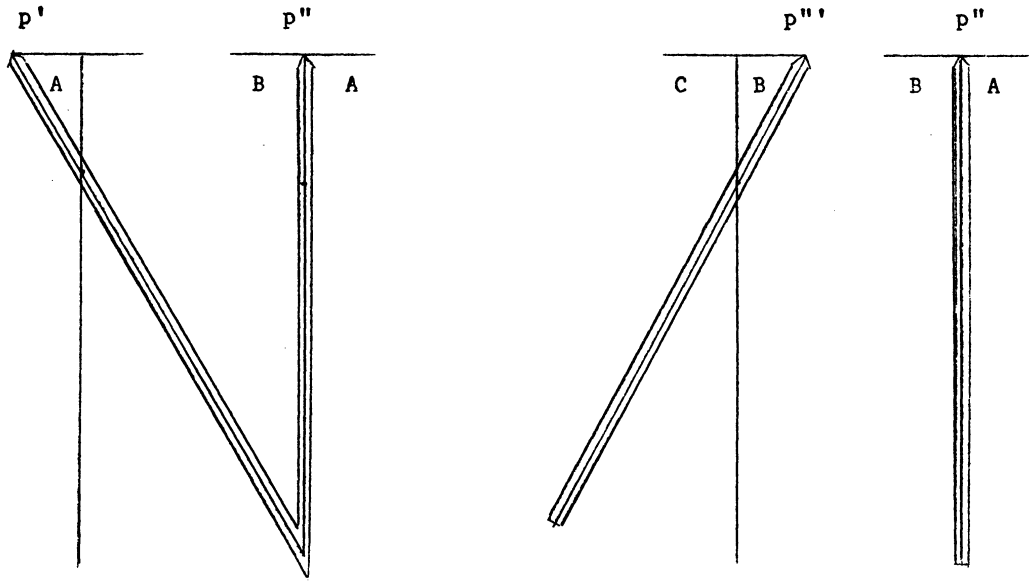


FIGURE 5-41

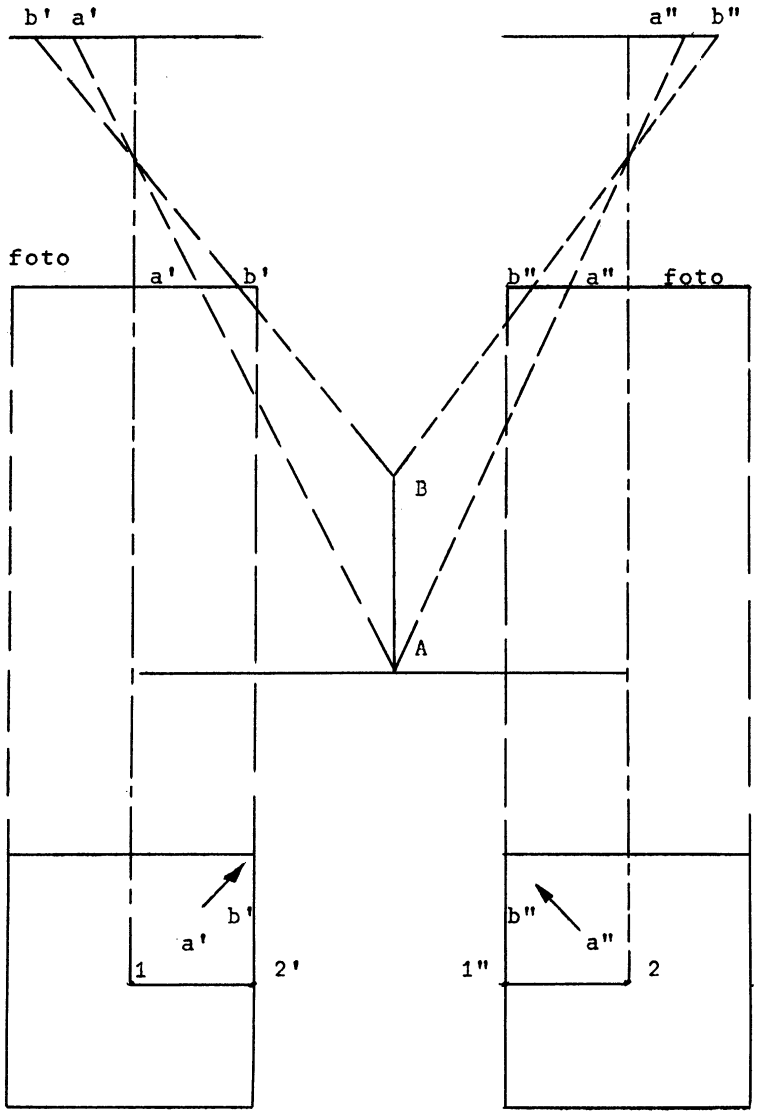


FIGURE 5-42

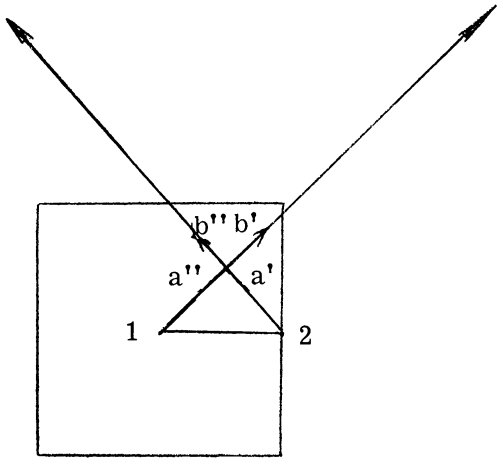


FIGURE 5-43

SECTION IV
TELEVISION AND MAGNETIC RECORDING

1. FUNDAMENTALS OF TELEVISION*

1.1. INTRODUCTION TO TELEVISION

The fundamental aim of a television system is to extend the sense of sight beyond its natural limits. A system to perform this function uses devices patterned after the principles of operation of the human eye. This fact must be kept in mind whenever television is to be used as a remote sensing aid. Many of the techniques employed and the compromises used are based on the properties of human vision, and other devices connected to the television system may not produce the desired responses. The specific meaning of this statement will become clearer as we proceed to explain how a television system works.

Keep in mind that the purpose of a television system is to arouse the same visual sensations that an observer would experience if he were physically present at the scene where the camera is located. It is not possible to understand the theory and operation of such a system without taking into account both the physical processes by means of which light energy is modified to produce images and the psychological processes by means of which the human observer interprets images. Thus, a television system calls upon the science of physics for knowledge about producing the images in electrical form and the transmission and reproduction of the images, while the science of psychology gives us information as to persistence of vision, the response of human sensory systems to illusion, and the importance of visual acuity and flicker. We will talk more about each of these subjects as we proceed.

1.2. ASPECTS OF DIRECT VISION

When the human eye views a scene directly, the information conveyed from the scene to the observers mind is based on five characteristics:

1. the relative brightness of the areas,
2. the geometric structure, that is the relative sizes, shapes, boundaries, and positions of objects,
3. motion of the geometric structure,
4. the colors of the areas, and
5. their apparent positions along the line of sight.

---

*These lecture notes by R. H. McLaughlin were only slightly modified from the first draft form.

Another way of describing these five characteristics would be by the five terms:

1. the tonal content,
2. the structural content,
3. the kinematic content,
4. the chromatic content, and
5. the stereoscopic content.

These five aspects of direct vision have been listed in order of importance. For example, the first two serve to convey all the information in a monochrome photographic print or printed halftone. When a scene involving motion is to be represented the kinematic content must be displayed. This factor has been introduced in motion pictures and television. The fourth element, color, may be dispensed with for many purposes, but it is an essential part of any visual reproduction process which aims to recreate the experience of direct observation. The stereoscopic content is almost always dispensed with but when available it lends the final touch of realism to the reproduced scene.

1.3. VISUAL REPRESENTATION

Let's move now into a discussion of the methods of visual representation. First of all, a few remarks about illusions are in order. For instance, all the arts involving two-dimensional images including photography, painting, and television are based upon illusions. In other words, the representation of the scene is presented without the stereoscope content. The degree to which the three-dimensional illusion is created is a measure of the success of the representation. As a further example, take this simple sketch shown in figure 1-1. To someone familiar with metropolitan life in the United States, the representation is clearly the figure of a man sitting in an easy chair viewing a television set. However, a person residing in Central Africa who had never seen a television set, an easy chair, or even a pair of shoes would see the image for what it really is, nothing more than a single irregular line drawn on a sheet of paper. It should be kept in mind that sketchy, psychological clues often control perceptions. For example, in very low levels of illumination familiar objects in a room may be recognized by their barest outlines.

1.4. PICTURE ELEMENTS

As an introduction to the methods employed in television transmission let's take a look at what's done in photography and photoengraving. These methods differ in many essentials from television but they serve to introduce two important concepts, namely, picture element and

image repetition. When photographic or printed images are examined closely they are found to consist of elemental areas of light and shade corresponding to the illuminated areas of the original scene. Figure 1-2 is an example of photo-engraving which has been greatly enlarged. The printed scene is composed of many fine black dots displayed against a white background. The dots are all equally black but they differ in size and separation, one from the other, and are thus able to represent a variety of tonal values and details. The printed dots are examples of picture elements. A picture element is an elemental area of an image, having a uniform total value within itself, corresponding in size to the smallest detail which could be represented in the image. Professional 35 mm movie frames display about one million picture elements, 16 mm home movies about 200,000 elements, and 8 mm home movies about 50,000 elements. Figure 1-3 is another example of the dot structure in the highlights, middle tones and shadows of a halftone photoengraving.

1.5. IMAGE REPETITION

The concept of image repetition arises when motion must be detected in the reproduction of a scene. A motion picture film consists of a great many static pictures, each of which remains stationary in the projector for a brief time during which it is revealed to the eye. Between successive projections, the light is cut off by a shutter in the projector, and the next frame is moved into position. If the rate of image repetition is high enough, persistence of vision causes the sensation of light in the mind of the observer to carry over from one frame to the next. If the rate of repetition is high enough a continuous displacement due to motion in the original scene is reproduced by small discrete displacements which are not separately perceived by the eye. Thus the process of image repetition brings two problems:

1. the light from the image is, in fact, discontinuous, and flicker may be evident if the rate of repetition is too low,
2. the motion represented by repeated images is not continuous, and the discontinuity may become evident if insufficient samples of the motion are taken.

In motion pictures the basic frame repetition rate is 24 frames per second; a value chosen as a compromise between the consumption of film on the one hand, and the necessity of minimizing flicker and jerky motion on the other. This rate is not high enough to secure freedom from flicker even at low brightness levels, so the rate is increased to 48 per second by interrupting the light briefly during the projection of each frame. A similar technique known as interlacing is employed in television to double the effective frame rate.

1.6. ELECTRICAL TRANSMISSION OF AN IMAGE

With these ideas in mind, let's see now how an image may be transmitted electrically. Since, as we have seen, a picture consists of a great many separate elements which are perceived simultaneously by the eye, we might replace the sensitive elements of the eye with a large array of transducers which would convert the tonal shadings of the scene into appropriate voltage levels. Each of these transducers or sensors could then be connected by a separate pair of wires to a source of illumination at the receiving end. The source of illumination would then be controlled in brightness by the voltage coming from its associated transducer at the transmitting end. This method of transmitting visual information has been experimented with, but has proved practical only when the number of picture elements in the picture is small, and when the distance of transmission is short enough to make feasible the great number of separate circuits required. Animated electric signs in which lamp bulbs are used for the picture elements, each wired to a separate circuit, are to date the only practical application of this method of transmission. For pictures containing 100,000 picture elements or more, typical of today's modern television practice, it is impractical to provide the necessary number of individual circuits for this method of transmission.

Another possibility, again assuming the picture is subdivided into a finite number of elements, is to assign a different frequency to each element in such a way that the frequency characterizes the position of the element, and the amplitude of the component of frequency gives the brightness of that element. This might be accomplished in principle by constructing n oscillators each tuned to a different frequency, and each controlled in amplitude by the output of a photocell located at one of the n picture elements of the image area (see fig. 1-4). The n signals of different frequency are combined to form a single complex wave output which is transmitted to the receiver. A receiver consists of n tuned circuits to match the oscillators. Each tuned circuit includes a detector which controls one light source forming a picture element of the viewing area. In this way, the brightness of every picture element of the viewing screen reproduces the brightness of its counterpart at the image field of the camera.

A better way of employing a single communication channel is to send the picture element impulses one after another or sequentially at a rapid rate. This sequential method of transmission is used universally in television. The effectiveness of this method rests on that property of the eye known as persistence of vision. If the eye were instantaneous in its operation a sequential method would fail since the eye of the observer would then see each picture element individually and separately. Actually, however, the impression made by any one element persists in the mind for a small fraction of a second. During the interval of persistence of this

one element, all the others are presented successively to the eye in their proper positions. In other words the eye operates as though it were seeing all the elements at once, and the simultaneous aspect of direct vision is recreated artificially. An example of sequential transmission by scanning is shown in figure 1-5.

The persistence of vision is a complicated phenomena. The persistence of the sensation of light is very definite and marked for a short period after the light stimulus is removed, after which the sensation gradually dies away. Experience with motion pictures reveals that the projected image appears continuously illuminated if the individual light impulses from the projector are presented to the eye at the rate of 40-50 per second. If they are presented at a slower rate, the sensation decays to a low level between the light impulses, and flicker is apparent at the usual brightness levels required in theatres or halls. To avoid flicker in television it has been found desirable to transmit all the elements of the picture in approximately 1/30th of a second. It should be noted in passing that the perception of flicker increases with brightness level. It is therefore necessary to increase the repetition rate of the picture in order to reproduce very bright images. Typical brightness levels of television receivers are higher than those found in theatres since the ambient light cannot conveniently be reduced as much. This is one of the main contributing factors in the fact that the picture repetition rate in television is higher than that normally used in motion picture projection.

It is obviously necessary in the sequential method of transmission to select the picture elements from the scene in some orderly sequence, and to reassemble them in the image in the same sequence. Any convenient pattern of selection would suffice provided the same pattern were used at the transmitter and receiver. Many such patterns have been proposed such as radial scanning, spiral scanning, or sinusoidal scanning to name a few. The method universally used in practice is known as uniform linear scanning. Figure 1-6 shows the basic elements of this type of transmission. The picture to be transmitted is swept from left to right and from top to bottom in the manner of reading words on a printed page. A more detailed scanning sequence is shown in figure 1-7. When the transmission of the picture begins, all the picture elements present in the top most row are selected, one after another from left to right, and converted into corresponding electrical signals which are sent successively over the communication channel. When the first row has been scanned the picture elements in the third row, which lies parallel to, and one row removed, from the first row, are selected in the same fashion followed by the elements in the fifth row, and so on, until the bottom row of the picture is reached. Then the area is covered again as the elements are selected from the 2nd, 4th, 6th rows, and so on. In this manner the whole area of the picture is systematically explored

for the information it contains, in two sets of parallel lines. At the receiver, the electrical impulses are translated back again to light impulses, and these impulses are assembled before the eye in the same scanning sequence.

If the eye is to see the whole picture at once, it is necessary that the last picture element in the last row be presented to the eye while the impression from the first element in the first row still persists. In other words, the entire picture must be transmitted within approximately 1/30th of a second. Within this short time it is necessary that all the elements of the picture be presented so that the eye sees the picture all at once. This process is then immediately repeated, and a new picture, perhaps differing somewhat from the first, is sent in the same fashion. Thus at the end of one second, 30 complete pictures have been sent. Any motion which occurs in the scene is thereby divided into 30 smaller motions which appear to blend smoothly and continuously one into the next.

It will be noticed in the above description that the picture is scanned in two sets of parallel lines. That is, the odd numbered lines, followed by the intervening even numbered lines. This method of scanning presents a picture which is divided into two intervals similar to that used in motion pictures. The picture repetition rate is thus effectively doubled, as the picture information remains unchanged. This is called interlace scanning, and is now universally employed in the television systems throughout the world.

1.7. ANALYSIS OF IMAGES

Let's return now to the essential aspects of visual images, and discuss how each essential factor is dealt with. As a vehicle for this discussion, and in order to explain the development of the television system in quantitative terms, we will use the American Standard Broadcast System. This is similar, though different in some of its essential aspects, to other standard systems used throughout the world. However, the application of the principles just discussed would be similar in other systems. It will be convenient to discuss the essential features of visual images, not as we listed them before (in the order of their relative importance), but rather we will list them now in the order in which the picture structure may be logically developed. This list then becomes: the gross structure of the image, the fine structure, the continuity, and the tonal gradation.

1.8. ASPECT RATIO

The frame adopted for television images is a rectangle having its long dimension in the horizontal direction. There are many reasons for this choice. This rectangular form has been

used for centuries in the framing of works of art, with the proportions of width to height being selected to fit the subject. When the frame contains a variety of scenes in motion, the preference is for a rectangle wider than it is high. This is the form universally used for framing stages and motion picture screens. This ratio is chosen because in human affairs motion occurs primarily in the horizontal plane.

Two additional technical reasons support the choice of a rectangular frame. It is simpler to scan an image in parallel lines of equal length, and a large part of the program material is motion picture film. The sole disadvantage of the rectangular frame is the fact that the screens of some television picture tubes which might be used are circular in shape, and hence are not fully utilized by a rectangular image.

The frame dimensions in television are described by the aspect ratio which is the ratio of the frame width w, to its height h. This ratio has the standard value given by equation 1.

$$w/h = 4/3 = 1.333 \tag{1}$$

1.9. GEOMETRIC FORM

The second factor in the gross structure of a television image is the delineation of geometric form. It is obviously necessary that a square object viewed by a television camera be reproduced as precisely as possible as a square image, and similarly for other shapes. Scanning patterns at the transmitter and receiver which are precisely congruent and synchronous, satisfy this requirement. In practice this requirement is difficult to meet. It requires precise synchronization between the transmitter and receiver but with care in design it may be achieved.

1.10. RESOLUTION

Moving now to the fine structure, we will investigate the resolving power, or the resolution, of an image producing system. Resolution may be measured by counting the maximum number of adjacent parallel lines which may be resolved in the image, within a specified distance. In television, the structure of the image along each scanning line is fundamentally different from the structure at right angles to the lines. Therefore, it is necessary to examine separately, the vertical resolution and the horizontal resolution. By vertical resolution is meant the number of television resolution lines which can be resolved in the height of the frame. By horizontal resolution is meant the number of television resolution lines which can be resolved at right angles to the scanning line, measured in a distance equal to the height of the frame. If we take a frame, for example, (fig. 1-8) which has 30 vertical lines, and 20 horizontal lines measured

in a distance equal to the height of the frame, and this is considered to be the limit of the resolution, we may then say the following:

r_V = vertical resolution
= 30 lines
r_h = horizontal resolution
= 20 lines
A = aspect ratio
= 4/3
N = number of picture elements
$= A\, r_V\, r_h = 4/3 \times 30 \times 20 = 800$ (2)

This number is the figure of merit of the fine structure of the image.

1.10.1. VERTICAL RESOLUTION. Let's consider first the vertical resolution of the image. This is the maximum number of separately discernible horizontal white and black resolution lines which can be accommodated in the height of the picture. This figure is dependent on the number of scanning lines in the scanning pattern, since one line can represent, at best, one resolution line. However, not all the scanning lines are capable of representing separate resolution lines. In fact, in the worst case, the scanning lines may fail to reproduce any resolution line whatever. Consider for example, figure 1-9. As we can see the object represented is a vertical bar consisting of alternate white and black segments, the height of which are just equal to the thickness of one scanning line. If the white and black segments are wholly contained, each separately, within the width of one scanning line they are fully resolved. If, however, the scanning lines straddle the picture elements, the reproduction may lack detail completely. Experience has shown that on the average about 70% of the picture elements in the vertical dimension are separately reproduced. The vertical resolution, r_V, is therefore equal to a constant times the number of active scanning lines n_a as shown by equation 3

$$r_V = k n_a \tag{3}$$

1.10.2. HORIZONTAL RESOLUTION. Let's consider next the horizontal resolution of television images. This is the number of picture elements resolved along the scanning line. Consider figure 1-10. It may be seen that the horizontal resolution is determined by the maximum number of changes in voltage which can occur as the line is scanned. The rapidity with which the voltage can change is in turn a function of the maximum frequency which can be

transmitted. In other words the bandwidth of the channel over which the television signal is transmitted directly determines the horizontal resolution.

1.11. UNIFORM LINEAR SCANNING

It has been stated that television systems use uniform linear scanning. Let's take a little more detailed look at this method of scanning in order to determine some significant parameters which will be needed in determining the required communication channel bandwidth. The method of scanning the picture has been briefly mentioned before and was described as follows: The image is scanned beginning at the upper left-hand corner and proceeding from left to right across the top line of picture elements. The beam then is quickly returned to the left-hand side of the image, and proceeds once more from left to right scanning, in this instance, a row of elements parallel to, and displaced one line width downward from, the first line (fig. 1-7). The signal which is used to generate the linear motion across the picture is a sawtooth, having one linear velocity from left to right and a more rapid linear velocity from right to left, this portion being called the retrace (fig. 1-11). The simplest manner of generating the vertical motion is to use a second lower frequency sawtooth which is continually moving the beam down, even as it traverses the image from left to right. This causes the scanning lines to have a downward slope as they scan the picture elements. Both the horizontal and vertical sawtooth wave signals require a retrace interval which is a percentage of the total time period of the sawtooth waveform. The American Television Standards have set up 83% as the active scanning period of the horizontal sweep, and 93.5% as the active period of the vertical sweep (fig. 1-12). Let's now define the following symbols; let k_h be the fraction of the total line scanning period actively devoted for scanning the picture element. This will be as stated 83%. Let's define k_V as the fraction of the scanning field time during which the scanning lines are active and this figure will be equal to 93.5%.

1.12. RELATIONSHIP BETWEEN BANDWIDTH AND RESOLUTION

The next step in setting up a system of television standards is to determine the relationship between the bandwidth and the resolution. Let's begin with the bandwidth of the channel which has been established by the Federal Communications Commission, and with the frame rate which we have determined from considerations of flicker. The width of the communication channel established by the FCC is 6 megahertz and the portion of the channel available for the picture sideband is 4.25 megahertz. The frame rate we recall, has been established at 30 per second. A further implicit assumption can be made; that we wish to provide nearly equal resolution in the vertical and horizontal direction.

Computation of the Relationship Between
Bandwidth and Resolution

n = number of lines per frame

f = frame rate

$V = 1/2f$ = field period (secs)

nf = line scanning frequency (Hz)

$H = 1/nf$ = line scanning period (secs)

$H_a = k_h/nf$ = active duration of each scanning line (secs)

f_{max} = maximum frequency

$2f_{max}$ = number of picture elements per second

N_ℓ = number of picture elements per line

$$N_\ell = \frac{2f_{max}k_h}{nf} \text{ (number of cycles of } f_{max} \text{ per line)} \tag{4}$$

Then horizontal resolution is $N_\ell \times$ ratio $\frac{h}{w}$:

$$r_h = \left(\frac{h}{w}\right)\frac{2f_{max}k_h}{nf} \quad \text{since horizontal resolution is defined as number of elements in a width equal to the height} \tag{5}$$

The vertical resolution r_V is the number of lines, n, multiplied by the active percentage, k_V, and also by the vertical resolution factor:

$$r_V = k_V k n \tag{6}$$

For flexibility in design let the horizontal resolution exceed the vertical resolution by the ratio m. Then

$$r_h = m r_V \tag{7}$$

Now substituting (6) and (7) into (5)

$$m k_V k n = \left(\frac{h}{w}\right)\frac{2f_{max}k_h}{nf} \tag{8}$$

Solving for f_{max}:

$$f_{max} = \frac{m n^2 f k_V k}{2k_h}\left(\frac{w}{h}\right) = 1/2\ m n^2 f\left(\frac{k_V}{k_h}\right)\left(\frac{w}{h}\right) \tag{9}$$

when $K = 0.7$, $f = 30$, $w/h = 4/3$, and $\frac{k_V}{k_n} = \frac{93.5}{83}$, we have

$$f_{max} = 15.8\, m n^2 \tag{10}$$

for $m = 1$, and $f_{max} = 4.25 \times 10^6$ cps

$n = 519$ lines

Since, for interlaced scanning we must have an odd number of lines, and n must be composed of small odd factors for ease of generation of sync pulses, n is chosen as 525 lines ($525 = 3 \times 5 \times 5 \times 7$). Then with $f_{max} = 4$ mc

$$m = 0.92$$

1.13. COLOR TELEVISION

In studying television to this point we have considered the tonal content, the structural content, and the kinematic content of an image which is to be transmitted and reproduced. The next element on the list was the chromatic content and this will be discussed in this section. Frequent reference is made to the National Association of Broadcasters Engineering Handbook.

The chromatic or color content of an image is a combination of those properties of light which control the visual sensations known as brightness, hue, and saturation. The monochrome* television system transmits these three sensations combined as a value of brightness. In a color system, the brightness characteristic still controls that aspect of the color which places it on a scale ranging from dark to light. The second variable of color, hue, is that characteristic which permits a color to be described as red, yellow, green, or blue. Saturation refers to the extent which the color departs from the white or neutral conditions. Pastel colors for example, are low in saturation while very vivid colors are high in saturation.

1.13.1. PRIMARY COLOR REPRESENTATION.

In order to transmit and reproduce a color image we must add two additional characteristics to the one which is sent in the monochrome television system. One of the techniques employed to convey the necessary additional information is by making use of the so-called primary color process.

Experiments have shown that almost any color can be matched by the proper combination of no more than three primary colors. This is shown in figure 1-9 in the NAB (National Association of Broadcasters) Engineering Handbook, 5th Edition. The three colors used in color television are the primaries red, green, and blue. These are shown in the outer portions of the three large circles in the figure. In the center of the figure, where the three circles intertwine, the combination of the three colors produces the color white. In the area where the red and

*Monochrome— literally, one color: in other words, black-and-white television.

green circles overlap, yellow is produced, and in the area where the red and blue circles overlap, purple is produced. In the remaining portion of the picture where the green and blue circles overlap, bluegreen or cyan is the color produced. These are examples of the colors which may be produced by using combinations of the primaries mixed in equal proportions. By the use of dichroic mirrors it is a simple matter to separate the image at the camera into its three primary colors. One such type of mirror is a blue reflecting dichroic mirror. This mirror will reflect blue light while transmitting all other colors. An example of this type of color camera is shown as figure 1-11 on page 5-9 in the NAB Handbook. The basic procedure here is to use three television camera tubes all scanned in synchronism to produce three separate images corresponding to the red, blue, and green components of the original scene. Figure 1-10 of the NAB Handbook shows an object depicted in various saturated colors with the red counterpart shown in the upper right hand, the green counterpart in the lower left hand, and the blue counterpart in the lower right hand portion. Notice, for example, that the doorway which is white has equal components in the red, green, and blue portions of the signal. The yellow shutters have red components and green components, but no blue components. Figure 2-6 of the Handbook shows a single scanning line through the center of the same image with the amplitudes of the various color components displayed. Notice again that the white doorway is represented by full amplitude in each of the three signals whereas for example, the red chimney shows full amplitude in the red signal and zero amplitude in the green and blue signals. Thus we see that we have separated the colors into the red, green, and blue counterparts where they may now be processed for transmission to the receiver.

1.13.2. ENCODING AND TRANSMISSION. One possible method of approach is to simply transmit the red, green, and blue signals to the receiver where they may be combined using a three color kinescope for picture reproduction. The transmission may be accomplished either on three separate channels, or by frequency multiplexing onto a single channel. The three color kinescope is basically a cathode-ray tube containing three electron guns and three phosphors on the screen, one for each of the three primary colors. The guns are so arranged that the red gun illuminates only the red phosphor particles, and so on for each of the three guns and phosphor systems. A sketch of this device is shown as figure 1-12 on page 5-9 of the NAB Handbook.

Two difficulties to this fairly straight forward system appear, however: (1) it is desirable and, in fact, necessary that a color television system which is introduced after a monochrome television system has been in use, be compatible with the monochrome television system, and (2) simply transmitting the three primary colors would take basically three times the bandwidth of a monochrome television system. A compatible color television system is one in

which color transmissions may be received on ordinary monochrome television receivers without modification or special equipment. It is also desirable, if possible, to transmit the color television signals over the same channel which was used originally for monochrome transmission. We recall, that a channel 6 megahertz wide has been allocated for television transmission. This appears as in figure 2-14 on page 5-15 in the Handbook. What we desire to do is to transmit three signals in the bandwidth normally occupied by a single signal. The manner in which both of these objectives are realized will now be briefly described.

The first step in the process consists in generating three new signals from the original red, green, and blue components. The three new signals are called the M signal, I signal, and the Q signal. The M signal or luminance component corresponds very closely to a signal produced by a monochrome camera, and is therefore capable of rendering excellent service to monochrome television receivers. This is one of the principal aspects in obtaining the feature which we have called compatibility. The additional information required for full reproduction of color images is carried by the I and Q signals which are called the chrominance components. We will not take the time here to discuss the specific method by means of which these signals are generated, but the technique using resistance matrices and phase inverters is well described in the NAB Handbook for those of you who are interested in pursuing the subject further.

A second technique employed in a color television system is to take advantage of the fact that the eye has substantially less acuity in depicting variations of color than it has for resolving differences in brightness. This important characteristic of the eye is considered in setting up the composition of the I and Q signals. A bandwidth of approximately 1.5 megahertz was found satisfactory for the I signal which corresponds to color differences in the range from orange to bluegreen. For the color differences in the range from green to purple represented by the Q signal the eye has even less acuity and a bandwidth of only 0.5 megahertz is required. The M signal, as before, conveys the fine details and must be transmitted with the standard 4 megahertz bandwidth. Two additional techniques are employed to permit transmitting the full color television signal over the original allocated bandwidth. These techniques are called two phase modulation and frequency interlace. Time does not permit a complete discussion of these techniques, but again they are described in the reprint handout. In very general terms two-phase modulation permits two signals to be modulated upon a common carrier, and to be separated into the original two signals at the receiving end. Frequency interlace is a technique which permits adding several components of the chrominance and monochrome signals together without causing objectionable mutual interference. The frequency interlace technique is based upon two factors:

1. a precise choice of the color subcarrier frequency, and
2. the familiar persistence of vision effect.

The color subcarrier is made an odd multiple of one-half the line frequency, and therefore its apparent polarity reverses between successive scans of the same picture area. Since the eye responds to the average stimulation after two or more scans the interlace effect of the carrier signal tends to be self cancelling owing to the periodic polarity reversals. The frequency interlace technique is illustrated in figure 2-17 on page 5-16 in the NAB Handbook. This very brief description of the elements of the color television system is not intended to be comprehensive, but only to serve as a general introduction to the subject. The techniques and processes which have been only briefly mentioned here are described in the Handbook.

1.14. PROBLEMS

1. Compute the vertical resolution, horizontal resolution, resolution ratio, and number of picture elements per frame for the British Television System. The standards are n = 405 lines, f = 25 per second, w/h = 4/3, f_{max} = 2.75 mc, k_V = 0.95, k_h = 0.85, k = 0.7.
2. Compute the number of lines which would give equal vertical and horizontal resolution in Problem 1.
3. List the total number of lines per field, active lines per field, the number of fields per second, the number of lines per second, and the active duration of each line, using the standards in Problem 1.

1.15. PROBLEM SOLUTIONS

1. $r_V = n k_V k$

$= 405 \times 0.95 \times 0.7$

$= 270$ lines

$$r_h = \frac{h}{w}\frac{2f_{max}k_h}{nf}$$

$$= 3/4 \times \frac{2 \times 2.75 \times 10^6 \times 0.85}{405 \times 25}$$

$= 347$ lines

$r_h = m r_V$; $m = \frac{r_h}{r_V}$

$$m = \frac{347}{270} = 1.28$$

$N = 4/3 \times 70 \times 347 = 125{,}000$ elements per frame

2. Let m = 1

$$f_{max} = 1/2\ k\, m\, n^2 f(w/h)\ (k_V/k_h)$$

$$n^2 = \frac{2 \times 2.75 \times 10^6}{0.7 \times 25} \times 3/4 \times \frac{0.85}{0.95}$$

$$n^2 = 2.10 \times 10^5 = 21 \times 10^4$$

$$n = \sqrt{21 \times 10^4} = 4.6 \times 10^2 \cong 460 \text{ lines}$$

3. $\frac{405}{2} = 202\ 1/2$ lines per field

$202\ 1/2 \times 0.95 = 192\ 1/2$ active lines per field

$25 \times 2 = 50$ number of fields per second

or $\left.\begin{matrix} 202\ 1/2 \times 50 \\ 405 \times 25 \end{matrix}\right\} = 10{,}125$ cps number of lines per second

$$\frac{1}{10{,}125} \times 0.85 = 83.8\ \mu\text{secs}$$

TABLE 1-I. SCANNING SPECIFICATIONS,
FCC BLACK-AND-WHITE STANDARDS

$A = w/h = 4/3$

$n = 525$ lines

$n_a = 491 \pm 8$ lines

$H = 1/15{,}750 = 63.5\ \mu\text{secs}$

$V = 1/60 = 16{,}667\ \mu\text{secs}$

$H_a = (0.82 + 0.01 - 0.00)\ H = (52.1 + 0.6 - 0.0)\ \mu\text{secs}$

$V_a = (0.95 + 0.00 - 0.03)\ V = (15{,}820 + 0 - 500)\ \mu\text{secs}$

2. MAGNETIC RECORDING*

2.1. THE THEORY OF MAGNETISM

2.1.1. INTRODUCTION. The phenomenon of magnetism was studied in connection with the so-called permanent magnets for hundreds of years before the relationship between electricity and magnetism was established by Oersted and Faraday in the early 19th Century. In dealing with permanent magnets it was natural that the concept of poles should be made the starting point, since, seemingly only at the poles could magnets manifest their properties. Consequently, it was necessary to evaluate the strength of a magnet in terms of the strength of its poles. When the ability of an electric current to generate a magnetic field was recognized, the next logical step was to describe this field in terms of an equivalent pole strength. This terminology has persisted to the present day. The nomenclature of poles, and the definition of field strength are now firmly entrenched.

2.1.2. POLES AND MAGNETIC FIELD STRENGTH. If a bar magnet is suspended so as to be free to rotate, and if there are no other masses of iron or other magnets in the vicinity, the magnet will take up a position that is roughly north and south. The end of the magnet pointing in the northerly direction is called the north seeking pole or simply the north pole. The other end is designated as the south pole. If the north pole of one magnet is brought near the south pole of another magnet it is found that they attract each other. If the two north poles or two south poles are brought together they repel each other. In other words, like poles repel and unlike poles attract.

Although the centers of attraction are never actually points, good approximations may be obtained by using very long slender magnets. When this is done it is found that the forces exerted between the poles obey the inverse square law. That is, the force exerted by one pole upon another, is inversely proportional to the square of the distance separating the poles. To make possible the evaluation of the relative strength of magnets, a so-called unit pole is defined. This unit is an imaginary concept, and is defined to be an isolated pole of such strength that if separated from an equal and exactly similar pole by a distance of 1 centimeter in vacuum it repels it with a force of 1 dyne.

This concept is imaginary since an isolated pole has never been observed. If an attempt is made to obtain an isolated pole by breaking off one end of a bar magnet, it is always found that

*These lecture notes by R. H. McLaughlin were only slightly modified from the first draft form.

two new poles come into existence at the point of division. Each of the two parts of the original magnet now possess two poles. Furthermore, the new poles are of such polarity that a complete magnet is formed from each of the separate pieces, each having a north pole and a south pole (fig. 2-1). When two pole strengths are measured in terms of unit poles the force that will be exerted by one pole on the other is given by

$$F = \frac{m_1 m_2}{r^2} \tag{1}$$

m_1 and m_2 = pole strengths (unit poles)
r = distance between in cm
F = force in dynes

2.1.3. MAGNETIC INDUCTION. At any point in a region where magnetic forces would be exerted on a pole located at that point, a magnetic induction is said to exist. This induction is measured as the force in dynes that would be exerted on a unit north pole at that point, and it requires both direction and magnitude for its complete description. Any region throughout which such magnetic induction exists is said to contain a magnetic field. Figure 2-2 is a picture of such a magnetic field constructed by placing a piece of paper atop a bar magnet and sprinkling the paper with iron filings. When the paper is gently vibrated the iron filings configure themselves along the lines of magnetic force between the north and south poles of the bar magnet. It is seen that the filings appear to lie along curved lines joining the two poles of the magnets. These lines are dense in the vicinity of the poles and relatively sparse near the center.

2.1.4. LINES OF FLUX. This appearance suggests a very useful mathematical device for describing such a field by so-called lines of flux. These lines of flux are defined as lines drawn in a magnetic field in such a manner that their direction at any point is the same as the field at that point, and in such a manner that their number per unit area perpendicular to the direction of the field at any point is equal to the strength of the field at that point. One flux line is called one Maxwell. Induction is then measured in Maxwell's per square centimeter or gausses and is designated by B.

In accordance with their definition flux lines arise on north poles and terminate on south poles. For an isolated pole in free space these lines would issue radially from the pole and the number per square centimeter passing through a sphere of any radius centered on the pole would give the field strength due to the pole at that distance.

If the region around a long straight conductor that is carrying an electric current is explored with a small compass needle it is found to contain a magnetic field. The flux lines form concentric circles with a conductor as a center as shown in figure 2-3. If a conductor is wound in the shape of a long coil, the field lines are plotted as seen in figure 2-4. The flux lines surrounding each segment of the conductor reinforce each other in the direction of the axis of the coil, and create a field which is substantially uniform in the center of the coil, and which external to the coil bears a marked resemblance to that of the bar magnet. Also, it may be noticed, that the flux lines are continuous; that is, each flux line outside the coil is a part of a closed loop which closes through the inside of the coil.

It can be shown mathematically, that the flux lines around any current carrying conductor, of any shape, must always be closed loops. This is stated as $\nabla \cdot B = \text{div } B = 0$; that is there are no sources or sinks for lines of flux, and hence they must be continuous closed loops. In addition, the flux lines are always closed in a field produced by a bar magnet. The fact that the flux lines emerging from a bar magnet fan out as they leave the end, gives them the appearance of having originated from a pole. Actually, it may now be seen that a pole is in reality nothing but a region in a magnetic field in which flux lines converge or diverge. This makes it clear why a magnet always has two poles no matter into how many pieces it is broken.

2.1.5. ELECTROMAGNETIC INDUCTION. As electromotive forces produce currents, and currents produce magnetic fields, so magnetic fields themselves can be made to produce electromotive forces. If a bar magnet is thrust into a coil of wire, a momentary voltage will appear across the terminals of the coil. If the conductor is moved through a magnetic field, thus cutting the flux lines composing the field, a voltage will appear across the ends of the conductor. Other similar situations are possible but this serves to illustrate one common element which must be present; that is change in the magnetic field. This change may occur in either of two ways: (1) the first consists in changing the magnetic flux which encircles or links a conductor (fig. 2-5), and (2) the conductor or the field may move relative to each other, so that the conductor cuts across the flux lines composing the field (fig. 2-6).

These two methods of producing an electromotive force, V, by means of a magnetic field may be expressed in the following two equations:

$$V = -\frac{N}{10^8} \times \frac{d\phi}{dt} \qquad (2)$$

N = number of turns in coil

ϕ = flux lines; $\frac{d\phi}{dt}$ = rate of change of flux lines

$$V = \frac{B \ell v}{10^8} \tag{3}$$

B = flux density in lines/cm^2

ℓ = length of conductor in cm

v = velocity across field in cm/sec

Equation 2 states that the electromotive force is equal to the number of turns of the coil divided by a constant, times the rate of change of flux with respect to time. The negative sign indicates that the electromotive force generated causes a current to flow, which in turn produces a magnetic field opposing the rate of change. For example, if the voltage is produced by a steadily decreasing magnetic field, a field produced by the resulting current will be in the same direction as the generating field thus opposing the change. Let's illustrate the use of these relations with a simple example.

$N = 100$ turns

$$\frac{d\phi}{dt} = 10^6 \frac{\text{flux lines}}{\text{second}}$$

$$V = -\frac{100}{10^8} \times 10^6 = -1 \text{ volt}$$

B = 1,000 Gausses

$v = 10^4$ cm/sec

$\ell = 10$ cm

$$V = \frac{10^3 \times 10^4 \times 10}{10^8} = 1 \text{ volt}$$

2.1.6. PERMEABILITY. When certain substances are introduced into a magnetic field, the number of flux lines in the material is found to exceed greatly the number of flux lines that would otherwise be present at any given point. When an iron bar is placed in the interior of a current carrying solenoid, the external field of the solenoid is much increased. Materials that behave in this manner are said to have a high permeability. Permeability is by definition, the ratio of the flux density existing in the material in the presence of a given magnetizing force, to the flux density which would be produced in air or vacuum by the same magnetizing force. This is shown by formula (4) where the greek letter μ is the symbol for permeability, and is equal to the flux density divided by the magnetizing force. The permeability of the vacuum is by definition equal to unity. The permeability of air differs from this amount by so little that it is usually considered unity also.

$$\mu = \frac{B}{H} \tag{4}$$

μ = permeability

Substances with permeability greater than unity are called paramagnetic and substances with permeabilities less than unity are said to be diamagnetic. Iron and a few other materials, such as cobalt and nickel, possess high permeability, and these materials are called ferromagnetic materials. Permeabilities of 1000 and greater are observed. Furthermore, the permeability of materials such as iron is not constant, and is a function of the magnetizing force as we shall shortly see.

2.1.7. MAGNETIZATION. If the induction of a piece of iron is plotted against the magnetizing force for constantly increasing values of the latter, and assuming a neutral nonmagnetized sample, the curve known as the normal magnetization curve results. This is shown in figure 2-7. The value of B increases slowly at first to a point known as the instep, increases rapidly to a second point known as the knee, and then further increases slowly in a region known as the saturation region. Beyond the saturation point the presence of the material itself makes no further contribution to the total induction. It may be seen from this curve that the B/H relationship is nonlinear, and that no one number suffices to describe the quotient of these two quantities, and hence to designate the permeability.

2.1.8. HYSTERESIS. If the value of H is decreased from the saturation point the induction does not retrace the path which it followed as the magnetizing force was increased. It follows instead a new path such as ab in figure 2-8 which intersects the induction axis considerably above the origin. This value of induction remaining in the iron when the magnetizing field strength has been reduced to zero is called the retentivity of the sample. The induction remaining in ferromagnetic materials after they have been subjected to a magnetizing field is called the remanence. The retentivity is simply the maximum value of remanence which a sample can have.

To reduce the induction in a sample to zero after it has been subjected to a saturating field strength, it is necessary to apply a magnetizing field in the opposite direction. As this counter-magnetizing force is applied the B/H coordinate will follow the path cd (fig. 2-8) into the negative magnetizing force region. If at any point such as 2, 3, or 4, along this path the magnetizing force is removed, the BH coordinate will move toward the B axis along the short curves, as shown, to the points B_2, B_3, or B_4. These points represent the remanence values for the particular cycles shown.

It should be noted particularly that these remanence values depend entirely on the magnetic history of the sample. If, for example, the field strength H_2, H_3, or H_4 had been applied initially to the magnetically neutral material, the remanences would have been given by B'_2, B'_3, or B'_4 instead of B_2, B_3, or B_4. The point H_4 at which the induction in the sample becomes zero as the magnetizing force is decreased from saturation is called the coercive force. This point is also a measure of a value known as the coercivity of the material. This is one of the most important characteristics of a piece of magnetic material especially in the field of magnetic recording. It represents the force necessary to demagnetize the material.

If the magnetic field strength is again decreased from the coercivity point to H_5, the induction now becomes negative, and a point known as the negative saturation region is reached. If the magnetizing force is again reduced, a new value of remanence is reached as the force becomes zero. As the magnetizing force is again increased to H_1 a path is traced which again reaches the positive saturation region. The curve thus traced out is called the hysteresis curve for the material. If at any point in the hysteresis cycle the magnetizing force is made to undergo a small cyclic reversal a so-called minor hysteresis loop will be traced out. As shown in figure 2-9, this loop will be appended to the major loop which will be otherwise unaffected.

2.1.9. MAGNETOMOTIVE FORCE AND RELUCTANCE. Magnetic materials also possess a property analogous to the conductivity of the conductor in electrical circuits. An electrical conductor contains a property known as resistance, whereas this property in iron or other magnetic materials is known as reluctance. For a material of uniform cross section, the reluctance is given by equation 5. Analogous to the voltage in an

$$R = \frac{\text{length}}{\mu \times (\text{cross sectional area})} \tag{5}$$

electric circuit we have the magnetomotive force (mmf). This relation is given in equation 6. The Gilbert is the official unit (e m u)

$$\phi = \frac{\text{mmf}}{R} \tag{6}$$

of magnetomotive force, but another unit commonly employed is the ampere turn. In turns of ampere turns we have the relation given by equation 7. One Gilbert is thus seen to be equal to $4\pi/10$ ampere turns.

$$\phi = \frac{4\pi NI}{10R} \tag{7}$$

Let's take now a simple example illustrating these relationships.

Example 2

ring of iron, mid-dia. = D = 100 cm (average diameter)

cross sectional area = A = 1 cm^2

μ = 1,000, N = 100 turns, I = 10 amps

air gap = G = 0.1 cm

$$R_I = \frac{\pi \times 100}{1000 \times 1} = 0.314 \text{ reluctance of ring}$$

$$R_G = \frac{0.1}{1 \times 1} = 0.1 \text{ reluctance of gap}$$

$$R = 0.314 + 0.1 = 0.414$$

$$\text{Flux} = \frac{4\pi \times 100 \times 10}{10 \times 0.414} = 3.040 \text{ Maxwells}$$

Since A = 1 cm^2

$$B = 3{,}040 \frac{\text{Maxwells}}{cm^2} = 3{,}040 \text{ Gausses}$$

The distinction between magneto-motive force and magnetizing force should be kept clearly in mind. Magnetizing force is the magneto-motive force per unit length. When the unit length is the centimeter, the magnetizing force becomes the Gilbert per centimeter or the oersted.

2.1.10. DEMAGNETIZATION. When a hysteresis curve is drawn for a ring it is assumed that the curve is drawn for a closed sample of the material. When a curve is drawn for material which contains an air gap part of the applied magneto-motive force is expended as a drop across the gap. The resulting effect is that the hysteresis curve is flattened, and any remanence is reduced. The new curve is called a sheared hysteresis curve and is shown in figure 2-10. A term demagnetization effect has been applied to this phenomenon. It is important in magnetic recording particularly when high frequencies are considered.

2.1.11. HARD AND SOFT MAGNETIC MATERIALS. Two types of ferromagnetic materials are important in magnetic recording: the hard magnetic materials which are employed for the recording mediums, and the soft magnetic materials which are used for the pole pieces of the recording and reproducing heads. A hard magnetic material is one with high coercivity which will retain its magnetism when relatively high counter-magnetizing forces are present. Such a material also resists the demagnetizing effect. Soft magnetic materials are those which combine a high value of permeability with a low value of coercivity. For a summary of names and conversion factors for magnetic units see table 2-I.

TABLE 2-I. NAMES AND CONVERSION FACTORS FOR MAGNETIC UNITS

Quantity	MKS Name	Unrationalized CGS or emu Name	MKS Unit / emu Unit
Magnetizing Force (Magnetic Field) H	ampere-turns per meter	$\frac{\text{gilbert}}{\text{cm}}$ or oersted	$4\pi \times 10^{-3}$
Magnetomotive force mmf	ampere-turns	gilbert	$\frac{4\pi}{10}$
Magnetic flux ϕ	weber	maxwell	10^8
Magnetic flux density, (Force per unit pole) (Magnetic induction) B	$\frac{\text{weber}}{\text{m}^2}$ or tesla	$\frac{\text{maxwell}}{\text{cm}^2}$ or gauss	10^4
Pole strength m	ampere meters or unit poles	abampere centimeter or unit poles	10

Permeability of vacuum (μ_o) = $4\pi \times 10^{-7}$ MKS System = 1 CGS system

Conversion factor $4\pi \times 10^{-7}$

Substance	Type	Relative Permeability, μ_r
Air	Paramagnetic	1.0000004
Water	Diamagnetic	0.999991
Aluminum	Paramagnetic	1.00002
Nickel	Ferromagnetic	600
Supermalloy	Ferromagnetic	1,000,000

2.2. THEORY OF MAGNETIC RECORDING

2.2.1. INTRODUCTION. Three basic elements are required to make a magnetic recording and later reproduce it. These are:

1. A device which can respond to an electrical signal and create a magnetic pattern in a magnetizable medium,
2. A magnetizable medium which will conform to and retain the magnetic pattern,
3. A device which can detect such a magnetic pattern and convert it once again to the original electrical signal.

These three elements are shown in figure 2-11. They are the record head, the magnetic tape, and the reproduce head. In addition to these basic elements electronic amplification, and mechanical tape handling are required.

2.2.2. DIRECT RECORDING. The most straightforward manner of recording on magnetic tape is called the direct record process. The record head is fed with a signal current which flows through the winding and produces magnetic flux in the head core material. To perform as a record head, the core is made in the form of a closed ring which has a short non-magnetic gap in it. When the non-magnetic gap is bridged by the magnetic tape, the flux detours around the gap to the tape, completing the magnetic path through the core material. The magnetic tape itself is a ribbon of plastic upon which tiny particles of magnetic material have been uniformly deposited. As the tape is moved past the gap in the record head the magnetic material or oxide of the tape is subject to a flux pattern which is a function of the signal current in the head winding. As the tape leaves the head gap each particle retains the state of magnetization that was present at the training edge of the head gap.

To reproduce the signal the magnetic pattern on the tape is moved across the reproduce head. Again a small nonmagnetic gap in the head core is bridged by the magnetic oxide of the tape. Magnetic lines of flux are shunted through the core, and these lines are proportional to the magnetic gradient of the pattern on the tape which is spanned by the gap. The voltage induced in the head winding follows the law of electromagnetic induction which we considered in Section I. This law states that induced voltage is proportional to the number of turns on the reproduce head, and the rate of change of flux flowing through the core. Notice that the reproduced voltage is not proportional to the magnitude of the flux but to the rate of change of flux.

Suppose a sinusoidal signal is to be recorded on tape. This may be described by the following relation shown in equation 8.

$$V = A \sin \omega t = A \sin (2\pi f)t \tag{8}$$

The record amplifier is designed with a constant current output so that the current is the record head winding is proportional to the applied voltage at all frequencies within the desired frequency band. The flux through the record head core is thus proportional to the applied signal voltage. The tape then retains this flux pattern, and regenerates it in the reproduce head core. The voltage in the reproduce head winding is therefore given by equation 9

$$V_{repro} \propto \frac{d\phi}{dt} \tag{9}$$

where $\phi = kA \sin (2\pi f)t$

and $\frac{d\phi}{dt} = kA\, 2\pi f \cos (2\pi f)t$

Then:

$$V_{repro} = kAf \cos (2\pi f)t \tag{10}$$

Thus the reproduce head acts as a differentiator, and the reproduced signal is actually the derivative of the recorded signal—rather than the signal itself. This fact imposes two well known limitations on the direct record process: (1) The output of the reproduce head is proportional to the signal frequency and to maintain the proper amplitude relationships a 6 db per octave rise must be compensated for in the reproduce amplifier. This is shown in figure 2-12. The process of shaping the reproduce amplifier output is known as equalization. (2) The other limitation occurs as the recorded frequency approaches zero. At some point the output voltage from the reproduce head falls below the inherent noise level of the overall recording system. So there is a low frequency limit in the direct record process below which reproduction cannot be made.

2.2.3. D-C BIASING. A recording made by the process just described would be found to have poor signal-to-noise ratio and high distortion. The reason for this may be seen from the consideration of the magnetization curve discussed earlier. If we plot a head-to-tape transfer characteristic as drawn from a magnetization curve a result similar to figure 2-13 will be obtained. The high distortion caused by this type of transfer characteristic is apparent. Fortunately, there are two fairly linear segments in the transfer curve, one on each side of the origin, with their centers about halfway to saturation. If the recording can be confined to one or both of these straight sections low distortion can be realized. Early investigators, in the field of magnetic recording, conceived the idea of using d-c bias in order to permit operation on the straight portion of the characteristic. The method first employed was to pass enough d-c

current through the record head winding so that a certain constant level of magnetizing force was produced. This caused the operating point to be established on the straight portion of the characteristic in the first quadrant. This type of biasing was used in some early magnetic recorders, and did produce improved operation, but was hampered by a restricted dynamic range. An improved method of d-c bias moved the magnetic medium into the saturation region before recording began and then reduced the field so as to place the operating point on the back slope of the hysteresis curve in the third quadrant (see fig. 2-14). This method of bias yielded two significant improvements: (1) erasures of previous recorded material by taking the magnetic medium to saturation, and (2) much improved dynamic range by using the long straight portion at the back of the hysteresis curve.

2.2.4. A-C BIASING. A still better method of bias is provided by the high frequency a-c bias used in modern recorders. The clearest visualization of a-c biasing can be gained by describing the process known as anhysteretic magnetization.

Let's consider a flexible bar magnet made by cutting off a length of blank tape, say 4 cm long. This bar can be magnetized in the interior of a solenoid carrying a known amount of direct current. The resulting permanent magnetization, left after the current is removed, can be measured by means of a flux meter. Performing this experiment and plotting the resulting magnetization gives us the familiar magnetization curve. Now let's run the experiment again, except this time we will add an alternating magnetization current which we then reduce to a zero value before turning off the direct current. When this procedure is employed the series of curves shown in figure 2-15 is obtained, in which, as the a-c field is increased, the magnetization curve becomes increasingly more linear. Notice that in addition we have increased the sensitivity; that is, the magnetization, for a given d-c magnetizing current. Thus, with this system an undistorted recording can be made having high signal-to-noise ratio. In this experiment the d-c represents the signal to be recorded, and the a-c represents the a-c bias. The principal difference between this experiment and actual tape recording is that the a-c field decreases while the d-c field remains constant. If we were to use a magnetic ring core head on a tape recorder and magnetize the tape as it is pulled past the head we find that the a-c and d-c fields die out together. Using this approach in our experiment figure 2-16 shows that the curves are as before except that beyond a certain critical a-c field strength the magnetization decreases. This magnetization process is exactly equivalent to what actually happens in a tape recorder at low frequencies. At high frequencies the process becomes more complicated because the signal field is changing while a particle of tape passes across the recording gap. Increasing bias current increases the output up to the point of maximum sensitivity, also called

peak bias and then further increases in bias current decrease the output. The choice of the best bias current for practical operation of a tape recorder depends on several factors. The bias current must be selected as a compromise between sensitivity, frequency response, and distortion.

Several features of a-c bias operation are worth noting:

(1) Bias and the signal are linearly mixed or added together—it is not a modulation process.

(2) The proper amplitude for the bias is dependent upon the exact transfer characteristic of the tape, and should be adjusted for each particular type of tape. Two much bias will greatly reduce the high frequency response, while inadequate bias will cause increased distortion of the low frequencies.

(3) Bias frequencies are usually not critical but should be at least 3.5 times the highest frequency to be recorded, to minimize interaction with the harmonics of the signal.

(4) Care must also be taken to provide a harmonically pure sine wave bias current to minimize distortion products. Bias currents from 1 to 20 milliamperes are common, and they may be from 5 to 30 times the signal current depending on the tape and head characteristics.

Several factors combine to limit the high frequency response of tape recorders, and before these are discussed we should understand what is meant by recorded wavelength.

If a sine wave signal is recorded, the magnetic intensity of the recorded track will vary sinusoidally. The distance along the tape required to record a complete cycle is called the recorded wavelength and the wavelength parameters are shown in equation 11.

$$\lambda = \frac{v}{f} \tag{11}$$

where λ = recorded wavelength

v = tape speed or velocity

f = signal frequency cps

Several ways of using this relationship are shown in the following example:

1.2 MHz response @ 120 ips

$$\lambda = \frac{120}{1.2 \times 10^6} = 10^2 \times 10^{-6} = 10^{-4} \text{ or } 1/10 \text{ mil}^*$$

*A mil is one thousanth of an inch.

$$D = \frac{1.2 \times 10^6}{120} = 10^4 \text{ or } 10{,}000 \text{ cycles/in.}$$

where D = packing density.

The wavelength of 1/10th mil is also said to be the limit of the machines resolution. Both packing density and resolution can be used to describe a recorder's response independent of tape speed, and are thus more descriptive of a recorder's capability than a frequency specification at a given tape speed.

Five factors contribute to the frequency limitation of tape recorders. These are (1) gap effect, (2) recording demagnetization, (3) self-demagnetization, (4) penetration losses, and (5) head losses. As we have seen before, the reproduce head output increases with frequency up to a point, and then decreases rapidly to zero. This decrease is primarily the result of gap effect, and occurs as the recorded wavelength becomes shorter and shorter until it eventually equals the reproduce gap dimension. At this point there is no magnetic gradient spanned by the gap, and thus no output voltage. This is the most important restriction on high frequency response. Figure 2-17 is a graphical representation of the gap effect.

Recording demagnetization causes a decrease in the reproduced signal level at the shorter wavelengths. This loss occurs as the name applies in the recording process. When short wavelengths are being recorded the recording signal may vary considerably as the particle is leaving the recording field, and a corresponding reduction in a remanent magnetization will result.

Self-demagnetization occurs in the magnetic medium when the external magnetizing force is removed. This is most pronounced when the magnetic poles are crowded close together as with short wavelengths.

Penetration losses occur because the full depth or thickness of the magnetic coating on the tape is only magnetized at long wavelengths. As the wavelengths become shorter, only the surface layer of the particles are effectively magnetized. Thus, since short wavelengths influence fewer particles, there is less intensity in the recorded pattern, and the reproduced output falls off.

The losses in the head increase with frequency like any a-c driven ferromagnetic material. Core losses and winding losses both act to reduce the effective recording current at high signal frequencies. These losses are the result of hysteresis and eddy currents in the core material, and the distributed capacity of the windings.

Figure 2-18 shows the flux distribution in the recording medium, and the flux density on the surface B_S of the recording medium. In the bottom of the figure the concept of halfwave magnets is introduced. A halfwave magnet is a bar magnet of the same material and cross section as the recording track of the medium, and of the length equal to 1/2 wavelength of the signal to be recorded. Notice also that the plot of flux and flux density on the surface, illustrates the 90^o phase shift inherent in the recording-reproducing cycle. The flux on the tape is directly proportional to the recording current in the record head, and the surface flux density of the recording medium produces the voltage induced in the reproduce head. In other words, for sinusoidal input the recorded flux is proportional to sin ωt of the recorded signal, whereas the reproduced signal is proportional to cosine ωt.

In discussing the magnetic characteristics of ferromagnetic materials, it was seen that the transfer characteristics becomes nonlinear gradually as magnetization approaches saturation. This gives the recording process a graceful limiting, or, in other words, increasing the recording level above normal will gradually increase distortion before hard limiting occurs. To define the maximum signal level which can be recorded, it is necessary to state the maximum distortion which can be tolerated. In practice the specified maximum signal level is usually specified as the 1% total harmonic distortion point. This is abbreviated THD.

Dynamic range or signal-to-noise ratio is quoted in decibels, and is the ratio of the maximum signal for a given THD to the minimum signal which can be recorded. The minimum signal is determined by the noise level of the entire system over the bandwidth of interest.

Let's digress at this point for a moment, to review the definition and the meaning of the decibel. It will be recalled that the following definition describes a decibel:

$$db = 10 \log_{10} \frac{P_1}{P_2} \tag{12}$$

P_1 = Power "one"
P_2 = Power "two"

Notice that the basic definition of a decibel is in terms of a ratio of two powers. Remember, further, that a decibel is nothing more than a ratio, that is, to speak of the amplitude of a signal as 20 db has no meaning, anymore than saying the signal has an amplitude of 10 times. Signal amplitudes expressed in decibels must be referred to some reference level. This is normally indicated by attaching an additional letter, and saying for example, 20 dbV. This means that the signal amplitude is 20 decibels greater than one volt rms.

There is frequently considerable confusion on the subject of the decibel, because of the supposed existence of two types of decibels. We sometimes hear references to the power decibel and the voltage decibel. There is, in fact, only one decibel as defined by the equation we have written down. One may however, use this same expression to express the amplitude relationship of two voltages at the same impedance. Let's see how this would be done.

$$P = VI = V\left(\frac{V}{R}\right) = \frac{V^2}{R}$$

$$db = 10 \log \frac{E_1^2}{E_2^2} \text{ or,}$$

$$db = 20 \log \frac{V_1}{V_2}. \tag{13}$$

Thus we see, if the measurements are made at the same impedance, the ratio of the two voltages in question can be expressed in db by taking 20 times the logarithm of the ratio of the voltages. This gives rise to the expression "the voltage db," whereas in reality it is the same power expression previously given, but employed by measuring the ratio of the voltages at the same impedance.

This is the method generally used to measure the signal-to-noise ratio from the tape recorder. The system noise is measured across the output impedance as so many volts. The output of the tape recorder system, when reproducing the maximum record level is then measured as a second voltage. Twenty times the logarithm of this voltage ratio is then the signal-to-noise ratio of the tape recorder channel in decibels. This can be illustrated by the following example.

assume: reproduce level = 1 v rms
noise level = 0.01 v rms
output impedance = 10 K ohms

$$S/N = 20 \log \frac{1}{0.01} = 20 \log 100$$

$$S/N = 20 \times 2 = 40 \text{ db}$$

This example states that the signal-to-noise ratio is 40 decibels, or, equivalently, that the ratio of the signal to noise is 100. This may also be expressed by saying that the output voltage is 100 times the noise voltage. This same result would be achieved, if one chose instead to use

the power ratios, by computing the power available when reproducing the signal, and the power available when reproducing the noise, and using the basic defining formula.

Tape recorders exhibit another feature which must be recognized and understood when using magnetic recording for instrumentation purposes. This peculiarity is amplitude instability of varying degrees. The effect is seen in the reproduced signal as short random reductions or dropouts in signal level. These dropouts are caused by imperfect tape to head contact. If the amplitude of the signal being recorded must be carefully preserved, some means of reducing the effect of this tape recorder feature must be found. We will discuss how this is accomplished as we explore the various methods of recording signals on tape. It explains, however, why amplitude modulation is not used with a magnetic tape recorder.

2.2.5. PROBLEMS

1. Find the flux density at a distance of 10 cm in air from a north pole with a strength of 1000 unit poles. Also find the force on another north pole of equal strength at this distance. Assume the south poles are at a large distance. In which direction does the field act, and in which direction is the force on the equal pole?

2. How many flux lines issue from a pole of pole strength m? (The surface area of a sphere of radius r is $4\pi r^2$.) Assume this is an isolated pole in free space.

3. A certain tape recorder has a specification of 6 MHz response at 1500 ips. What is the recorded wavelength? What packing density does this represent? What is the limit of the machines resolution?

4. Using the problem worked in class, verify that the signal-to-noise ratio is the same when the computation is based on the signal power and the noise power rather than on the voltage ratio. Parameters given were:

   Reproduce output level = 1 v rms
   Noise level = 0.01 v rms
   Output impedance = 10 K ohms

   The signal-to-noise ratio obtained previously was 40 db.

2.2.6. SOLUTIONS

1. One definition of flux density is force per unit pole. Hence, since

$$F = \frac{m_1 m_2}{r^2}$$

the force per unit pole is

$$\frac{F}{m_1} = \frac{m_2}{r^2}$$

so

$$B = \frac{m_2}{r^2} = \frac{1000}{10^2} = \frac{10^3}{10^2} = 10 \text{ Gauss}$$

The force on an equal pole is:

$$F = B \times m_1 = 10 \times 1000 = 10{,}000 \text{ dynes}$$

or

$$F = \frac{m_1 \times m_2}{r^2} = \frac{1000 \times 1000}{100} = \frac{10^6}{10^2} = 10^4 \text{ dynes}$$

The field acts radially outward from the pole. The force on the equal pole repels the pole.

2. The induction at r cm from a pole of strength m is

$$B = \frac{m}{r^2} \text{ Gauss or } \frac{\text{Maxwells}}{cm^2}$$

since the area of a sphere of radius r is $4\pi r^2$, the total flux is

$$\phi = \frac{m}{r^2} \times 4\pi r^2 = 4\pi m \text{ Maxwells (lines)}$$

3. The recorded wavelength is

$$\lambda = \frac{v}{f}$$

$$\lambda = \frac{1500}{6 \times 10^6} = 250 \times 10^{-6} = 0.25 \text{ mil}$$

The packing density is

$$D = \frac{f}{v} = \frac{6 \times 10^6}{1500} = \frac{6 \times 10^6}{1.5 \times 10^3} = 4000 \text{ cycles per inch}$$

The limit of the machines resolution is 0.25 mil.

4. The reproduce signal power is:

$$P_R = \frac{V^2}{Z} = \frac{1^2}{10^4} = 10^{-4} \text{ watts}$$

the noise power is:

$$P_N = \frac{(10^{-2})^2}{10^4} = 10^{-4} \times 10^{-4} = 10^{-8} \text{ watts}$$

then

$$S/N = 10 \log \frac{10^{-4}}{10^{-8}} = 10 \log 10^4 = 10 \times 4 = 40 \text{ db.}$$

2.3. COMPONENTS OF A MAGNETIC RECORDING SYSTEM

2.3.1. TAPE RECORDER HEAD. The most important element in any recorder is its heads. There are record heads, reproduce heads, and erase heads. There are rare cases such as some FM carrier recorders where a single head is used for both record and reproduce functions, but the great majority of instrumentation recorders use separate record and reproduce heads. The reason for this is that separate optimization can be made in each head, for the function which it is to serve.

The construction techniques of tape recorder heads are most carefully guarded proprietary secrets in the recording industry. Many aspects of the assembly of high resolution record and reproduce heads are close to being a black art rather than a science.

Figure 2-19 is an enlarged view of the essential features of the tape recorder head. All heads, whatever their purpose, are designed with this basic construction. The core consists of two c-shaped half sections made from a number of bonded laminations of thin, high permeability, ferromagnetic material. The surfaces of the half sections which interface are lapped and polished very carefully, and the gap material is deposited on one surface. The two cores and their windings are then joined together to complete the head. One head is used for each track of the tape in multitrack operation. This requires that several heads must be assembled together with intertrack shields to form a head stack. Figure 2-20 is another illustration of the component pieces making up a single track head. The core laminations, the gap separators, and the shield are shown, as well as the windings and the two c-shaped pieces which make up the core. Figure 2-20 also shows a cutaway of a typical multitrack head. This configuration

contains seven individual separate heads, one for each of seven tracks, assembled together with intertrack shields, and molded in plastic into a single assembly. Extreme precision is necessary in aligning the heads in a head stack since the gaps of all heads must fall within a 100 microinch band. Figure 2-21 is a sketch of several head and gap specifications. The head stack is finished by contouring and lapping for optimum head to tape contact, and then it is mounted on a precision machined mounting base by which it is secured to the tape transport. The complete head assembly thus includes a base with a mounting surface whose plane is perpendicular to the gap line at the contact surface of the head stack. It is because of the precision mounting base, that the head stacks may be removed for cleaning, and new head stacks may be installed without factory adjustment.

Such precision mounting of the heads is necessary to reduce the short wavelength losses due to azimuth misalignment. These losses occur when the reproduce head gap is not precisely parallel to the record head gap. In order to insure interchangeability of tapes between various recording machines, the alignment of individual gaps within a head stack shall be within ±1 minute of arc referred to a straight line perpendicular to the direction of tape travel. The consequences of azimuth misalignment are shown by equation 14.

$$\text{Loss (db)} = 20 \log \left[\frac{\sin\left(\frac{\pi\omega \tan A}{\lambda}\right)}{\frac{\pi\omega \tan A}{\lambda}} \right] \qquad (14)$$

ω = track width

A = angle of misalignment

λ = wavelength of recorded signal.

This equation has the familiar (sin x)/x as a component, and graphs of this function with the ratio of track width to wavelength as a parameter are available. To show the significance of this azimuth misalignment effect, consider the following two examples:

(1) $\left.\begin{array}{l}\omega = 50 \text{ mils} \\ \lambda = 0.2 \text{ mil}\end{array}\right\} \omega/\lambda = 250$

$A = \left(\frac{1}{6}\right)^{o} = 10 \text{ min.}$

L = 9 db

(2) also

$\left.\begin{array}{l}\omega = 250 \text{ mils} \\ \lambda = 1 \text{ mil}\end{array}\right\} \omega/\lambda = 250$

L = 9 db

A standardized track and head geometry which guarantees compatibility between instrumentation tape recorders has been issued by IRIG. The standards have been issued as document 106-66 which was published in March 1966 and are shown in figure 2-22. The track width shall be 50 mils with a track spacing 70 mils center to center. Seven tracks will be used on 1/2" tape, and 14 tracks on 1" tape. The tracks as a group will be centered on the tape. This means, for example, that on 1/2" tape, track #4 will be located in the center of the tape. Standard head placement locates the head for alternate tracks in separate head stacks. Thus, to record all tracks of a standard width tape, two record head stacks are required, and, in addition to reproduce all tracks, two reproduce head stacks will be used. The two stacks of a record or reproduce pair, shall be mounted so that the center lines to the head gaps on each stack are parallel, and shall be spaced 1.5 inches apart along the tape path ±1 mil.

The tracks are numbered from left to right as you look down the tape in the direction of travel on the oxide side. Thus, track #1 is the first track on the left in the first head stack. The second head stack thus contains all even numbered stacks, with track #6 on the right hand edge. All machines of recent manufacture use this standard track numbering. In previous years, however, Ampex recorders used a different track numbering. This was essentially the reverse of the IRIG track numbering. The Ampex system numbers the tracks, 1 through 7 from right to left as you look down the tape in the direction of travel on the oxide side. This means that track 1 is the first track on the right in the first head stack.

These two systems are compatible except for identification numbers in the case of 1/2" tape, but are completely incompatible in the case of 1" tape. To be specific, see the chart shown as figure 2-23, which relates the track numbers in one system to the track numbers in the other system. In the case of 1" tape each recorded track will be reproduced, but recorded track 1 will reproduce as track 14, track 2 will reproduce as track 13, and so forth. This serves to double the original record head separation during playback. This comes about in the following way: the even numbered tracks are recorded by the #2 head stack at a point 1.5 inches downstream from the odd numbered tracks in either format. When reproduced with head stacks on the opposite format, the even numbered recorded tracks are reproduced by the #1 headstack, and 3" additional tape travel is necessary to get the time coincident data of the odd numbered tracks, beneath the number 2 reproduce head stack. At 60 ips tape speed a 50 millisecond delay is introduced in data which was time coincident when recorded.

It was explained earlier that the record process takes place at the trailing edge of the record head gap. Because of this, the record head gap length has little effect on frequency

response. Most machines use a record gap length of from 2/10ths to 5/10ths mil. It is, however, very important to the recorder's high frequency capability to get a sharp well-defined gap edge.

The high frequency response of the tape recorder, as we have seen, is determined by the gap length of the reproduce head. For a machine to reproduce a 1 MHz signal at 60 ips it must "see" recorded wavelengths on tape as short as 60 micro-inches. This requires a reproduce head gap length of approximately 20 micro-inches. The gap must also be sharply defined, and have adequate depth to withstand the abrasive effect of the tape. Unfortunately, as the gap length decreases, the output voltage from the head winding also decreases, thus lowering the signal-to-noise ratio of the recorder. The output voltage is reduced, because the shorter gap width has low reluctance, and tends to shunt some of the flux that would otherwise flow through the head core.

The recording and reproducing of short wavelengths requires intimate contact between the oxide particles and the head gap to eliminate spacing loss. It has been found experimentally, and verified theoretically, that the spacing loss is given by equation 15.

$$\text{Loss, (db)} = 55\,\frac{d}{\lambda}$$

d = spacing

λ = wavelength (15)

From this relation we can see that at a frequency of 1 mil wavelength, a half-mil spacing results in a drop of more than 20 db. The 1 mil wavelength results when recording for example, 60,000 Hz at 60 ips. To restrict the loss to 6 db for a 1 mil wavelength, the separation or spacing must be kept to less than 111 micro-inches. Since a dust particle might easily approach this size the possibility of spacing loss, and the importance of keeping the recorder extremely clean is obvious.

2.3.2. TAPE. The medium presently accepted for magnetic recording is a tape consisting of a plastic base with a coating of minute ferrous oxide particles. The tape backing consists of thin sheets of acetate or Mylar film. Thicknesses commonly employed are 1.5 mils, 1 mil, and 0.5 mil. The base material supports the magnetic layer. The magnetic layer or coating is classed under two headings:

(1) Magnetic oxide

(2) Binder material.

The most suitable magnetic material is called gammaferric oxide γFe_2O_3. A sample of tape coated with this material is shown in figure 2-24. The particle length is from 2/10ths to 8/10ths microns, and the width is 1/2 to 1/6th of its length. One micron is 10^{-6} meters, or approximately 39 micro-inches. The coercivity of this material is approximately 260 oersteds.

Mylar base is more expensive, but much more stable than acetate base. Mylar has approximately twice the tensile strength of acetate, and nearly 6 times the tear strength. The most important difference is the greater stability of Mylar in the presence of temperature and humidity changes. Mylar has one-half the coefficient of thermal expansion, and about 1/15th the coefficient of humidity expansion.

2.3.3. TAPE TRANSPORT. The sole purpose of the tape transport is to move the tape by the heads at a constant speed, and to provide the various winding modes of operation required for tape handling. A transport must guide the tape past the heads with extreme precision, and maintain the proper tension within the head area to obtain adequate tape to head contact. Spooling or reeling of the tape must be done smoothly, in order to introduce a minimum number of perturbations into the head area. The tape transport must also move the tape from one reel to the other, quickly, in the fast forward or rewind mode. Even with fast speeds the tape must be handled gently and accurately, to maintain good tape pack without straining or distorting the tape. In going from fast modes to stop, precise control of the tape must be maintained, so that undue slack or stress is not incurred by the tape.

Figure 2-25 shows the simplest type of drive, called the open-loop drive. This is basically a sound design both simple and reliable, but has higher than desirable flutter characteristics. The inherent sources of unwanted tape speed variations in the head area are: (1) disturbances caused by the supply reel, the drive motor, or the brakes, and (2) erratic vibrations of the tape because of the long unsupported tape length in the head area. Figure 2-26 shows the closed-loop transport design which would seem to solve both these problems. The unsupported tape length is halved, and if both ends of the tape in the head area are under positive control from the capstan, the reeling function disturbances are eliminated. Actually, these desirable characteristics are only partly achieved.

If there were no slippage between the tape and the capstan at either point of contact, then any initial tension in the loop would be maintained. But the only elements which can supply tension are outside the loop. Hence, even in the closed-loop design there must be some creep between the tape and the capstan in order to maintain tension within the loop. This means

motional disturbances in the reeling functions still cause uneven tension variations in the head area. These effects are, however, greatly improved over the open-loop design.

An improvement in the closed-loop design uses a two diameter capstan, shown in figure 2-27, to maintain tension within the closed-loop. The two discrete diameters, which are only a few thousands of an inch different, cause tension to be generated within the loop. Special contoured pinch rollers are used, and tape entering the head area is forced against the smaller capstan diameter, while tape leaving the head is forced against the larger capstan diameter. This generates a speed differential between the tape entering, and the tape leaving the loop. Less tape tensioning is necessary from the reeling function, and more isolation is obtained in the head area. The drive, however, is not suitable for bi-directional operation.

Another version of a tape transport uses a zero-loop tape drive shown in figure 2-28. Many modifications of these basic designs have been used. The open-loop design may be made to resemble the performance of a closed-loop design, by adding a second capstan on the other side of the head assembly (fig. 2-29). Another drive which has been used, eliminates the pinch rollers, and uses a capstan with high surface friction as a turn around roller. This machine provides tape tension in the head area, and isolation from reeling disturbances, by means of vacuum controlled tape chutes. Time does not permit a discussion of the details of the mechanisms used to control tape tension, braking, and capstan speed stability. An excellent discussion of this subject is contained in the NASA Publication on magnetic tape recording (Athey, Skipworth W; Magnetic Tape Recording, Technological Survey, NASA, SP-5038, Washington, D. C., 1966).

In spite of the excellent mechanical design presently available in a wealth of instrumentation recorders, tape speed irregularities and deviations, cause magnetic tape recording errors. Relatively long-term errors, are termed drift—usually representing deviations below 0.1 Hz. The term wow is used to describe components between 1/10th Hz and 10 Hz. Flutter denotes speed variations occurring at frequencies above 10 Hz. In instrumentation tape recording, the term flutter is sometimes used to denote all three of these effects. Sometimes flutter is called instantaneous tape speed error.

Short term speed variations which are uniform across the tape, can be caused in many ways in a tape transport mechanism. Some of these are: (1) pulsations of the torque motors, (2) reel eccentricities, (3) irregularity in tape pack or tape physical characteristics, (4) vibrations in the tape caused by friction as it passes over fixed guides or heads, (5) mechanical runout of rotating parts, (6) slight cogging of capstan drive motor, (7) power line voltage transients, (8) pinch rollers with surface deformations, (9) sticky bearings, (10) reels and reel drive assemblies which have varying velocities, and a mass which is constantly changing.

The flutter spectrum of the well-designed machine is made up of a combination of small discrete sinusoidal components, and a more or less uniformly distributed noise signal. Because of the noise involved, a well-defined measurement of flutter is difficult. The flutter signal can be obtained as shown in figure 2-30 by recording an extremely stable reference sine wave, and passing the reproduced signal through an FM discriminator. The signal thus generated would be zero if the tape speed were exactly the same for reproduction as it is for recording. Any variations between the two tape speeds will deviate the frequency of the reference tone, and produce an output from the discriminator. Flutter components are customarily measured to at least 10 KHz, and the reference tone should be 5 to 10 times the maximum flutter frequency. One common form of flutter specification is cumulative flutter. This is made by passing the flutter signals through a variable cutoff low pass filter, and measuring the filter's output for increasing values of the cutoff frequency. Since the noise contribution in the flutter signal is essentially uniform, the shape of the curve rises with frequency. At any point where some rotating component produces a discrete contribution, there will be a small step function in the curve. Many published curves of this type are usually averages, and may not show the steps mentioned above. A typical set of cumulative flutter curves is shown in figure 2-31.

An rms measurement for flutter has long been used for audio machines, but peak-to-peak measurement is more useful for instrumentation machines. Actually a true rms value of flutter is almost impossible to attain, since the flutter signal contains a d-c component, a noise component, and many sine wave components. As a very rough approximation, the rms value can be assumed to be 1/6th to 1/4th of the peak-to-peak value.

The usual way of making a peak-to-peak measurement of flutter, is to display the flutter signal on an oscilloscope with the vertical scale calibrated for percent peak-to-peak flutter. Photographs of these traces are probably the most accurate measurement available of the recorders amplitude characteristic. A typical oscilloscope representation is shown in figure 2-32.

2.4. METHODS OF STORING DATA ON MAGNETIC TAPE

There are several distinctly different recording processes in common use. Each of these processes requires a different form of electronic coding and decoding of the signal information. Some of the more common recording processes will be described and explained.

2.4.1. DIRECT RECORDING. Normal analog signals are recorded on tape by the direct record method. The information lies in the magnetization level from point to point along the

tape. High frequency bias current, approximately 3 to 5 times the highest signal frequency to be recorded, is added to the signal at the record head. As we have seen, this bias current has the effect of making the magnetization curve of the oxide, linear. Direct recording utilizes the maximum bandwidth capability of a recorder, but is limited in accuracy by all the extraneous phenomena which cause amplitude changes. The data to be recorded is amplified, mixed with bias, and presented directly to the recording head as a varying current. This is basically the process we have discussed, in considering the basic theory of magnetic recording. There are a number of machines on the market at the present time offering a bandwidth capability of 400 cps to 1.5 MHz ±3 db with a signal to noise of 25 db. This operation is provided at a tape speed of 120 ips.

2.4.2. AUDIO RECORDING. The audio recording process is a special type of direct recording. The characteristics of the record and reproduce amplifiers are modified to conform to the particular characteristics of speech and music signals. It has been established that the spectral energy content in speech and music signals is not uniformly distributed over the signal frequency range. For this reason, pre-equalization circuits are incorporated in the record amplifier which pre-emphasize the extreme low and high ends of the frequency spectra. These are the frequencies at which the energy content of audio signals is low. By raising their level, it is possible to approach a constant flux recording at all frequencies (fig. 2-33). In this way, benefits can be achieved in the signal-to-noise ratio without sacrifice in distortion. Of course, the inverse frequency response characteristic must be introduced in the reproduce amplifier in the form of post-equalization (fig. 2-34). This counteracts the effect of the pre-equalization, and produces a final output signal which is a replica of the original signal, but having much improved signal-to-noise ratio.

The danger in using an audio recorder for instrumentation purposes, is that the instrumentation type signal does not, in general, have the peculiar spectral energy distribution of speech or music. The result is that the pre-emphasis in the record amplifier could result in serious distortion of the high and low frequencies. This could be overcome by reducing the recording level by a significant amount with resultant deterioration in the signal-to-noise ratio, but the best course of action is to avoid the use of audio tape recorders for instrumentation purposes.

2.4.3. FREQUENCY MODULATION RECORDING. We have seen some significant drawbacks in the direct recording process; namely limited low frequency response, amplitude instability, and only moderate signal-to-noise ratio. In order to record d-c and low frequency

data magnetic recording had to somehow provide d-c response, good d-c linearity, and better signal-to-noise ratios. This need is filled by the FM recording technique. Data recording using a frequency modulation process is accomplished by deviating a carrier frequency in response to the amplitude of a data signal, and recording it. The signal is recorded to saturation without bias. A d-c signal of positive polarity increases the carrier frequency, and the opposite polarity decreases it. An a-c signal, alternately increases and decreases the carrier, above and below its center frequency, at a rate equal to the data signal frequency. In the reproduce process, the carrier's amplitude instability is wiped out by limiting, and the data signal is reconstructed by detecting zero crossings. Residual carrier signals, and out of band noise are removed by a lowpass filter. FM recording is, however, extremely sensitive to tape speed fluctuation; in other words flutter, since, in either the record or reproduce mode, tape speed variations produce unwanted modulation of the carrier which is demodulated as noise.

A brief discussion of FM theory will assist in an understanding of the FM recording process.

f_c = carrier frequency

Δf = carrier deviation from f_c

f_s = signal frequency

$f_{s_{max}}$ = highest data signal frequency

$\frac{\Delta f}{f_{s_{max}}}$ = deviation ratio or modulation index

$\frac{\Delta f}{f_c} \times 100$ = percentage deviation

Deviation ratio is one of the most important factors in any FM process. Basically, the higher the deviation ratio, the more immunity the system will have to noise. In FM recording, however, there are practical limits to deviation ratios, since Δf is restricted by recorder bandwidth limitations, and $f_{s_{max}}$ must be kept high to accommodate the data signals. Common deviation ratios in use today range from 5 in the telemetry FM sub-carriers through 0.675 in wideband FM recording to 0.125 in video recording. FM broadcasting which enjoys an excellent noise immunity uses the deviation ratio of 5.

The percentage deviation $\Delta f/f_c \times 100$ is another important factor in FM recording. Telemetry sub-carriers use deviations of 7 1/2% and 15%, while wideband recording systems use deviations ranging from 20% to 75%.

The importance of the deviation percentage can be illustrated by an example.

Assume

signal = 7 1/2% deviation

flutter and wow = 1%

then

noise = $1/7.5 \times 100 = 13.3\%$

but for a signal with 40% deviation

Noise = $1/40 \times 100 = 2.5\%$.

Thus the 40% deviation system has 13.3/2.5 = 5.3 times or 14 db better signal-to-noise ratio.

The most common FM recording system is now referred to as single carrier FM recording. This system uses 40% deviation with a deviation ratio of 2.16. The original system (now called lowband) uses a center carrier frequency in KHz which is 9/10ths the tape transport speed in ips (inches per second). For example, 13.5 KHz is the carrier for 15 ips. The bandwidth in KHz is 1/6th the tape speed in ips. For example, 2.5 KHz is the bandwidth for 15 ips. A table listing the FM recording parameters is given in the NASA publication on page 36. Four categories of FM recording standards are shown. The presently accepted names for the various categories are lowband, intermediate band, and wideband. The last category is now called wideband, group 2. This recording method is only used on machines with 1.2 MHz or greater direct record response at 120 ips. Direct recording with a-c bias is used to record this particular FM signal, since distortion products cannot be kept out of band. It is also necessary to employ considerable electronic sophistication to produce octave bandwidth, voltage controlled oscillators, and discriminators.

FM recording has the advantages of good amplitude accuracy, d-c response, good d-c linearity, low distortion, and high signal-to-noise ratio. The price paid for this improvement over direct-record performance is greatly reduced frequency response for a given tape speed, as well as added complexity and cost in the record and reproduce electronics.

Another very useful attribute of the FM recording method is the time base compression and expansion which can be achieved with this technique. As an example 20 KHz data recorded at 60 ips on a 108 KHz carrier can be reproduced at 1 7/8 ips with the data frequency spectrum reduced by a factor of 32. Combinations of changing tape speeds and rerecording, can provide time base changes and frequency shifts of well over 200.

2.4.4. VIDEO RECORDING. Video tape recording is described in some detail in the NAB reprint, but some of the more important points will be covered now. We have seen that a television signal contains energies in a bandpass from direct current to more than 4 MHz. In our study of tape recorders so far, we know that the head gap may be reduced or the tape velocity increased, or both, in order to attain a given high frequency performance. If an effective head gap of 0.25 mils is assumed the required head to tape velocity must be in excess of 2000 ips, or more than 110 mph. Even if the gap width could be reduced to 0.025 mils, the relative head to tape velocity required would be 200 ips. This would still require either reels of cartwheel size, or would indicate very short playing time. For example, a reel approximately 38 inches in diameter, containing 60,000 feet of tape, would be required for a playing time of one hour at such a tape velocity. It appears then, that a new concept is required. The technique used, employs a rotary drum containing 4 video heads mounted in quadrature on the periphery of a 2" diameter drum which rotates at 14,400 rpm (fig. 2-35). The heads trace a transverse path across the tape, which moves slowly in a longitudinal direction. The tape employed is 2" wide, and is oriented for transverse recording. The tape is cupped around the head drum by a concave tape guide. The relative head-to-tape velocity is approximately 1500 ips. The head used for this purpose is shown in figure 2-36.

During the recording process all heads are fed in parallel. The dimensions employed provide an overlap between successive heads, since each head describes an arc across the tape of approximately 120°. The appearance of a section of a recorded tape is shown in figure 2-37. Figure 2-38 is another picture showing the track configuration on the tape. The complete tape includes the video track and audio track, a cue track, and a control track. Each frame occupies 1/2 inch of tape and one field occupies 1/4 inch. Since the linear tape motion is at a velocity of 15 ips, and head rotational rate is 14,400 rpm or 240 rps, and, since 4 tracks are traced transverse to the direction of linear tape motion during each revolution of the head drum, 960 tracks are laid down during each second of recording along 15 inches of tape. More than 17 horizontal television lines are recorded on each transverse track. Because of the 120° arc across the tape traversed by each head, some of the information is duplicated between the end of one track and the beginning of the next. During this overlap period a switching system selects a horizontal blanking interval in which to accomplish the switching from one head to the next. Thus, either 16 or 17 horizontal television lines of video information are reproduced from each transverse track. The switching is done in this manner to prevent switching transients from appearing in the picture. One mil base tape is used, and 4800 feet can be wound on a reel 12 5/8" in diameter. This provides sufficient tape for a 64 minute recording or reproduction.

Although we have now achieved the necessary high frequency response in the neighborhood of 6 MHz, the lowest frequency which can be recovered with adequate signal level is now in the neighborhood of 3000 Hz. Furthermore, if the video signal were recorded directly, or recorded by amplitude modulation, the signal level will vary approximately 10% due to small tracking errors in the servos. To overcome these objections frequency modulation of a carrier is employed.

The FM system has, however, some unusual characteristics. The center carrier for this system is approximately 5.25 MHz with a deviation of approximately 500 KHz. The highest useful modulating frequency is slightly over 4 MHz. The modulation index is approximately 0.125. Thus, it is seen, that an FM signal which is deviated 500 KHz from a carrier in the neighborhood of 5 MHz by modulation frequencies in the neighborhood of 4 MHz through a system whose bandpass is sharply limited to approximately 6 MHz, only the lower sidebands will be fully recoverable. The video magnetic recorder operates, therefore, as a single sideband FM system.

The highest frequencies are distorted through an apparatus of this type, but the only detectable effect upon the picture is a small degree of aperiodic waviness in narrow picture elements which is integrated out by the eye. The synchronizing signals suffer serious distortion so they are stripped away and replaced with reconstituted synchronizing signals. For a more complete description of video recording see the section 6 part 5 in the NAB reprint.

2.4.5. OTHER TYPES OF RECORDING. There is currently available another rotary head tape recorder designed for instrumentation purposes rather than television recording. This machine has 10 Hz to 5.5 MHz bandwidth and a signal-to-noise ratio of 40 db peak-to-peak at full bandwidth. Special techniques have been used to eliminate the switching transients between segments of the rotary head.

Time does not permit a discussion of the various methods of recording pulses on magnetic tape. Several such methods are available, such as, (1) digital or pulse code modulation, (2) pulse duration modulation and (3) frequency shift keying.

3. SOURCES OF INFORMATION

Athey, Skipworth W., "Magnetic Tape Recording," National Aeronautics and Space Administration, 1966.

Begun, S. J., "Magnetic Recording," Rinehart and Company, 1955.

Davies, Gower L., "Magnetic Tape Instrumentation," McGraw-Hill, Inc., 1961.

Fink, Donald G., "Television Engineering," 2nd Edition, McGraw-Hill, Inc., 1952.

Howard, James A. and Ferguson, Lester N., "Magnetic Tape Recording Handbook," Application Note 89, Hewlett Packard Co., 1966.

Kraus, John D., "Electromagnetics," McGraw-Hill, Inc., 1953.

Lowman, Charles E. and Angerbauer, George J., "General Magnetic Theory," Ampex Corp., 1961.

McKnight, John G., "Biasing in Magnetic Recording," Ampex Readout, Nov. 1967.

NAB Engineering Handbook, 5th Edition, National Association of Broadcasters.

Stewart, W. Earl, "Magnetic Recording Techniques," McGraw-Hill, Inc., 1958.

"Telemetry Standards" (Revised March 1966), Ampex Corp., 1966.

Wentworth, John W., "Color Television Engineering," McGraw-Hill, Inc., 1955.

Zworykin, V. K. and Morton, G. A., "Television," 2nd Edition, John Wiley and Sons, Inc., 1954.

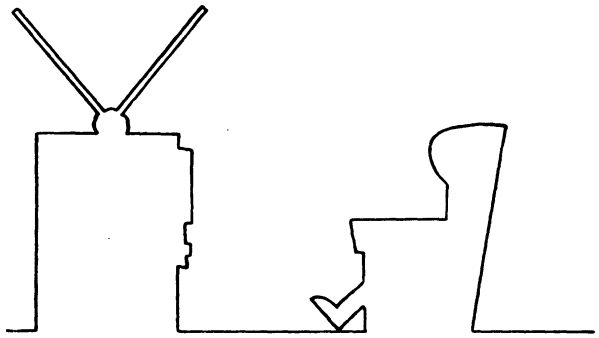
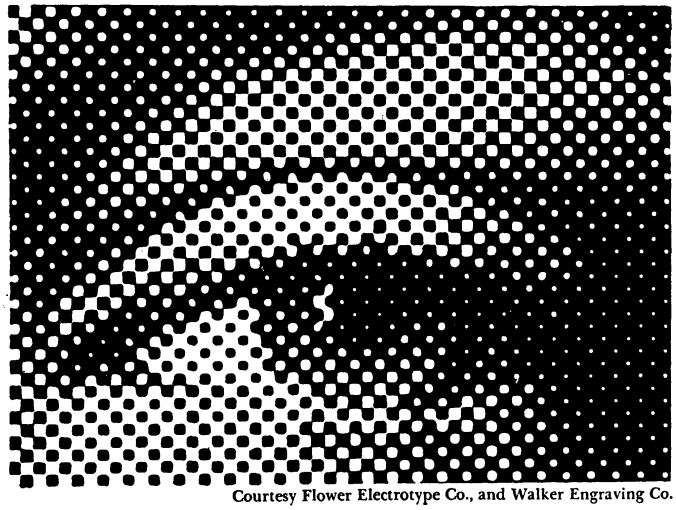


FIGURE 1-1. SKETCH TO ILLUSTRATE HOW PERCEPTION MAY BE CONTROLLED BY VAGUE PSYCHOLOGICAL CLUES. (From Color Television Engineering by John W. Wentworth. Copyright 1955 by McGraw-Hill. Used with permission of McGraw Hill Book Company.)



Courtesy Flower Electrotype Co., and Walker Engraving Co.

FIGURE 1-2. ENLARGED PHOTOENGRAVING ILLUSTRATING DOT STRUCTURE.

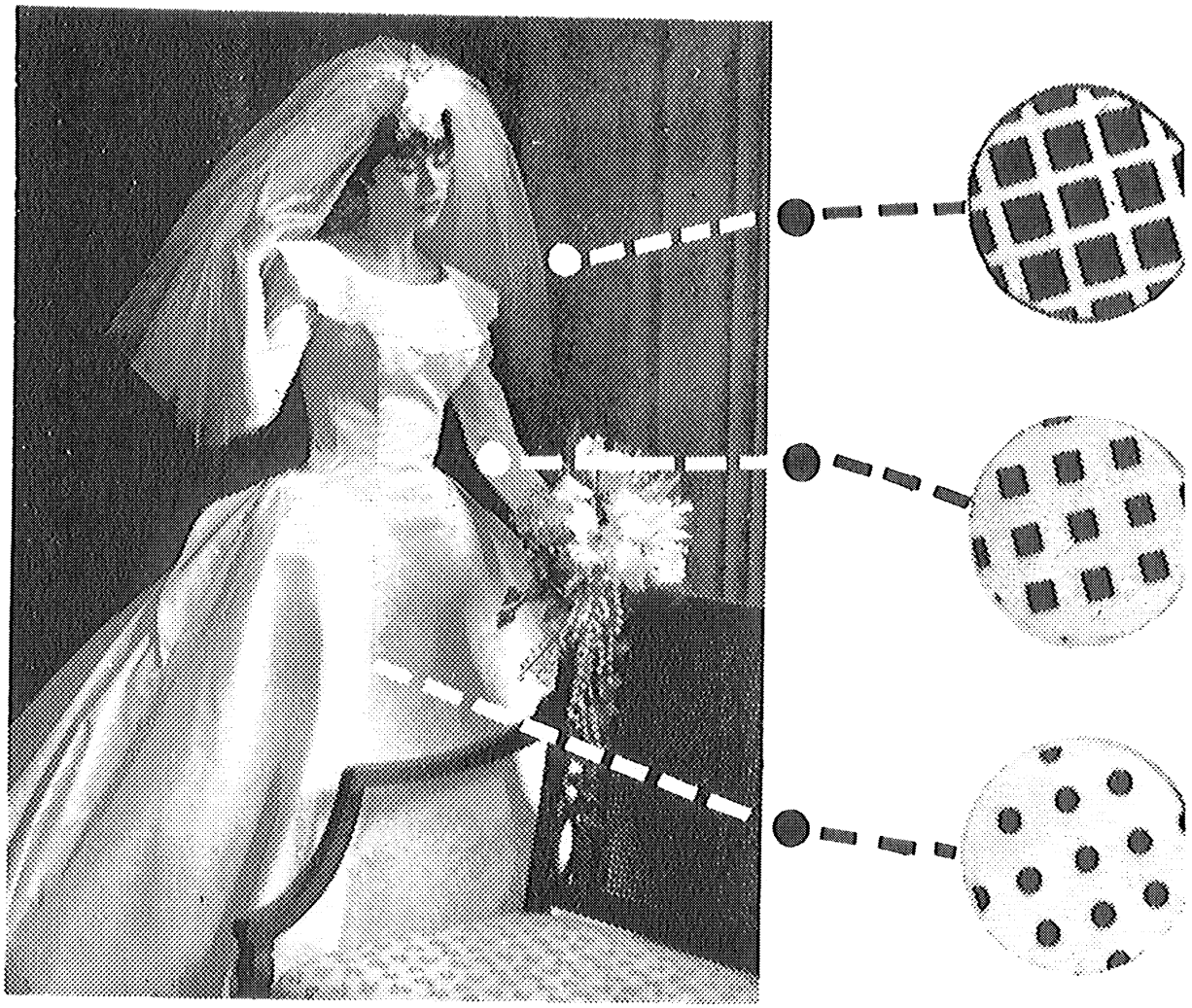


FIGURE 1-3. ILLUSTRATION SHOWING THE DOT STRUCTURE IN THE HIGHLIGHTS, MIDDLE TONES AND SHADOWS OF A HALFTONE NEGATIVE MADE BY THE PHOTOGRAVURE PROCESS

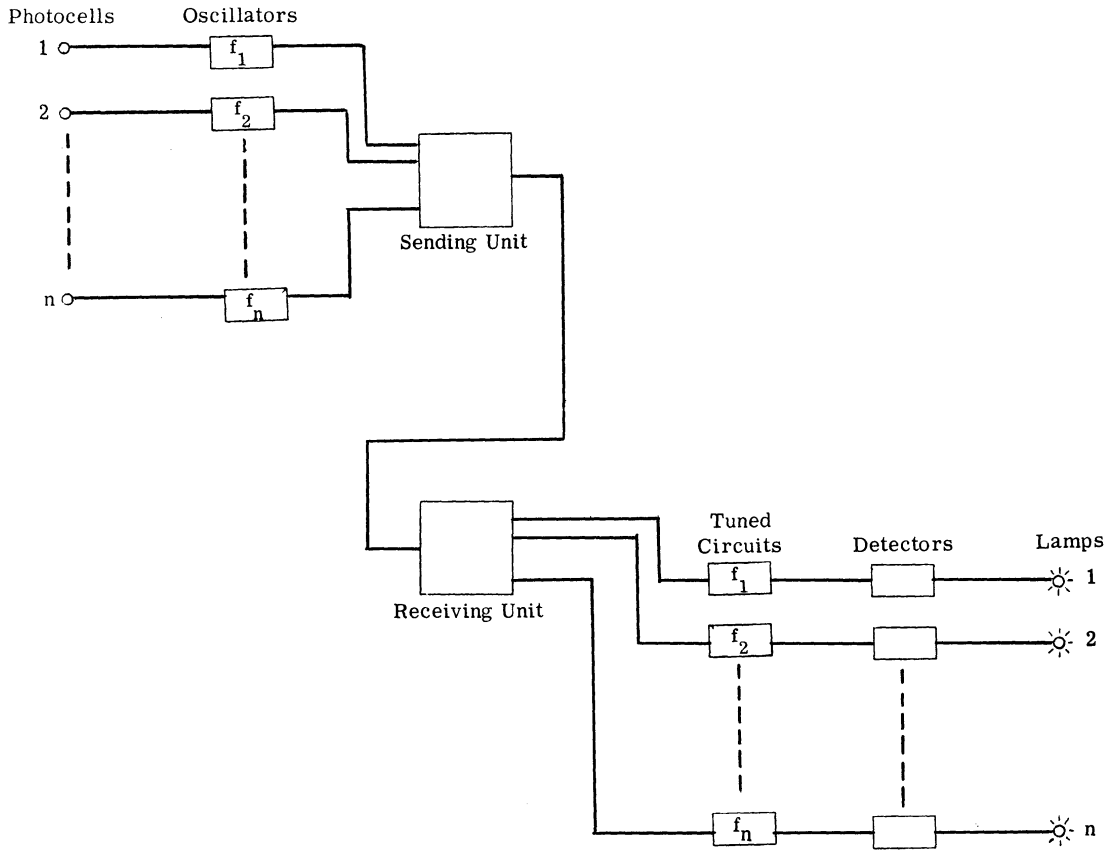


FIGURE 1-4. A POSSIBLE SYSTEM TO TRANSMIT PICTURE ELEMENTS ELECTRICALLY OVER A SINGLE CHANNEL

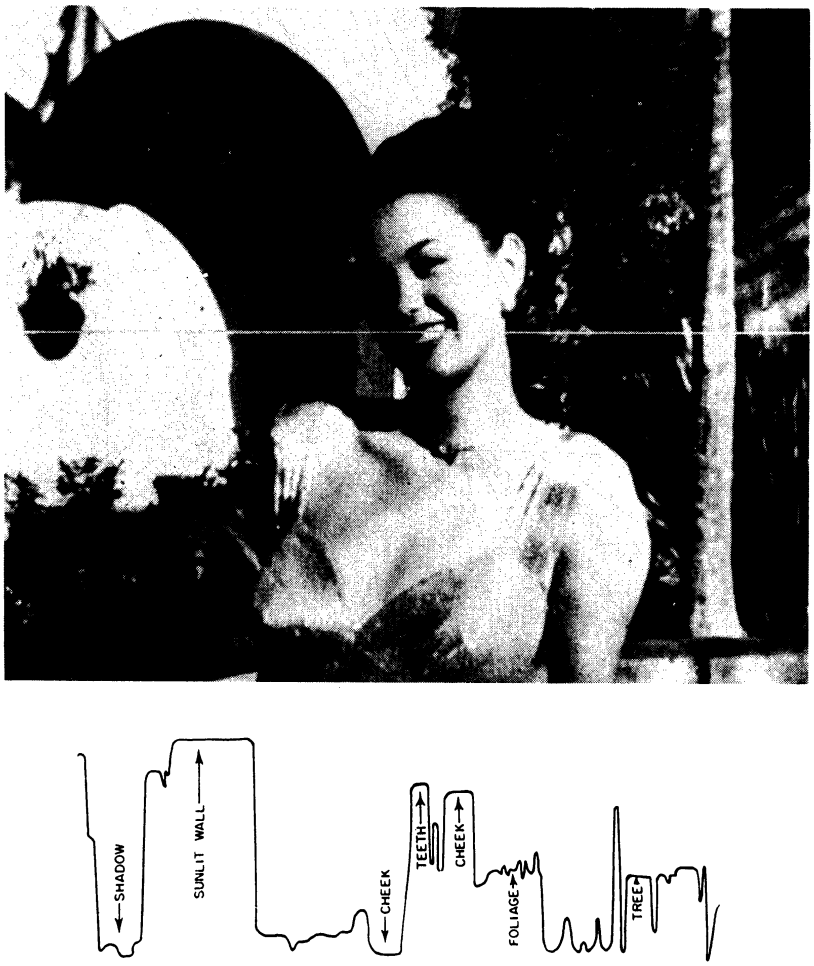


FIGURE 1-5. THE SCANNING TECHNIQUE. The line through the center of the picture represents one scanning line. As the scanning spot explores this line from left to right, it generates a signal proportional to the brightness at each point. The diagram shows the signal plotted as signal amplitude (vertical) against distance along the line (horizontal). (From Television Engineering by Donald G. Fink. Copyright 1952 by McGraw-Hill. Used with permission of McGraw-Hill Book Company.)

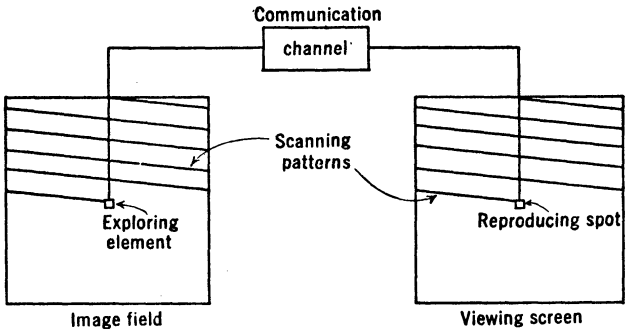


FIGURE 1-6. FUNCTIONAL REPRESENTATIVE OF A TELEVISION SYSTEM. (From Television by V. K. Zworykin and G. A. Morton. Used with permission of John Wiley and Sons.)

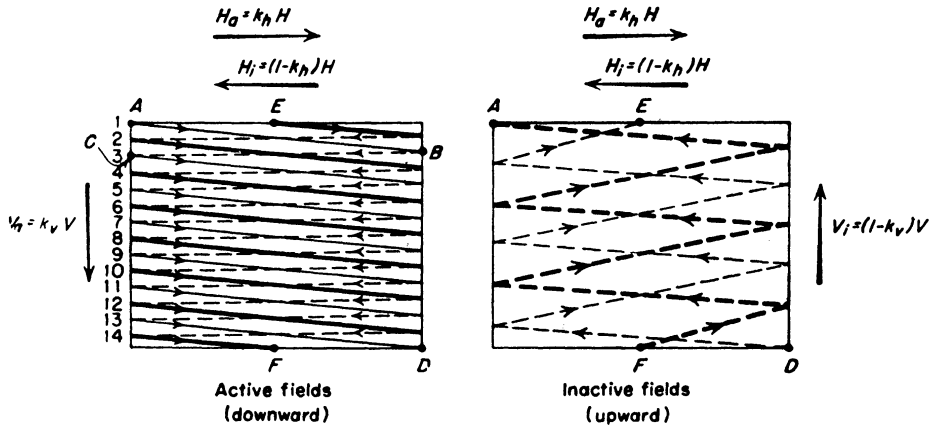


FIGURE 1-7. GEOMETRY OF THE INTERLACED SCANNING PATTERN. (From Television Engineering by Donald G. Fink. Copyright 1952 by McGraw-Hill. Used with permission of McGraw-Hill Book Company.)

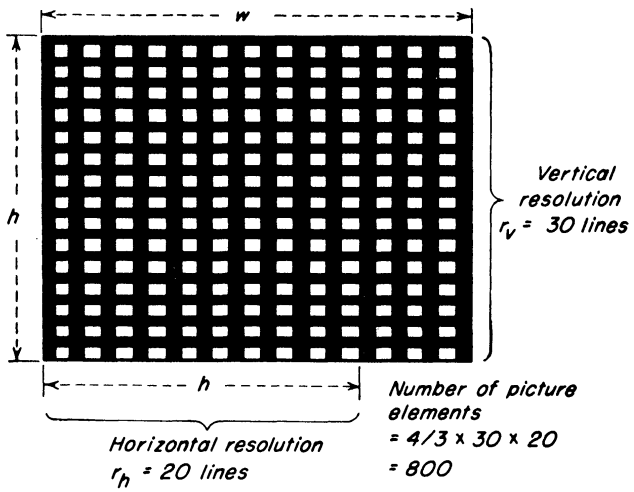


FIGURE 1-8. VERTICAL AND HORIZONTAL RESOLUTION. Vertical resolution is expressed as the number of white and black bars in the height of the picture; horizontal resolution as the number of bars in 3/4 the width of the picture (aspect ratio 4 by 3). (From Television Engineering by Donald G. Fink. Copyright 1952 by McGraw-Hill. Used with permission of McGraw-Hill Book Company.)

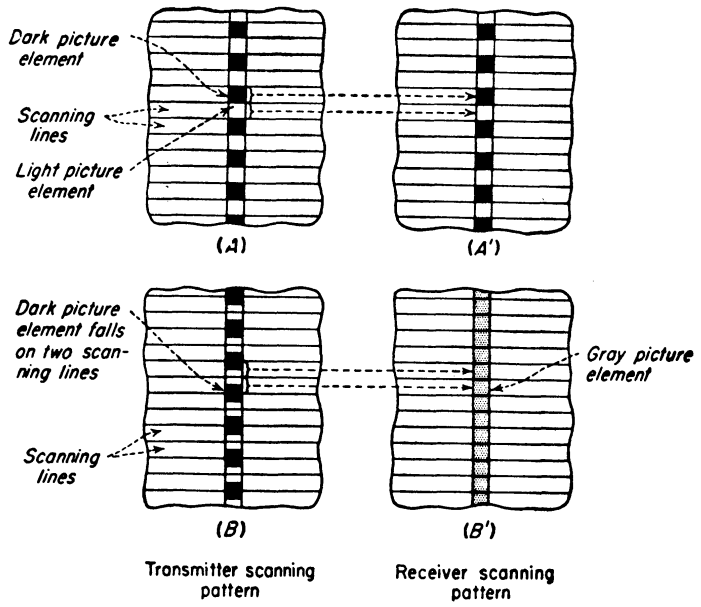


FIGURE 1-9. RELATIONSHIP OF ACTIVE SCANNING LINES TO VERTICAL RESOLUTION. If the scanning lines pass directly over the picture elements, the reproduction A' is like the original A. If, however, the scanning lines straddle the picture elements as in B, the reproduction B' may wholly lack detail. (From Television Engineering by Donald G. Fink. Copyright 1952 by McGraw-Hill. Used with permission of McGraw-Hill Book Company.)

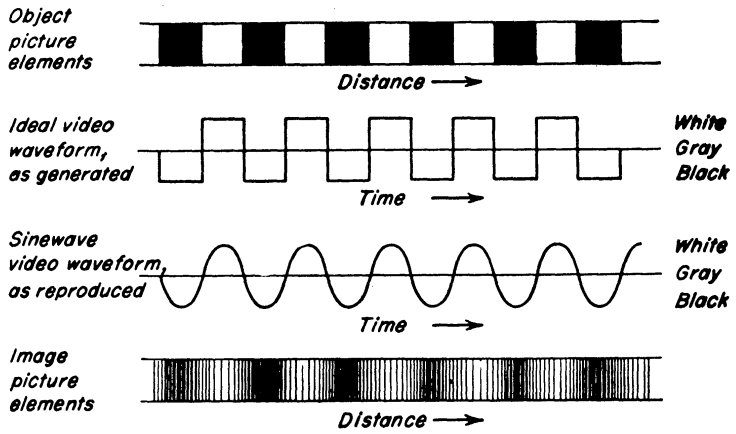


FIGURE 1-10. REPRODUCTION OF HORIZONTAL DETAIL: IF THE OBJECT CONTAINS ALTERNATE BLACK AND WHITE SQUARES, THE IDEAL VIDEO WAVEFORM IS A SQUARE WAVE. Owing to the limited frequency band, the square wave is reproduced at the receiver as a sine wave, and the image picture elements are correspondingly blurred, as shown at the bottom of the diagram. (From Television Engineering by Donald G. Fink. Copyright 1952 by McGraw-Hill. Used with permission of McGraw-Hill Book Company.)

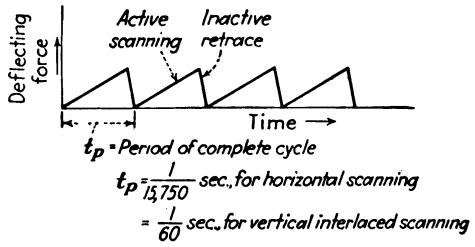


FIGURE 1-11. DEFLECTING FORCE USED IN SCANNING AS A FUNCTION OF TIME (SO-CALLED "SAW-TOOTH" WAVE). (From Television Engineering by Donald G. Fink. Copyright 1952 by McGraw-Hill. Used with permission of McGraw-Hill Book Company.)

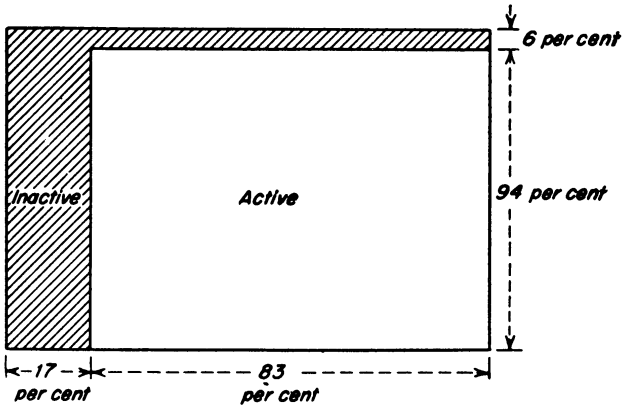


FIGURE 1-12. PORTIONS OF THE IMAGE LOST DURING THE INACTIVE SCANNING INTERVALS. (From Television Engineering by Donald G. Fink. Copyright 1952 by McGraw-Hill. Used with permission of McGraw-Hill Book Company.)

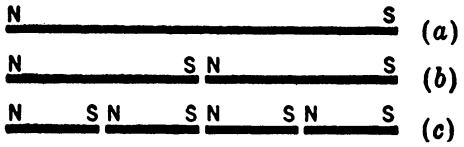


FIGURE 2-1. NEW POLES APPEAR AT EACH POINT OF DIVISION OF A BAR MAGNET. (From Electromagnetics by John D. Kraus. Copyright 1953 by McGraw-Hill. Used with permission of McGraw-Hill Book Company.)

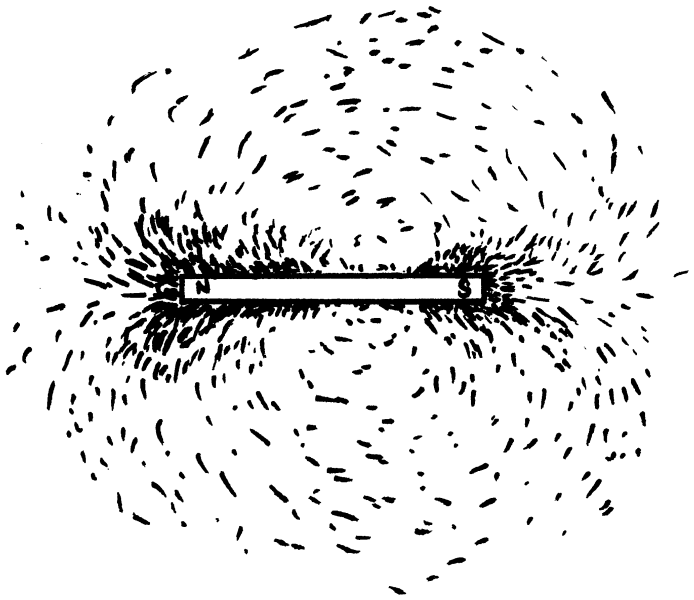
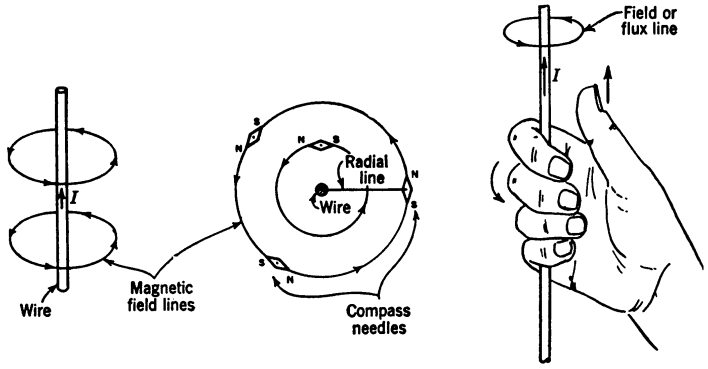


FIGURE 2-2. IRON FILINGS ALIGNED BY THE MAGNETIC FIELD OF A BAR MAGNET



(a) Magnetic Field Around Wire Carrying a Current

(b) Cross Section Perpendicular to Wire. The Current is Flowing Out of the Page

(c) Right-Hand Rule Relating Direction of Field or Flux Line (Fingers) to Direction of Current I (Thumb)

FIGURE 2-3. (FROM ELECTROMAGNETICS BY JOHN D. KRAUS. COPYRIGHT 1953 BY McGRAW-HILL. USED WITH PERMISSION OF McGRAW-HILL BOOK COMPANY.)

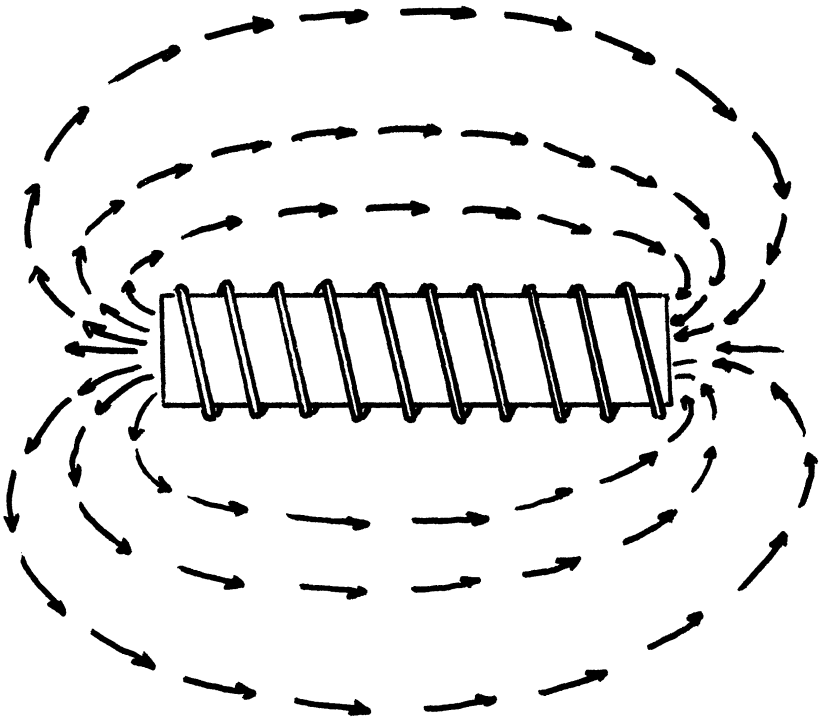


FIGURE 2-4. PLOT OF MAGNETIC FIELD OF A SOLENOID CARRYING A CURRENT

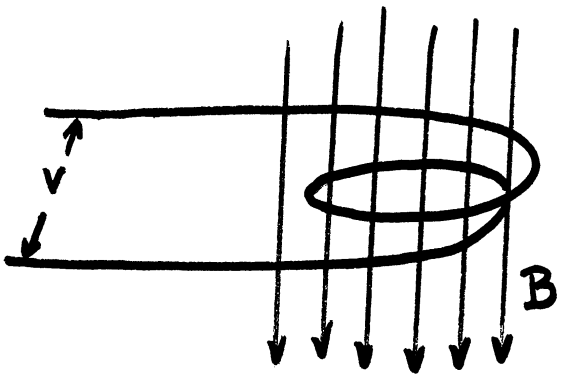


FIGURE 2-5. THE VOLTAGE V IS DEVELOPED ACROSS THE COIL ENDS WHEN AN INCREASING OR DECREASING MAGNETIC FIELD PASSES THROUGH THE COIL.

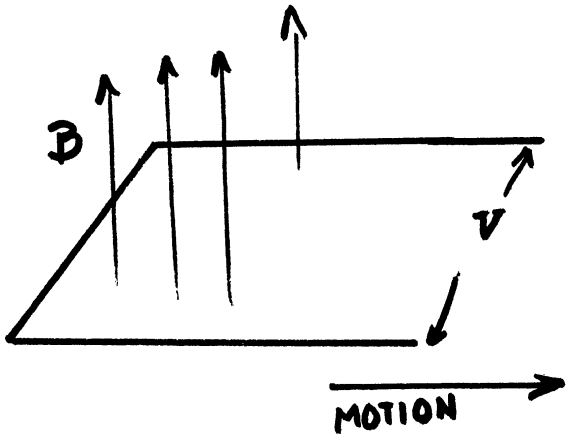


FIGURE 2-6. THE VOLTAGE V IS DEVELOPED ACROSS THE LOOP ENDS WHEN THE LOOP IS MOVED ACROSS A FIXED OR STEADY MAGNETIC FIELD.

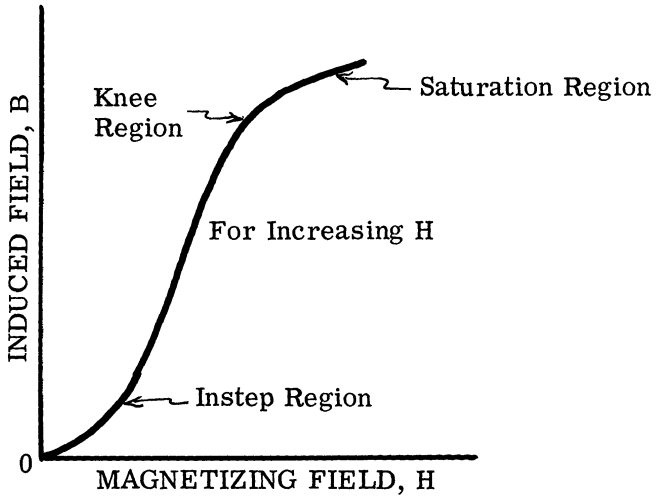


FIGURE 2-7. A NORMAL MAGNETIZATION CURVE. An increasing magnetizing force H produces in a virgin magnetic material an induction B in accordance with the normal magnetization curve.

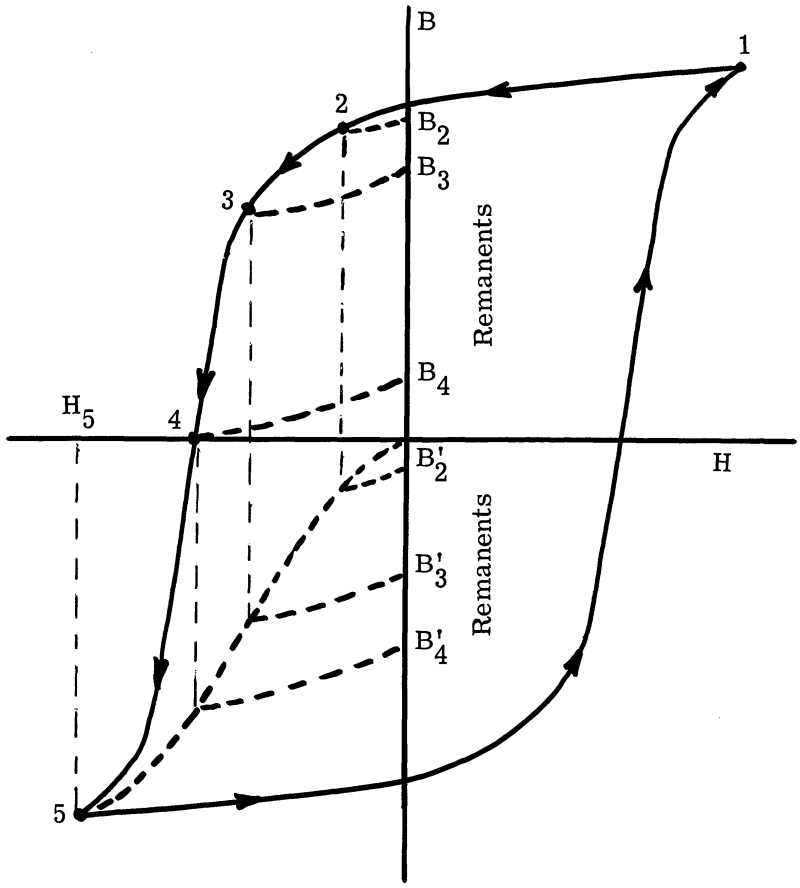


FIGURE 2-8. A HYSTERESIS CURVE. When the magnetizing force H is cyclically increased and decreased, the magnetic induction B in a ferromagnetic material varies in accordance with the hysteresis loop.

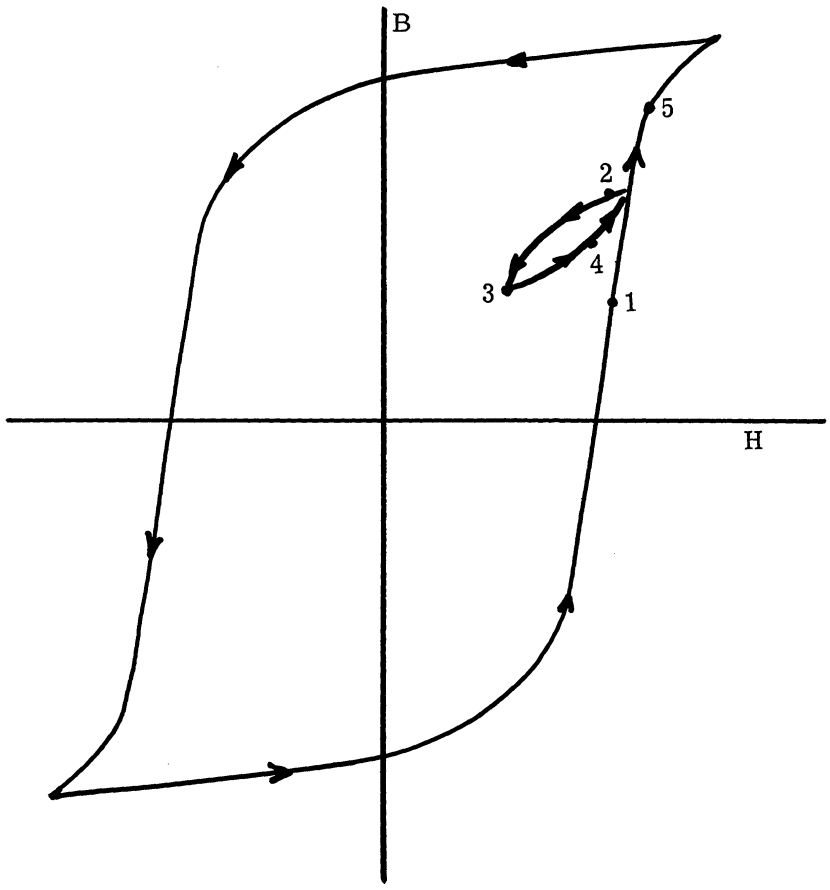


FIGURE 2-9. THE INDUCTION B VARIES IN ACCORDANCE WITH THE MINOR HYSTERESIS LOOP WHEN THE MAGNETIZING FORCE UNDERGOES A SMALL CYCLIC REVERSAL AT ANY POINT IN THE MAGNETIZING CYCLE OF THE MAJOR HYSTERESIS LOOP

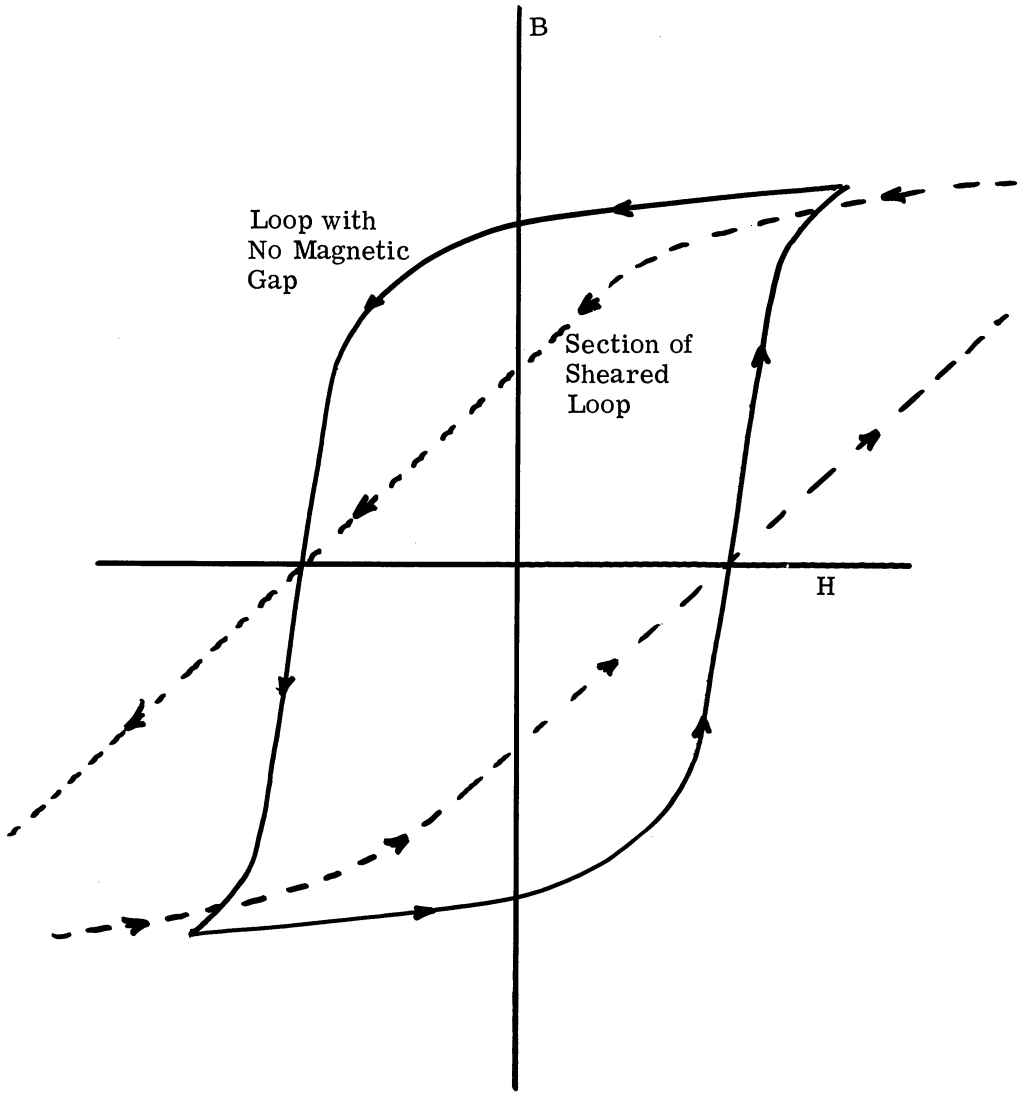


FIGURE 2-10. A SHEARED HYSTERESIS CURVE. In a magnetic circuit of ferromagnetic material having an air gap, the magnetization curve is a sheared hysteresis loop which has two points of zero magnetic induction (coercivity) common with the hysteresis loop of a closed ferromagnetic circuit.

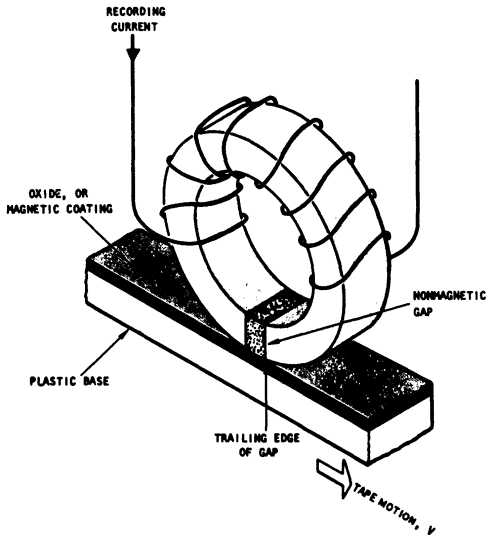


FIGURE 2-11. SIMPLIFIED DIAGRAM OF THE MAGNETIC RECORDING PROCESS. (From Magnetic Tape Recording Handbook, by James A. Howard and Lester N. Ferguson. Used with permission of Hewlett Packard Company.)

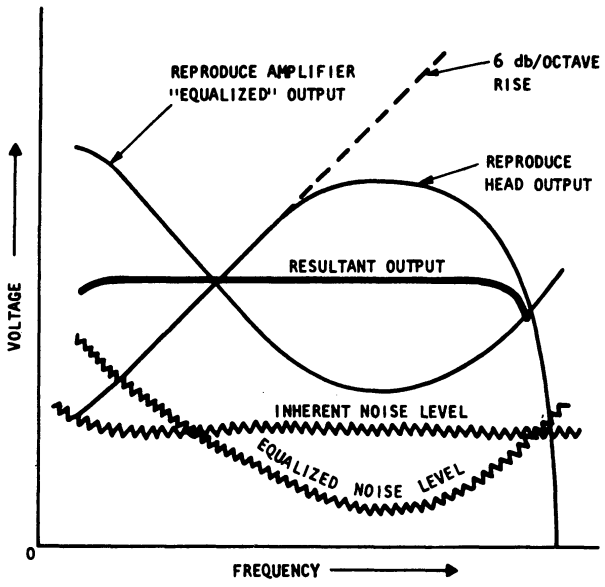


FIGURE 2-12. REPRODUCE CHARACTERISTICS, SIGNAL TO NOISE RATIO, AND LOW FREQUENCY LIMITATION OF DIRECT RECORDING. (From Magnetic Tape Recording Handbook, by James A. Howard and Lester N. Ferguson. Used with permission of Hewlett Packard Company.)

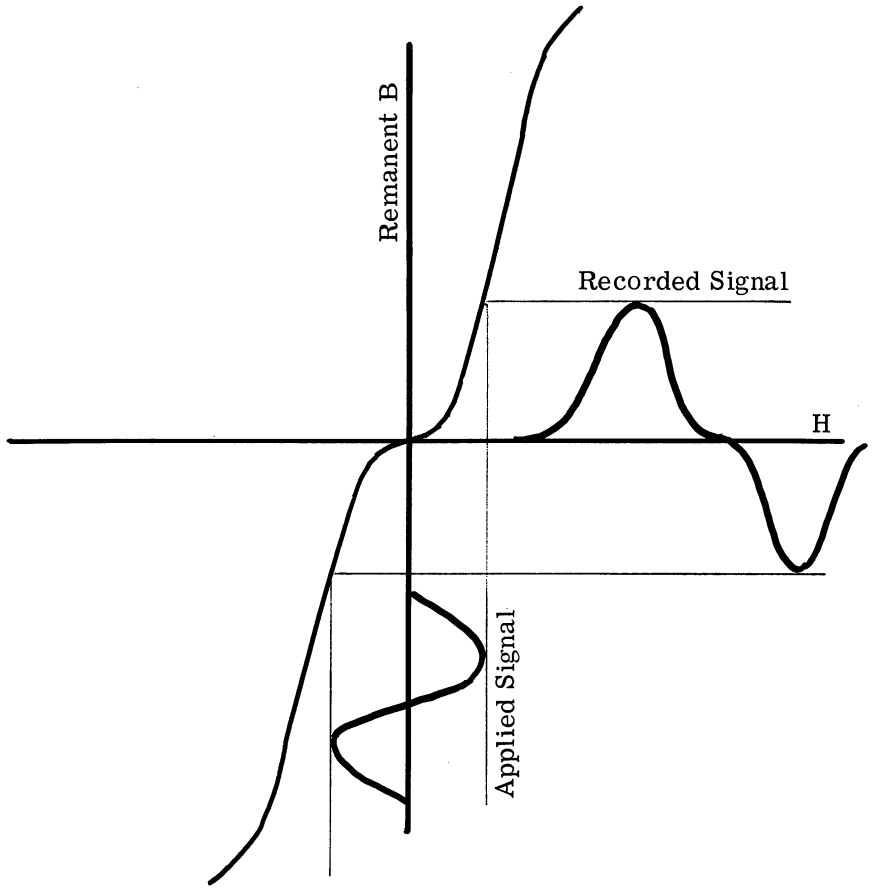


FIGURE 2-13. HEAD-TO-TAPE TRANSFER CHARACTERISTIC FOR LONG WAVELENGTHS OF A VIRGIN MAGNETIC MEDIUM AS FOUND BY PLOTTING REMANENT MAGNETIC INDUCTION VALUES AGAINST MAGNETIZING FORCES (NO BIAS APPLIED). A sinusoidal magnetizing force (applied signal) leaves a distorted remanent-induction wave pattern in the recording medium.

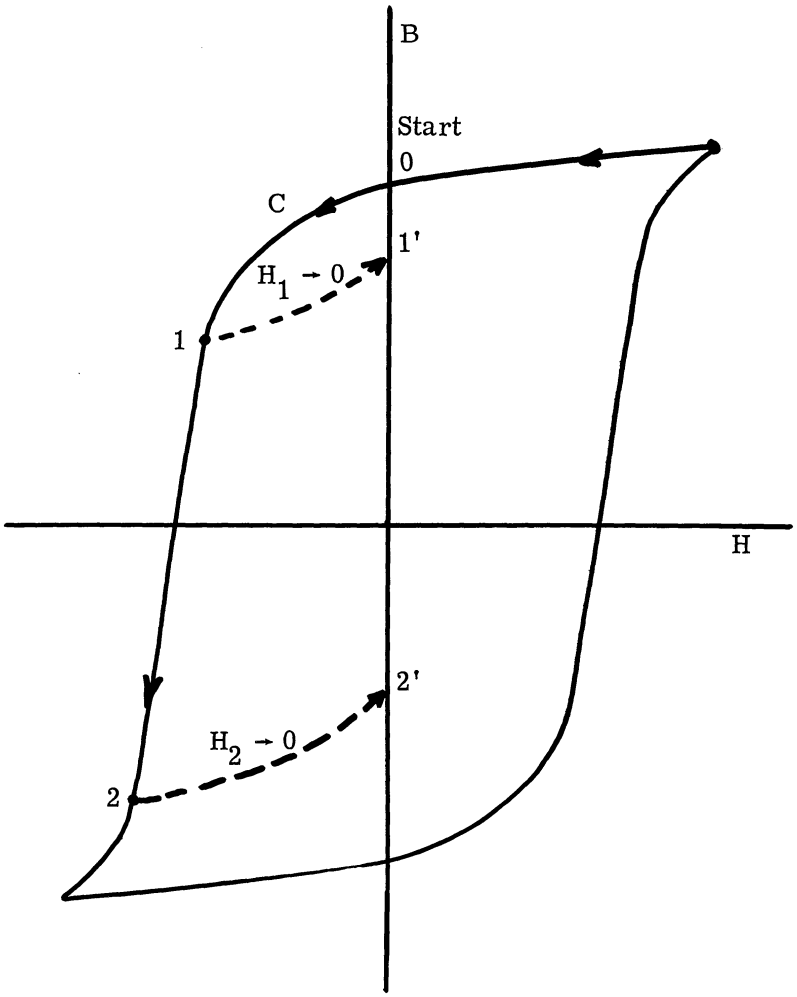


FIGURE 2-14. DIRECT-CURRENT BIASING. If elements of a previously saturated magnetic-recording medium are subjected to a unidirectional magnetizing field of opposite polarity, the induction of each element will follow branch C of the hysteresis curve while under the influence of the recording head.

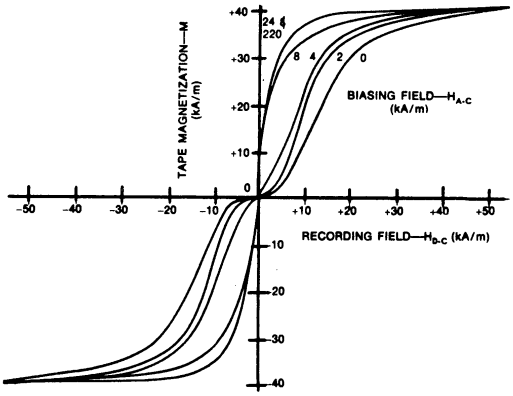


FIGURE 2-15. CHANGE IN THE REMANENT INDUCTION WHEN VARIOUS ELEMENTS OF a-c BIAS ARE ADDED TO A d-c SIGNAL. As the a-c bias is increased the transfer characteristic becomes more linear, and the sensitivity is increased. Increasing the a-c bias beyond a certain value (24 KA/m) has no further effect. The a-c bias is removed before the d-c signal is turned off. (Used with permission of Ampex Corp.)

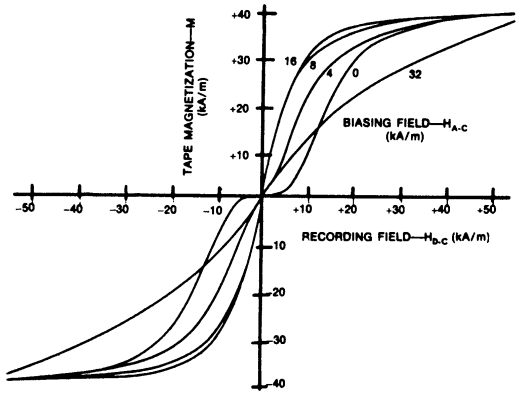


FIGURE 2-16. CHANGE IN THE REMANENT INDUCTION WITH VARIOUS AMOUNTS OF a-c BIAS AS a-c AND d-c SIGNALS DIE OFF TOGETHER. Bias value reaches a peak as bias is increased, and then falls off. (Used with permission of Ampex Corp.)

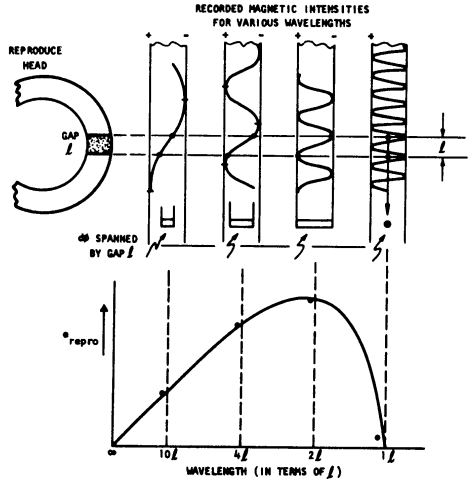


FIGURE 2-17. A GRAPHICAL REPRESENTATION OF GAP EFFECT SHOWING THE RELATIONSHIP BETWEEN THE REPRODUCE HEAD GAP LENGTH ℓ, THE RECORDED WAVELENGTH, THE REPRODUCE FLUX GRADIENT $\frac{d\phi}{dt}$, AND THE REPRODUCE OUTPUT. (From Magnetic Tape Recording Handbook, by James A. Howard and Lester N. Ferguson. Used with permission of Hewlett Packard Company.)

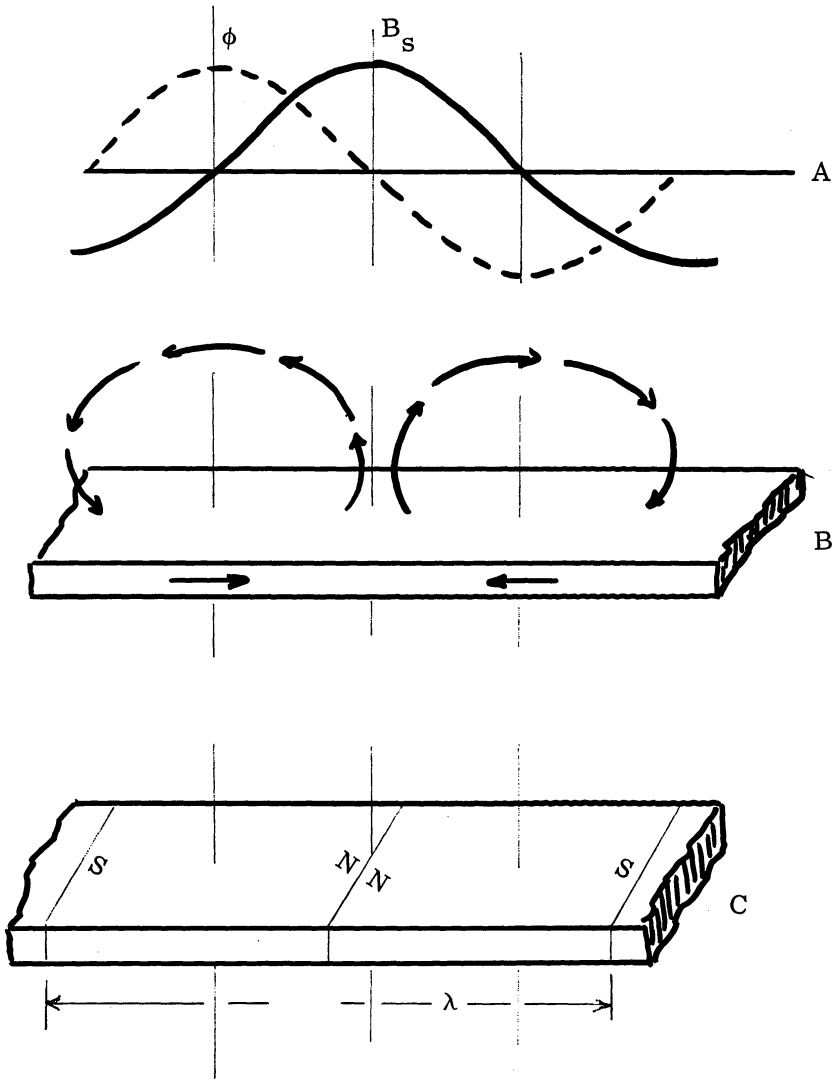


FIGURE 2-18. FLUX DISTRIBUTION (ϕ) IN THE RECORDING MEDIUM AND FLUX DENSITY ON THE SURFACE (B_S) OF THE RECORDING MEDIUM. In many cases it is permissible to represent a recorded sinusoidal signal by two magnets with opposite poles facing each other, where the length of each magnet equals $\lambda/2$.

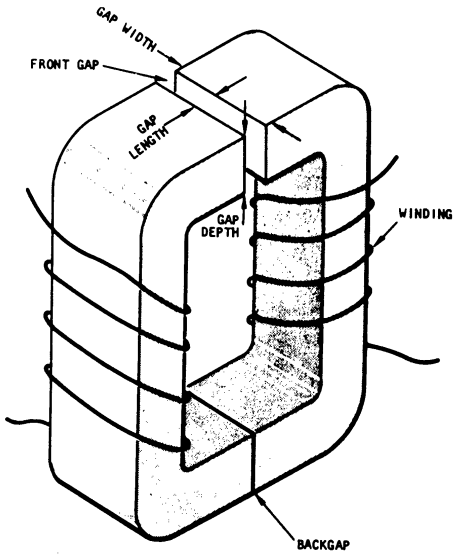


FIGURE 2-19. ENLARGED SKETCH OF A TAPE RECORDER HEAD SHOWING THE DEFINITION OF THE DIMENSIONS. The gap length is measured along the track in the direction of the tape travel. The gap width determines the track width on the tape (by definition). (From Magnetic Tape Recording Handbook, by James A. Howard and Lester N. Ferguson. Used with permission of Hewlett Packard Company.)

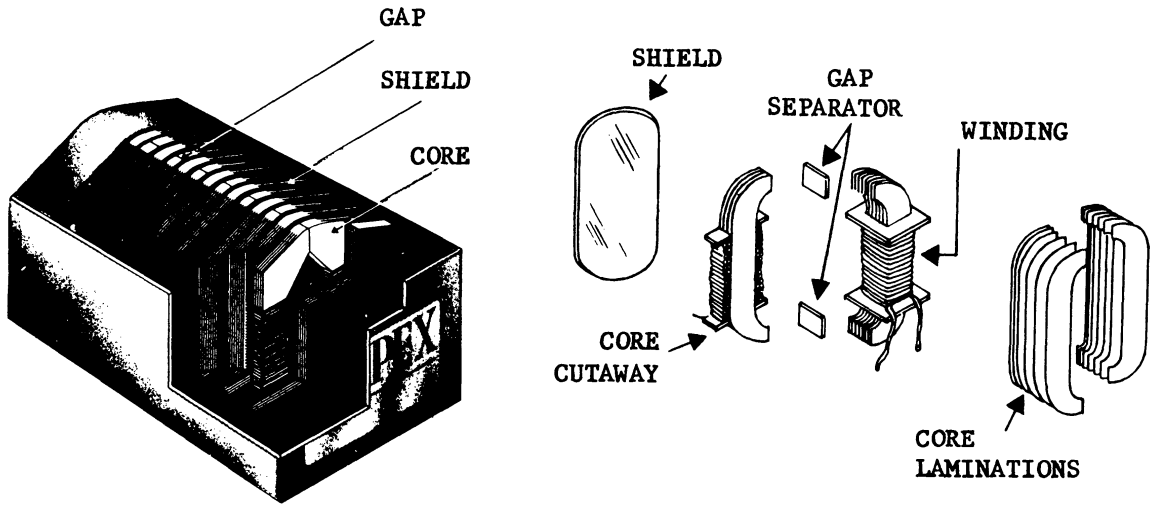


FIGURE 2-20. CUTAWAY VIEW OF A TYPICAL MULTI-TRACK HEAD STACK. (Used with permission of Ampex Corp.)

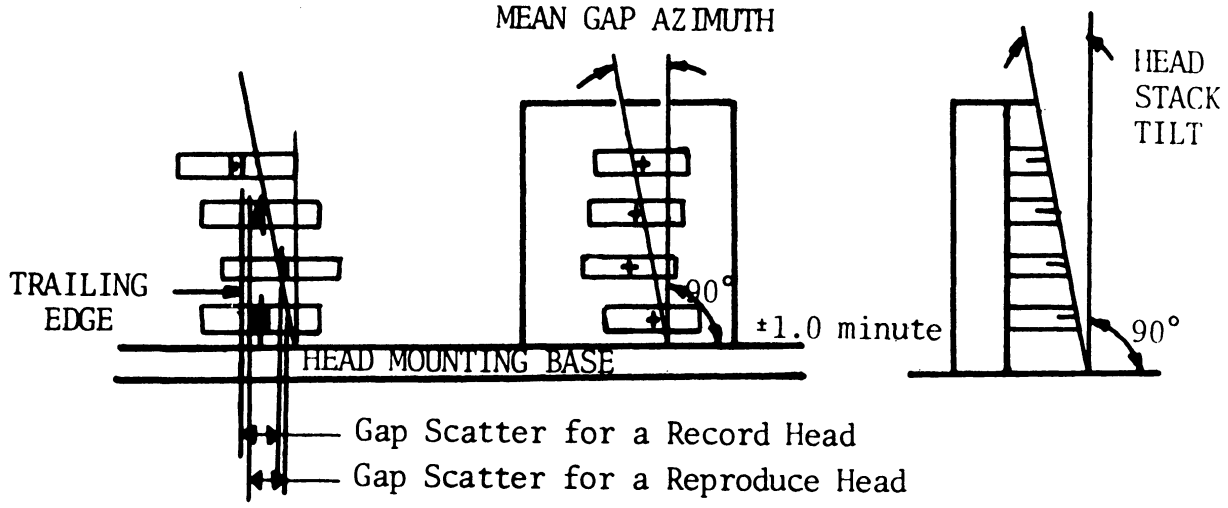
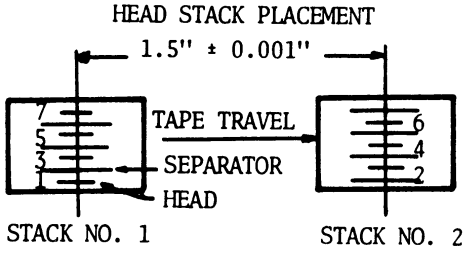
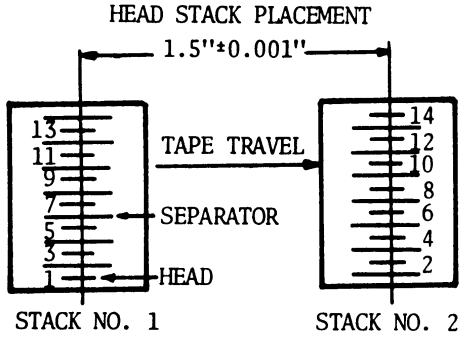


FIGURE 2-21. GAP SCATTER, MEAN GAP AZIMUTH, AND HEAD STACK TILT SPECIFICATIONS FOR MULTI-TRACK HEADS. (Used with permission of Ampex Corp.)

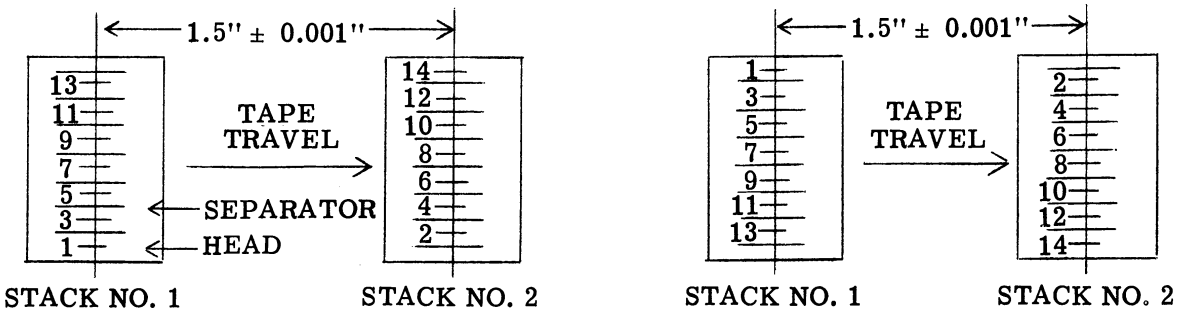


7-TRACK ANALOG HEAD
AND STACK CONFIGURATION



14-TRACK ANALOG HEAD
AND STACK CONFIGURATION

FIGURE 2-22. STANDARD ANALOG TRACK FORMATS FOR 1/2" AND 1" TAPES AS SPECIFIED BY IRIG* 106-66. (Used with permission of Ampex Corp.)



IRIG STANDARD

OLD AMPEX STANDARD

FIGURE 2-23. COMPARISON OF IRIG AND OLD AMPEX STANDARD HEAD CONFIGURATIONS FOR 1" ANALOG TAPE. Interchanging tapes between these two types serves to introduce a 3" spacing between even and odd tracks (double the interleaved head stack spacing).

*Inter-range Instrumentation Group

FIGURE 2-24. SKETCH ON AN ELECTRON MICROSCOPE PRESENTATION OF MAGNETIC OXIDE PARTICLES, SHOWING TYPICAL SIZE AND SHAPE. (Used with permission of Ampex Corp.)

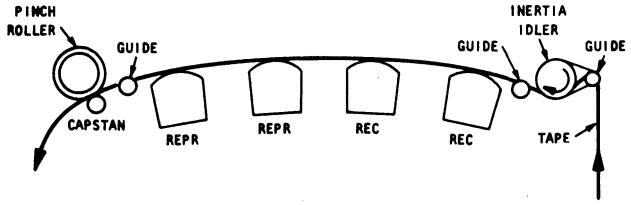


FIGURE 2-25. THE SIMPLEST TYPE OF TAPE TRANSPORT DRIVE CALLED THE OPEN-LOOP DRIVE. (Used with permission of Hewlett Packard Company.)

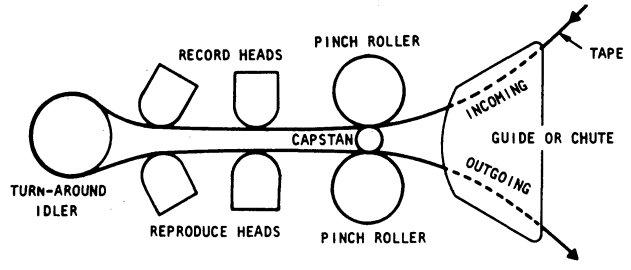


FIGURE 2-26. THE CLOSED-LOOP DRIVE. (Used with permission of Hewlett Packard Company.)

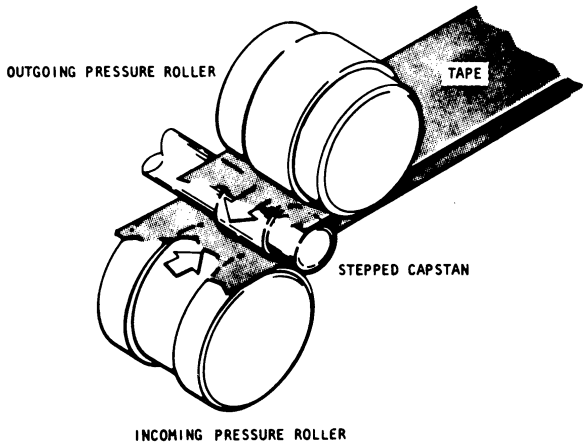


FIGURE 2-27. THE TWO-DIAMETER CAPSTAN WITH MATCHING PINCH ROLLERS. This is a method used to maintain tension within a closed tape loop. (Used with permission of Hewlett Packard Company.)

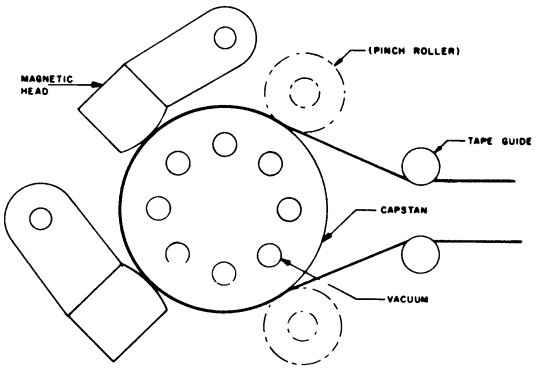


FIGURE 2-28. THE ZERO-LOOP TAPE DRIVE. The magnetic heads are spring loaded against the capstan reducing the unsupported tape length to a minimum. (Used with permission of Ampex Corp.)

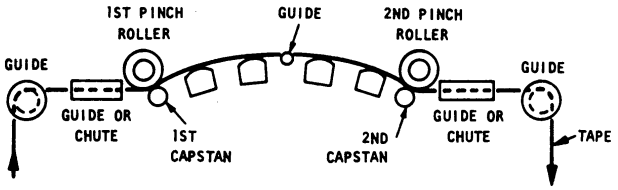


FIGURE 2-29. THE DUAL-CAPSTAN CLOSED-LOOP DRIVE. This design may function as a normal closed-loop drive or in the manner of the two-diameter capstan drive by using different peripheral speeds for the two capstans. (Used with permission of Hewlett Packard Company.)

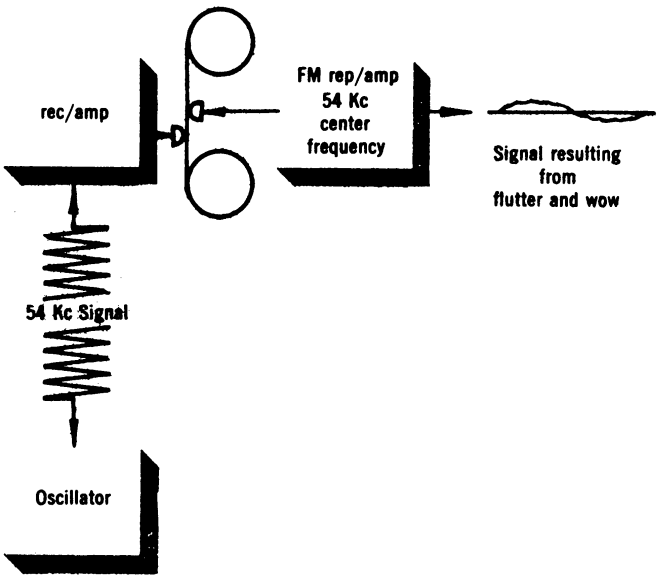


FIGURE 2-30. ONE METHOD USED TO MEASURE FLUTTER. Any derivation from zero output (above the inherent noise level of the rep/amp) is due to flutter. (Used with permission of Ampex Corp.)

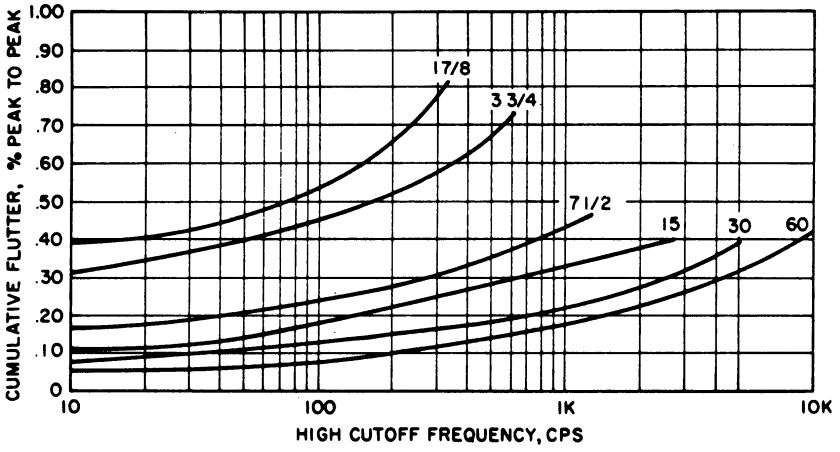


FIGURE 2-31. A TYPICAL SET OF CUMULATIVE FLUTTER CHARACTERISTICS MEASURED WITH A FIXED LOWER BAND LIMIT AND A VARIABLE HIGH CUTOFF FREQUENCY. (NASA)

FIGURE 2-32. PICTURE OF PEAK-TO-PEAK FLUTTER AS IT APPEARS ON THE FACE OF AN OSCILLOSCOPE. Upper spikes represent tape speed-ups during playback (higher frequencies of the recorded carrier, producing positive voltages) and lower spikes represent speed reductions. The lower frequency (rate-of-change) speed variations may be seen as a modulation envelope of the groups of spikes. (Used with permission of Ampex Corp.)

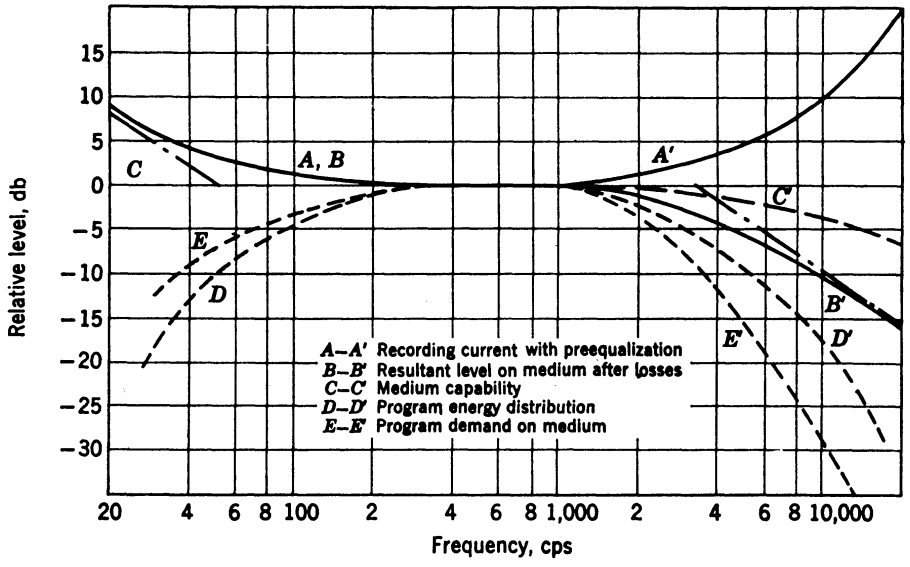


FIGURE 2-33. PREEQUALIZATION IN AUDIO RECORDING. (From Magnetic Recording Techniques by Earl W. Stewart. Copyright 1958 by McGraw-Hill. Used with permission of McGraw-Hill Book Company.)

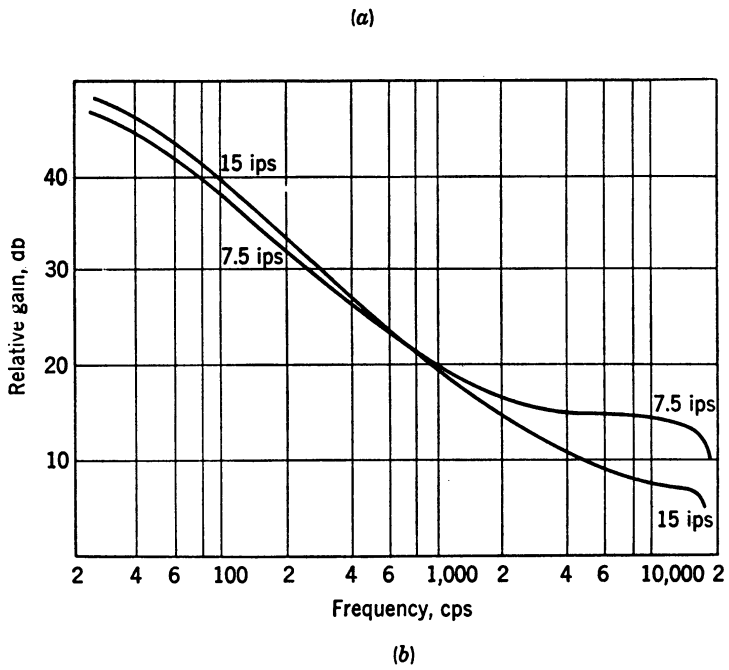


FIGURE 2-34. AUDIO REPRODUCE CURVES WITH THE PROPER POSTEQUALIZATION TO PRODUCE OVERALL FLAT FREQUENCY RESPONSE FROM THE RECORDED FLUX ON THE MEDIUM AS SHOWN BY CURVE BB' IN FIGURE 2-33. (From Magnetic Recording Techniques by Earl W. Stewart. Copyright 1958 by McGraw-Hill. Used with permission of McGraw-Hill Book Company.)

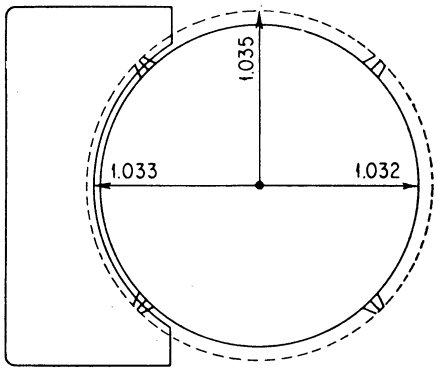


FIGURE 2-35. PROFILE VIEW OF THE ROTARY DRUM AND THE FOUR VIDEO RECORD-REPRODUCE HEADS. Note that the pole pieces of the heads protrude about 3 mils above the surface of the drum. The heads are forced into intimate contact with the heads by the pressure of the female guide. (Used with permission of the National Association of Broadcasters.)

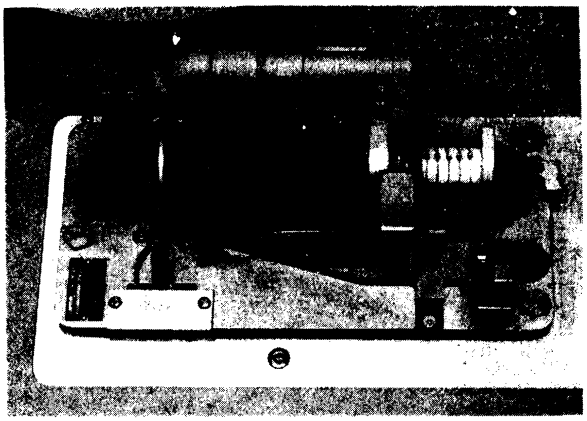


FIGURE 2-36. THE VIDEO ROTARY-HEAD ASSEMBLY. The long central cylinder is the motor which spins the heads at 14,400 RPM. The head drum is opposite the female guide, and is followed by the ceramic slip ring and brush structure. The large hose and tube supply vacuum for pressing the female guide against the rotary head. (Used with permission of the National Association of Broadcasters.)

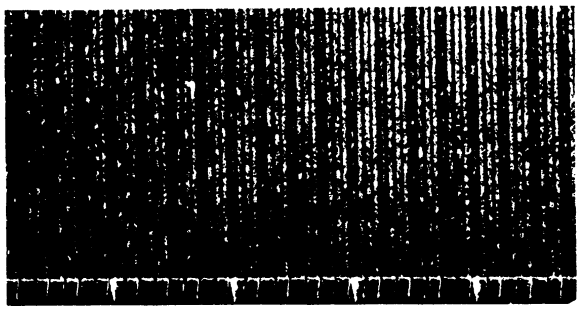


FIGURE 2-37. LOWER SECTION OF RECORDED VIDEO TAPE AFTER TREATMENT WITH MAGNETIC PARTICLES TO RENDER THE RECORDING VISIBLE. Video tracks are 10 mils wide, 15 2/3 mils center to center. The control track is at the bottom. Bright traces on the control track are editing pulses, occurring once each field, 1/4" apart. (Used with permission of the National Association of Broadcasters.)

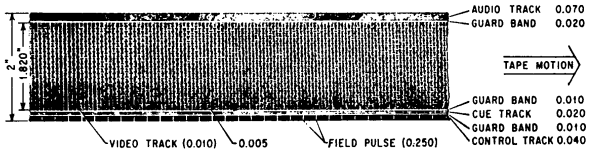


FIGURE 2-38. CURRENTLY USED TRACK DIMENSIONS FOR RECORDED TV TAPE. (Used with permission of the National Association of Broadcasters.)

SECTION V
OPTICAL-MECHANICAL SCANNERS

1. INTRODUCTION

In other sections we have discussed the basic nature of electromagnetic radiation as a radiated force field, with its associated wave motion and transfer of energy. We have examined the generation, absorption, reflection, transmission and detection of this radiation. A thorough understanding of the fundamental principles involved in each of these phenomena is essential to successful application of remote sensing techniques. However, the process of detection is somewhat special, representing the "handle" whereby we sample this force field and derive meaningful information from various parts of the electromagnetic spectrum. Through detection (and subsequent quantitative measurement of the properties) of this radiation we can obtain information concerning the nature of our environment, its material constituents, and their interrelations.

As detection and measurement technology advances, performance limitations arise in other areas, such as atmospheric absorption or rate of data interpretation. Nevertheless, historically, detector limitations have determined not only our use, but even our awareness, of certain aspects of the electromagnetic force field. For example, our rather complete familiarity with one particular portion of the electromagnetic spectrum, that which we know as "light" or "visible light", is due to the availability of a high performance detector (the human eye) ideally adapted to respond to energy in the insolation force field. That the eye is well adapted to this task is readily appreciated upon comparison of the human visibility curves with that for solar illumination (see fig. 1-1). The amazing amount of detailed information in what we "see", all the myriad variations in tone and shape, is contained in just the "visible" portion (somewhat less than one octave) of an infinite electromagnetic spectrum.

The silver halide crystal, which constitutes the "sensitive element" of photographic film, is a second radiation detector of rather widespread familiarity. Like the human eye, it too responds to only a restricted (although somewhat larger) portion of the electromagnetic spectrum.

The function of a camera is to create an image or miniature replica at the "image plane" (the plane of the film emulsion) of the actual scene, as observed in the "object plane". The density, D, in each minute area of the film following the irreversible exposure-development process is related to the irradiance of radiation falling on that particular area, and this in turn is proportional to the irradiance of radiation reaching the camera lens and, hence, the radiance of the corresponding region of the actual scene.

If a very small detector element is placed in the image plane of a camera so that the detector output signal is at all times proportional to the irradiance at that position in the image plane, then the value of the output signal can be used to vary the intensity of a point source of light (such as a cathode ray tube spot) which in turn can be recorded by a second camera to produce the density, D, on the film in the second camera. Now, if the detector element is moved across the film plane of the first camera in some area-covering systematic manner and if the cathode ray tube light spot is moved in exactly the corresponding manner on the cathode ray tube, the total exposure of the film of the second camera will result in an image of the image in the first camera focal plane. Such a scanning system is called an image plane scanner. The camera orientation is held fixed relative to the scene and the scanning function is performed on the image plane side of the objective optics. It is possible and quite usual to perform the scanning function in a different way. If the small detector element is fixed in the center of the image plane, the camera itself can be rotated in azimuth so that the irradiance at the center of the image plane is due to the radiance of the parts of the scene toward which the camera is directed at each moment. When the desired azimuthal rotation has been completed, the polar angle of the camera is changed slightly and the azimuthal rotation is repeated for an adjacent line of position in the scene. The cathode ray tube light spot must change in position to correspond to camera orientation. However, rotation of the complete optical system with high speeds could create difficult engineering problems. Frequently, the camera is pointed at an external flat diagonal mirror so that only the reflection of the external scene is viewed. As the external mirror is rotated, the portion of the scene viewed by the camera appears to rotate relative to the fixed camera. The same scanning function is accomplished by rotating a single mirror on the object plane side of the objective optics. Such a scanner is called an object plane scanner.

There are four principal reasons why optical-mechanical scanners are useful for remote sensing:

1. Detector elements exist which can respond to radiation of any wavelength of electromagnetic radiation which can be brought to a reasonable focus. The photographic film has a restricted spectral response.
2. Detector elements exist which produce output signals accurately in proportion to the incident irradiance. These detectors can be calibrated so that absolute irradiance values can be established. Radiometric measurements can be made.
3. The entrance slit of a spectrometer can be placed at the position normally occupied by the detector element so that the spectral distribution for each point can be measured over the entire spectral range capacity of the spectrometer (degree of polarization can also be measured).

4. The output signals are usually in electrical signal form so that direct storage of quantitative data can be made on tape, and signal analysis using analogue and digital computers can be made either at the time of reception or later to produce processed map-like displays of items of interest.

Loss of excellent geometric fidelity generally occurs for scanners, however, due to the inaccuracy of positioning the cathode ray tube spot relative to the scanner position; and a complete recording of a given area requires considerably more time than the exposure time of a good aerial mapping camera. Image-motion compensation becomes a critical factor for scanners.

Although applications for both image plane and object plane scanning exist, most optical-mechanical scanners are of the latter type, due to several interrelated effects. First, reflective optics are more effective than are refractive optics for use over a large wavelength range. There is no chromatic aberration and no transmission loss. However, reflective optics are affected by serious aberrations as one moves "off-axis", in the image plane. Consequently, in order to provide spectral versatility it is desirable to use reflective optics and to scan their field of view on the object plane side of the objective mirror while reducing aberration effects by maintaining the detector (or detector array) fixed, on or near the optic axis of the objective mirror.

Within the basic subdivisions of image and object plane scanning are a great many scanner designs for specific applications. Scan patterns can be classified as circular or spiral, cycloidal, and rectilinear or raster. Each has certain characteristics that suit it to a particular task. For instance, a simple circular or rectilinear pattern is sufficient to create an image, whereas a more complex, repetitive scan pattern would be better suited to a nonimaging search or homing device.

In addition, instruments have been devised to perform a specific scanning task without internal mechanical scanning, deriving their scan from platform motion alone. An example is the so-called "push broom scanner", where in a linear array (or mosaic) of detectors is oriented normal to the vehicle ground track so that the fields of view of the detectors are "swept" along the terrain in a manner analogous to a push broom.

It is apparent then that an "optical-mechanical scanner" represents no single easily defined device, but rather an extensive set of instruments which have been evolved to meet the widely diverse needs of many applications.

2. THE OPTICAL-MECHANICAL SCANNER

2.1. CLASSIFICATION OF OPTICAL SENSING TECHNIQUES: SCAN MODES

A radiometer containing a single detector element measures radiation arriving at the receiver aperture from a small field of view equal to the size of the projected image of the detector in the object plane. This small field of view is called the instantaneous field of view to distinguish it from a total field of view obtained by moving the radiometer itself, or by scanning this instantaneous field of view across the object plane. If this instantaneous field of view is held fixed relative to the object plane, a single detector element can measure temporal changes in the radiation arriving from a given spatial point. Such a device is useful to detect the presence (and/or change) of a radiating body in its field of view or as a monitor to record temporal events such as a fluctuation in stellar radiation or the rising temperature of a molten metal. As such it is quite properly a part of remote sensing instrumentation, and a member of a more extensive set of radiometric devices which measure the spatial and/or spectral distribution of radiation. Each specific remote sensing application will include circumstances such as the type of data to be obtained, the duration and repetition of observation, and the desired data format, which will determine the device design best suited to the task.

Because of the very large number of optical sensing techniques which may be devised, it is desirable to develop some means of characterization of different types, their individual capabilities and the general data format they provide. While a complete diagrammetric presentation is impractical (if not impossible) figure 2-1 illustrates the major points of distinction in a classification of optical sensors—e. g., discrete vs. distributed sensor elements, scanning vs. a fixed field of view, platform motion if any, and the spectral interval or intervals to be observed. Only a single basic type has been pursued to its logical conclusion; that is, each dotted line indicates a continuation into a family of instruments similar to that shown for the discrete sensor, one-dimensional optical-mechanical scanner. The final output of any selected path through the diagram is a data format attainable with that particular device design. For example, a discrete-sensor system employing a single detector element and a one-dimensional scan normal to the ground track of a moving platform can generate a strip map (a line-scanned image) of the terrain. The addition of detector elements in the form of a linear array can result in either a "parametric" strip mapper which generates a series of images, each associated with some value of a specific parameter (such as wavelength), or a framing sequence (such as a real time "moving window" display of the terrain).

The "optical-mechanical scanning" referred to here is a relatively high-speed repetitive scan about a reference axis within the system and is used to systematically cover a specific portion of the object plane. In addition to such high-speed scanning, slower speed "pointing" with respect to the platform may also be employed with any of these systems to center the scan about a desired point or to increase the aerial coverage.

Members of a given "family" of instruments can be considered as extensions of a more basic design as shown in figure 2-2. Here a simple non-scanning radiometric device employing N discrete detectors is systematically evolved into a complex multispectral sensor capable of generating N simultaneous images of a scene. Each image is a graphic presentation of the radiation received in a particular wavelength band. With the addition of platform motion, such as in an airborne application, these N images result in a set of multispectral "moving-window" presentations of the terrain passing below the aircraft. This information can be stored on film as a series of motion-picture sequences and/or stored electronically, via tape recorder, for later processing and analysis (for example the generation of "spectral signatures" of objects in the scene).

Of the many designs and scanning techniques available a simple line scanner is the most widely used, especially for the airborne remote sensing application. This device may be used to illustrate the scanner-scene geometry governing the formation of an image from an airborne or spaceborne platform.

2.2. THE SIMPLE LINE-SCANNER

Consider a simple Cassegrain telescope which focuses radiation on a single square detector positioned on-axis and in the focal plane (see fig. 2-3). If now a plane mirror is positioned in front of the telescope objective, inclined at an angle to the optical axis and rotated about that or a parallel axis, the telescope field of view will undergo a conical scan motion. In the normal case where the scan mirror is inclined 45^{o}, the telescope field of view will sweep out a circle in a plane normal to the telescope and scan mirror rotational axes. If the scanner is placed in a vehicle with these axes horizontal, the rotating field of view will periodically intercept a strip on the ground (fig. 2-4). Further, if the rotational-optical axes are parallel to the vehicular motion, successive strips will be normal to the ground track and separated (center-to-center) by a distance Vt, where V is the vehicular velocity and t the time for one revolution of the scan mirror.

A square field of view (as defined by a square field stop or square detector) of solid angle $\Omega = \beta^2$ steradians will define a strip of width βh directly below the vehicle. Thus to make the scanned strips contiguous (at the nadir) we must have:

$$Vt = \beta h \qquad (1)$$

As the field of view moves away from the nadir, due to rotation α about the optic axis, the strip width w grows proportional to the greater distance to the object plane, as $(1/\cos \alpha)$ or:

$$w = \beta h/\cos \alpha \tag{2}$$

Also the initially square (at the nadir) field of view becomes distorted due to a more rapid growth in the dimension normal to the ground track. This dimension is affected, not only by the greater object distance but also by the changing angle of projection of the instantaneous field of view on the ground, according to $1/\cos^2\alpha$.*

Since $t = 2\pi/\dot{\alpha}$, where $\dot{\alpha}$ is the angular velocity of the scan mirror we can rewrite equation 1 as:

$$\dot{\alpha} = \frac{2\pi (V/H)}{\beta} \text{ radians/sec.} \tag{3}$$

Now, any given scene point will be in the field of view for a "dwell-time" τ given by:

$$\tau = \beta/\dot{\alpha} = \beta^2/2\pi(V/H) \text{ sec.} \tag{4}$$

In this simple scanner there will be a "dead-time" during a part of each revolution when the field of view is directed away from the ground and, except in very special designs, into the housing or carrier vehicle. This dead-time can be decreased by using multifaced scanning mirrors (see fig. 2-5) and/or advantageously used for calibration and synchronization (see section 2.2.1 and also appendix A). The use of a multifaced scanning mirror reduces the angular velocity required of the shaft turning the mirror for contiguity by the number of faces (n) while simultaneously increasing the dwell time on each element of the scene by reducing the dead time. As will be seen later, this has important effects on detector response time requirements and on the signal to noise ratio achievable for a given set of operating characteristics and scene radiance levels.

2.2.1. PRINCIPLES OF OPERATION, PERFORMANCE EQUATIONS. The fundamentals of operation of a scanner can be understood by consideration of this simple form of object plane scanner. Referring again to figure 2-4, radiation, either emitted or reflected, from the terrain within the instantaneous field of view propagates upward in all directions. That part of this radiation which is intercepted by the instrument aperture (the scan mirror of fig. 2-3 or 2-5) is focused on the detector element by the telescope optics. As the scan mirror rotates the instantaneous field of view moves to an adjacent position and the instrument

* This property along with the fact that the aperture of an object plane scanner is always normal to the direction of view results in no cosine fourth fall off effect inherent in the instrument as would be found in the photographic camera and the image plane scanner.

now receives radiation from that part of the scene. The detector responds to this incident radiation by producing an output signal which is proportional to the power received, and varies as the instantaneous field of view moves throughout the scene. By rapidly spinning the scan mirror, radiation is received and measured from a continuous line of instantaneous fields of view of length corresponding to the total scan angle θ. The varying detector output can be amplified and recorded to produce an image of the scene. The recording can be directly on a photographic film by using the amplified detector output signal to modulate a light source whose illumination is focused and scanned across the film, in a manner which duplicates the scanning of the instantaneous field of view in the object plane. Alternatively, the output signal can be recorded in electronic form on magnetic tape, together with appropriate positional information, for later processing and analysis.

Considering for the moment a single-faced scan mirror as shown in figure 2-3, it is obvious that there will be a "dead time" during a part of each revolution when the mirror is oriented to receive radiation from above (normally the interior of the instrument housing). During this portion of each revolution calibration sources can be used to provide a known power input to the detector. Thus, the signal voltage output resulting from radiation from the scene is compared to that caused by radiation from the known calibration source. The radiance of the ground, L_g, can therefore be obtained from the calibration equation:

$$V_g - V_c = R_{0,L}(L_g - L_c). \tag{5}$$

More accurate values require an accounting for transmission losses and radiation emitted by the intervening atmosphere, which are not considered in the simple expression presented here.

As with any measuring device, the scanning radiometer is ultimately limited by its ability to detect a given signal level in the presence of noise; that is, to provide a usable signal-to-noise ratio for signal levels of significance to the measurement at hand. Since the detector output signal varies as scene radiance we can define a "noise-equivalent" spectral radiance, $[L_\lambda(NE)]$, for the scanning radiometer as that scene radiance, in the spectral band of observation, which results in a signal level equal to system noise. In order to derive an expression for this radiance, let L_λ be the spectral radiance of an instantaneous field of view as observed from an altitude h. Assume an instrument aperture A, a focal length f, a square detector of edge length s, and a detector noise-equivalent-power NEP. Considering a ground position near the nadir, the instantaneous field of view (an image of the detector projected onto the object plane) is a square of edge length $S = s(h/f)$.

Ignoring atmospheric losses and emission, the power reaching the instrument aperture from this instantaneous field of view is:

$$L_\lambda S^2 \frac{A}{h^2} \Delta\lambda = P_\lambda \Delta\lambda \tag{6}$$

Assuming this also is the power reaching the detector element, the noise-equivalent spectral radiance is found by setting $P_\lambda \Delta\lambda = (NEP)_{ave.}$ for the wavelength interval of observation:

$$L_\lambda(NE) = \frac{(NEP)_{ave.} h^2}{S^2 A \Delta\lambda} \tag{7}$$

The expression for NEP can be written in terms of the detector detectivity, D^*, so that:

$$NEP = \frac{\sqrt{s^2(\Delta f)}}{D^*} \tag{8}$$

where (Δf) is the electronic fluctuation frequency bandwidth required to record the varying signal. This bandwidth is determined by the number of resolution elements (individual instantaneous fields of view) which must be scanned per second in order to satisfactorily cover the scene.

Assume that complete coverage (contiguous scan lines at the nadir) is desired. For a given platform motion, the number of resolution elements which must be scanned per second increases rapidly as the system resolution is improved (instantaneous field of view is decreased). This is because the number of resolution elements per line increases and more lines are needed to cover the same area in the scene. The amount of time, τ, available for observation of each scene element (dwell-time) is the inverse of the number of elements which must be scanned per second. Also, in order to amplify the signal without serious distortions the electronics must pass all frequencies up to f_{max} where:

$$f_{max} \cong \frac{1}{2\pi\tau} \tag{9}$$

Hence, $f_{max} \cong \frac{1}{2\pi}$ (number of resolution elements scanned per second).

If the angular extent of a scan line is θ and the angular extent of the instantaneous field of view (angular resolution) is β, the number of resolution elements per line is clearly θ/β. The linear width of a scan line (in the direction of platform motion) at the nadir is $h\beta$, a distance which is traversed at the platform velocity V, in time $t = \frac{h\beta}{V}$. Thus for contiguity of the scan lines (at nadir) one must scan $V/h\beta$ lines per second. The total number of resolution elements covered per second is then the number of elements per line times the lines per second or $\frac{\theta}{\beta}\left(\frac{V}{h}\right)$. Thus f_{max} becomes

$$f_{max} = \frac{1}{2\pi}\left(\frac{\theta}{\beta^2}\right)\left(\frac{V}{h}\right) \tag{10}$$

and the noise-equivalent radiance (by combining eqs. 7, 8, and 10 and the fact that $s = f\beta$ and $S = h\beta$) is:

$$L_\lambda(NE) = \frac{\sqrt{s^2\left(\frac{1}{2\pi}\right)\frac{\theta}{\beta^2}\left(\frac{V}{h}\right)}}{D^*} \cdot \frac{h^2}{s^2 A \Delta\lambda}$$

or

$$= \frac{f\sqrt{\frac{\theta}{2\pi}\cdot\frac{V}{h}}}{\beta^2 A D^* \Delta\lambda}. \tag{11}$$

This formula shows the significant relationships of important airborne scanner properties which apply to all spectral bands from the ultraviolet through the infrared. As a review, these are, in the order presented in equation 11: the noise-equivalent spectral radiance, $L_\lambda(NE)$ (that scene radiance, in the wavelength interval of observation, which generates an output signal equal to system noise); the optical focal length, f; angular extent of scan, θ; platform velocity, V; platform altitude, h; angular extent of instantaneous field of view (angular resolution), β; instrument collecting aperture, A; detectivity of detector, D^*; and spectral band of observation, $\Delta\lambda$. This or a similar performance equation is derived in many books and articles (c.f. ref. 1-3) but seldom is expressed in exactly the same manner. No one form has received general acceptance, and the designer must define the various parameters in terms most meaningful to an understanding of the particular application at hand. For example, $\beta = s^2/f$ and $A = \frac{\pi D^2}{4}$ where D is collector diameter, or f/No. = f/D, etc.

A scanning radiometer viewing a scene having a large spectral radiance L_λ will produce a large signal compared to that generated when viewing a hypothetical zero level of spectral radiance. However we more often wish to record small differences in rather large values of radiance between adjacent positions in the scene. In this case a "noise-equivalent difference" is defined as that change in scene radiance from position 1 to position 2 which results in a signal change just equal to the system noise level. In the ultraviolet, visible and near infrared portions of the electromagnetic spectrum, received radiation is due primarily to a reflectance of solar illumination by objects in the scene. In this case it is customary to speak of a noise-equivalent reflectance difference (NE $\Delta\rho$) as that change in reflectance between two scene points which generates a signal-to-noise ratio of one (S/N = 1.0). In the thermal infrared region (beyond about 4 microns, see fig. 2-6) the energy is predominantly self-emission due to the temperature of objects in the scene. Here we may

define a minimum temperature difference between position 1 and position 2, the noise-equivalent temperature difference (NEΔT), such that again, S/N = 1.0. In either case the noise-equivalent difference is a measure of the system's sensitivity to changes in the significant parameter governing scene radiation, and is useful in comparing performance of various instruments in a given application.

2.3. COMPARISON OF THE SCANNER AND THE PHOTOGRAPHIC CAMERA

Other sections on photogrammetry and the photographic process have shown that aerial photography can do an excellent job of establishing the relative position of objects in the scene for the generation of very accurate maps. Reconnaissance cameras provide detailed recordings of the distribution of reflected radiation in black and white, color and "false-color" presentations. Satellite-based cameras have provided important geological information with relatively inexpensive photographic equipment. Compared to the photographic camera of comparable resolution, the optical-mechanical scanner is an expensive, highly complex instrument. Why then, with the demonstrated capability of the photographic process, should one consider the use of a scanning instrument in remote sensing applications?

Clearly no single answer can evaluate all the relative advantages and limitations of either system and a final selection must be based upon the needs and goals of a given application. However, a primary reason for the use of scanning devices is that both spectral and spatial radiometric measurements of good accuracy can be obtained in a form suitable for transmission, storage and automatic computation and analysis. In addition, imagery can be obtained outside the photographic region of the spectrum; detectors in general have a wider dynamic range than photographic film; and the detection process is reversible and readily amenable to calibration thereby yielding quantitative data.

The photographic process is seriously limited in:

1) Precision—the multiple steps required to produce the image and the tendency of the process to become non-linear in almost every property make reproducible quantitative results very difficult;
2) Spectral Range—the photographic process is limited to wavelengths shorter than about 1 micron by self exposure due to internal radiation;
3) Data Format—spectral-spatial information is recorded in the form of chemical deposits (silver or dyes) while computation equipment normally requires an electrical signal with the spectral and spatial information accurately correlated.

However, while scanners are excellent for obtaining radiometric data throughout the entire optical region, they are not competitive with simple camera systems for obtaining spatial or photogrammetric information. This comes from the fact that film is a fixed mosaic

of many detectors and has a much higher data rate than the detector or detector array of a scanner. This fixed array and high data rate of film permits a simple aerial camera to achieve a higher spatial resolution with less distortion than practical scanning systems. As a result, the distance between two objects in a frame can be measured with an accuracy set by lens distortion, image motion during exposure, image blur, and the stability of the film base with regard to shrinkage or stretching. These factors are relatively easy to control compared with the factors which introduce distance distortion by the optical scanner as it scans the scene line by line. Aircraft motion, e.g., roll, pitch, and yaw, together with velocity, and height variations affect the relative position of objects within a scene as observed in the scanner imagery. Compared with the camera distortion factors, precise control of these parameters is extremely difficult. However, scanners are able to produce some photogrammetric data.

2.3.1. COMPARATIVE PHOTOGRAMMETRY. As a further comparison of the photographic vs. scanning process, one should consider various factors as they affect the relative photogrammetric capabilities. Again the relative significance of various effects will depend upon the particular application, and no attempt has been made to weigh their importance here.

The Camera

1. Distortions include aberrations and image motion effects.
2. Aberrations where known can be compensated and parallax can be used to advantage in the generation of stereo imagery.
3. All elements of the scene are viewed simultaneously in mapping cameras and image smear due to platform motion affects all points in the scene.
4. Framing cameras have excellent geometric fidelity for mapping purposes, with a single scale factor relative to the reference plane for the entire image.

The Optical-Mechanical Scanner

1. Aircraft motion must be precisely known to compensate for distortion.
2. Each scene element is observed at a different time.
3. The scale factor along the ground track is linear in distance while that along a scan line (normal to ground track) is linear in angle of view.
4. Shadows may be a function not only of the sun's position during but also prior to observation in the case of thermal images.
5. Moiré fringes often result in viewing scenes having row structure.

Figure 2-7 illustrates the geometric relationships which exist between objects in the object plane and their image in the image plane for the mapping camera (2-7a) and the line scanner (2-7b).

2.4. THE THERMAL INFRARED

Of the various distinctions between operational characteristics of the photographic camera and the optical-mechanical scanner one of the most important is that of spectral range. As noted in the preceding section, photography is limited to wavelengths shorter than about 1 micron due to exposure by the camera's own internal radiation. The discovery, in 1800 by Sir William Herschel, of energy beyond the red end of the visible spectrum and the development of detectors which respond to this radiation with sufficient sensitivity for meaningful measurements led to an entirely "new" source of information within the electromagnetic spectrum, the thermal infrared. Whereas the information derived through the photographic process concerns such parameters as sun angle, surface roughness, and reflectivity, we are now confronted with a whole new set of scene parameters such as temperature, emissivity, specific heat and thermal conductivity. It is not surprising then, that unique and significant information should be contained in these "new" scene parameters.

The thermal infrared region of the spectrum may be thought of as extending from about 4 microns on the lower end (as limited by available energy of emission, see fig. 2-6) to about 22 or 24 microns at the upper (here limited by atmospheric transmission, see fig. 2-8). In this region, imagery is a pictorial display of radiation due primarily to self-emission of objects in the scene. As in the more familiar photographic image, objects are recognized on the basis of shape, texture and context within their environment. Similarly, objects are contrasted in their environment or "background" with other adjacent objects by a difference in radiance level. However, in this case there are primarily two interrelated properties which affect the value of L_λ (the spectral radiance) of a surface; its absolute temperature, T, and emissivity, $\epsilon(\lambda)$. The spectral distribution of radiated power is primarily established by blackbody radiation laws, but can be significantly modified by the spectral dependence of emissivity. In addition, thermal radiation from other objects can be reflected and contributes substantially to the radiance observed for a given surface.

The temperatures of many common objects in nature normally differ from one another by a few degrees as a consequence of their characteristic thermal properties. Under some conditions a particular object could be expected to radiate less power per unit area than its adjacent background while at other times the opposite is true. There are, moreover, conditions under which radiances of the target and background are equal within the range of sensitivity of the scanner. This radiance difference is, however, strongly dependent on the physical structure of both the target and background and on their environment, both past and present. Because of this, IR image interpretation is often severely handicapped in making quantitative assessments of a potential object, since the necessary detailed information about the target and its environment is not usually available. Sometimes a

qualitative argument for the apparent contrast of an object can be made if the nature of the object and some information on the meteorological conditions obtaining prior to the observation time are known. At best, in the general case, only crude after-the-fact justifications for the relative appearance of an object against its background can be given. However, even detailed environmental information is of little use if the time-dependent temperature response of the object to its environment is not known. To be sure, human experience provides us with some feel for the gross characteristics of the response of familiar objects to various environments and intensive training can produce a finer appreciation for the more subtle responses. For example, water bodies tend to be thermally stable and for that reason moderate the local climate. Small, thin objects tend to undergo rapid and extreme temperature changes as the net heat exchange varies. A good approximate rule for interpreting purely thermal infrared images is that those areas which readily show frost in cold clear weather at night are those which will be dark in nighttime thermal images. Grass, tree leaves, and other light vegetation, also certain housetops and automobile rooftops tend to appear dark at night. Objects which tend to hold the heat of the day such as concrete structures and rock outcrops appear lighter at night. Usually the brightest objects in nighttime thermal infrared images are the bodies of standing water such as ponds and shallow lakes.

In daytime, thermal images, which are made by filtering out reflected sunlight, tend to show a reverse in this brightness order. Lightweight objects heat more rapidly and to higher temperatures, heavy dense objects heat more slowly, while water bodies heat most slowly of all. Healthy plants which, due to transpiration, maintain a reduced temperature during the day do not increase in "thermal brightness" as much as less vigorous or dead plants. Similarly, moist soils maintain a lower daytime temperature by water evaporation at the surface.

Six factors govern the thermal exchange of the environment:

1. Solar radiation
2. Sky radiation
3. Ground radiation
4. Heat storage and change of state heat
5. Heat conduction
6. Heat convection.

These factors are shown schematically in figure 2-9. Ignoring atmospheric emission, one factor, ground radiation, is intercepted by the airborne sensor and used for interpretation.

2.4.1. TYPICAL THERMAL IMAGERY. The utility of the infrared portion of the electromagnetic spectrum can best be judged by inspection of typical imagery. The imagery

in figures 2-10, 2-11, and 2-12 was made with The University of Michigan M-5 scanner,* which has been declassified by the U. S. Army Electronics Command.

Figure 2-10 is an 8-13 micron image of an agricultural scene in Texas. It was generated for USDA, Weslaco, Texas, for use in a study of the effects of irrigation practices on crop vigor and yield. Figure 2-11 shows two time-lapse images of a power plant some twenty miles downstream from a dam. Generated for TVA in connection with a study of water flow, the top image shows the river at the time a slug of water was released from the dam while the lower one shows the slug of cold water passing the power plant about 20 miles and 24 hours after its release. Figure 2-12 is typical of imagery generated at Point Barrow, Alaska, in October, 1967, under a grant from the Arctic Institute of North America. These data were collected for a heat budget study of the area. It is interesting to note that the thermal imagery differentiates between snow covered lakes and dried lake beds in the polygonal ground structure.

As a more detailed example of the parametric relations involved in the interpretation of infrared imagery, consider the temperature history of several objects as shown in figures 2-13 and 2-14. The materials in figure 2-13 undergo two major crossover points during a 24 hour cycle. Those materials with low heat capacity and high absorption are first to increase in temperature after sunrise and undergo the widest temperature change. Water with its high heat capacity undergoes relatively little change in temperature. Figure 2-14 suggests that temperature may be used to discriminate between two different crop types if thermal imagery is obtained at the proper time of day.

The small diurnal temperature change of water accounts for its changing appearance (top center) in the bog shown in figure 2-15. The land heats and cools faster than the water so that in the upper image the most pronounced (lightest) tone is the "relatively" warmer water. Just after sunrise, middle image, the relative tone between the water and the surrounding woods is much less distinct, while in the mid-morning image the water is, relatively, the coolest material in the scene. This variable nature of infrared imagery just described should serve as a caution signal for the interpreter of thermal data. There are many conditions and phenomena that can influence the temperature of an object. Perhaps the most significant use of infrared (thermal) imagery is in those applications where an anomaly occurs (that is, an unexpected contrast exists). For example, the detection of a forest fire or volcanic activity is a natural application of infrared sensing. It can be used

* This scanner was made available to The University of Michigan by U. S. Army Electronic Command.

also to map the distribution of a fire line beneath smoke or to delineate a snow bridge over a crevasse or a land-sea interface hidden by snow. Other potential applications include the mapping of oceanic currents, the detection of insect infested trees (which have a higher temperature than healthy trees), the mapping of geologic faults and the mapping of the extent of the thermal pollution in rivers and lakes.

2.4.2. THE THERMAL MODEL. Present scanners have temperature sensitivities of less than one degree C or radiance sensitivities on the order of 10^{-12} watts/m^2 sr. It is clear, therefore, that in order to utilize more efficiently the quantitative information obtainable through infrared remote sensing, and indeed to utilize the sensor optimally, an accurate analysis of the response of objects to time-varying thermal environments is required.

To provide valid estimates of the temperature or radiance of a general three-dimensional target in terms of its environment without introducing very extensive approximations would require much more detailed information about heat transfer coefficients than is presently available even from laboratory-type measurements. The general one-dimensional case (a road, building roof, truck top, etc.) is, however, mathematically tractible and could yield valid estimates of the target and background temperatures and radiances. Many objects and backgrounds can be considered planar, hence the importance of the one-dimensional case. Studies have been conducted to determine the degree to which a reasonable one-dimensional model can be expected to accurately predict time-dependent surface temperatures or radiances. A thermal model developed at the University of Michigan, as a part of such a study, is presented in appendix B.

2.5. MULTISPECTRAL INSTRUMENTATION

In the foregoing we have discussed the optical-mechanical scanner primarily as a device to generate imagery in spectral regions beyond the range of the photographic camera. A major portion of our discussion has been concerned with the scanning of an image of a single detector element (the instantaneous field of view) across the object plane. Such a scanning device receives power in a single spectral band, as defined by the detector response and/or selective spectral filtering prior to detection. The resulting imagery is, then, a graphic presentation of scene radiance in a specific spectral band, wherein objects are recognized and identified primarily on the basis of shape, texture and their environmental relationships to other objects.

In addition to this spatial distribution of radiation, information is contained in its spectral distribution, temporal variation, state of polarization and variations in these parameters with angle of observation. While each of these parameters is a unique source of information and

of significant value in specific applications, the variation in spectral characteristics appears to offer the greatest potential for a wide variety of remote sensing situations.

The basis for this belief in the potential of spectral discrimination rests first on the variability of available measurements. Chemists make extensive use of spectra to identify and differentiate complex materials. Botanists are beginning to use similar techniques. Existing spectra indicate that when measurements are taken over sufficiently wide wavelength intervals, each material will exhibit unique spectral structure. The success of a few applications of very simple spectral discrimination confirms the utility of the methods, e.g. the use of infrared camouflage detection film to capitalize on the spectral variations in reflectivity of green plants and the use of infrared wavelength bands to detect forest fires.

These examples represent a very simple application of the idea of spectral discrimination; that of selecting a single spectral band wherein a given material exhibits some anomalous behavior relative to its environs. Thus, by sensing electromagnetic radiation in this particular band, the contrast between a specific object (or class of objects) and its background can be enhanced. This technique appears limited in its discrimination potential for remote sensing applications when compared to more elaborate multispectral methods. For example, consider a film-filter combination (or a single filtered detector) receiving a total radiant power P_T in its sensitive spectral band, given by: $P_T = P_1 + P_2 + P_3$ where P_1, P_2 and P_3 represent the contributions to the total power contained in each of three smaller subbands within the total measurement band. Film exposed to this radiation will have a density determined by the total power received, P_T. Similarly, the single detector will respond with an output signal proportional to this same total power. For simplicity, assume the power level in each subband (P_1, P_2 and P_3) can have only the values one or zero. Thus P_T (and either the film density or detector output) can take on only four distinct values, namely zero, one, two or three. However, as shown in figure 2-16 there are eight possible conditions (or combinations) for the radiation received in the three subband total. Thus the integration of a single band detection process has caused a loss of 50% of the potential information available in this single case. As the number of bands increases, together with the range of levels in each, the loss becomes even greater.

Color film is perhaps the most familiar advance in an attempt to extract more of the spectral information contained in the available radiation field. In order to better understand the response of various films in a specific application, consider the spectral reflectance curve of figure 2-17, typical of green vegetation. Also shown are the spectral bands of response for four types of film; panchromatic, filtered IR Aero, Ektachrome, and IR Ektachrome. Conventional black and white film observes green vegetation as having an average reflectance of about 10%, and fails to distinguish the unique spectral distribution of the

reflected radiation. The limited number of gray levels available clearly offers limited potential for distinguishing vegetation from other materials.

The conventional color film covers the same spectral interval but in three separate bands as indicated in figure 2-17. Thus more information is available than from any one single band since the relative brightness of a scene in three wavelength bands is preserved, recorded and presented to the viewer.

Another film found to be useful in vegetation detection is IR Ektachrome. Since vegetation is seen to exhibit a very marked reflectance in the infrared region, its image on this "false-color" film is red. This film is considerably better for distinguishing vegetation from other materials than IR Aero Film; e.g. "white" objects reflect equally in the green, red, and infrared thereby remaining white in the photograph while vegetation appears red. Good examples of the use of such false color presentations are presented in reference 4.

These color films merely display spectral reflectance variations as averages in three separate wavelength bands, i.e. they retain individual measures of P_1, P_2 and P_3 in the example of figure 2-16. By using an optical-mechanical scanner these spectral bands need not be limited to the photographic region. Indeed, the information need not be limited to spectral variations constituting a "reflectance signature" since the scanner can include thermal emission bands as well.

Early attempts at multiband sensing with optical-mechanical scanners utilized beam splitting and additional detectors (fig. 2-18) or a combination of beam splitting and dual scanning instrumentation (fig. 2-19). However, as detector fabrication technology advanced it became practical to construct multielement detector arrays in a variety of designs. By selective filtering, each element of such an array can be made to respond to a different wavelength band. If a linear array is oriented so that the image of the element on the object plane lies along a scan line, each element will scan across a given scene point in succession, and each can generate an image of the scene radiance in its particular spectral band.

Alternatively, the individual signals can be stored in electronic form and processed to indicate a variety of interrelations prior to printing a final image. In practice, however, this latter approach suffers from two prime difficulties; (1) proper registration of the signals from many separate detector elements is difficult to attain and (2) the establishment of desired spectral bandwidths by multiple filters is not always possible.

Both these difficulties arise from the fact that such multiband optical-mechanical scanners utilize the individual detector elements (or a field stop in front of or imaged on the detectors) to define the instantaneous field of view in each channel. If however, the entrance slit of a multichannel spectrometer is used as a field stop, each detector in the spectrometer

observes the same resolution element of the scene (instantaneous field of view) but in a different wavelength region (as shown schematically in fig. 2-20). Registration is automatic in both space and time since the same physical stop defines the field of view for all wavelengths. Further, the usual problem of establishing spectral bandwidths by filters does not exist. The bands are established by stops in the dispersed spectrum which permit more exact tailoring than do filters.

The output signal from each detector element is a video signal corresponding to the scene brightness in the particular wavelength region of operation. These multichannel video signals can be tape recorded and analyzed in a number of ways to obtain information about the spectral characteristics of the scene. Also each video signal can be used to generate scene imagery which corresponds to the radiance variations encountered in that particular wavelength region. Figure 2-21 is an example of such multiband imagery. The video signals from the multiple detectors can be sampled to determine the spectral distribution of the radiation from a given scene point; figure 2-22 shows a sampled raw spectrum as obtained with a 12-Channel airborne system, while figure 2-23 shows a plot of reflectance as a function of wavelength as determined by this system, compared to that obtained with laboratory and field spectrometers. Such spectral data can be analyzed by high-speed digital computers to obtain summary statistics of the spectral characteristics of objects and their backgrounds; and to determine how the spectral signature varies with instrumental, operational, and environmental parameters. Figure 2-24 shows a computer printout of the effect of angle on the brightness of a row crop (a potential background) in a given wavelength channel. Finally, the spectral signatures can be used to identify or selectively enhance or suppress the brightness of objects in a scene on the basis of their spectral radiance or thermal properties (e.g. ratio of albedo to emittance). Thus, the multichannel video data can be fed to a signal processor that is designed to generate a single video signal whose intensity, for example, might be proportional to the probability that the spectrum measured is that of a given material or terrain. Figure 2-25 shows a conventional image made from the 0.8-1.0 micron video of the multispectral scanner and an image made by processing two spectral bands to recognize bare soil, rice, and safflower. The latter imagery is truly multispectral as opposed to multiband imagery of figure 2-21.

The discrimination potential of truly simultaneous multispectral sensing is enormous. Assume for example that in a given spectral region a sensor can distinguish 10 gray levels. Such a sensor can only provide 10 unique states or measurements of radiation level. If, by broad band sensing, independent signals are derived in the ultraviolet, visible and infrared then 10^3 different states are possible. Consider now that the 0.3-2.5 micron spectrum alone has been shown to be divisible into at least 30 separate bands, each providing sufficient

signal to noise to permit scanning of the scene and recognition of small differences in reflectance (ref. 5). Of course, in practice, the variation in source illumination and in reflectance of like objects is such that a considerably small number of different "signatures" will be realized. However, signals from detectors having the potential for 10^{30} different states can be processed by computers to determine how they correlate with those expected from known materials.

As an imaging sensor capable of obtaining spectral data the optical-mechanical scanning spectroradiometer has several unique features in addition to the automatic space-time registration and enormous discrimination potential. The detectors are stable, linear and have a dynamic range on the order of 10^6. They can be calibrated periodically by observation of sources of known radiance. This calibration procedure is considerably simpler than for a photocathode detector such as the vidicon, since each resolution element in a camera-type system is essentially a different detector.

Since data are being generated by scanning the scene, the resulting video signal also contains spatial information. The data are amenable to spatial discrimination and processing or the output can take the form of total area coverage of a given spectral response (e.g. acres of wheat).

The large dynamic range mentioned above, together with the linearity of the nonintegrating quantum detectors used in scanners has a further advantage over integrating detectors. The scanner output can readily be adjusted to compensate for a washing out of contrast such as that due to atmospheric scattering. Basically the apparent radiance of a scene (L) is given by:

$$L = \tau L_o + L_s$$

where L_o is the surface radiance, τ the atmospheric transmission and L_s the radiance due to atmospheric scattering. For the condition of uniform scattering within the scene, L_s is a constant, produces a d-c signal, and can be subtracted out of the resulting signal. Similarly, atmospheric attenuation can be compensated by appropriate gain changes. Such techniques make possible the generation of imagery (from space) free from the degradation of blue haze due to Rayleigh scattering.

The parametric equations governing performance, schematic diagrams, and a general discussion of the details of design for a generalized multispectral despersive scanner are presented in appendix C. A wide variety of potential applications in the field of remote sensing are presented in the next section.

3. THE OPTICAL-MECHANICAL SCANNER IN REMOTE SENSING

The remote sensor provides a non-destructive measurement of the natural environment under investigation. This measurement can take on many forms for myriad applications in the solution of problems currently facing agriculturalists, foresters, biologists, botanists, oceanographers, ecologists and the like. In fact there seems to be no practical end to the list of disciplines and their associated problem areas which are amenable to (at least partial) solution through the remote collection of data. The optical-mechanical scanner, with its capability for the collection of both spatial and spectral information is a powerful tool in the attack on these problems.

This section will attempt to present an admittedly incomplete but representative cross section of the potential applications of the optical-mechanical scanner to remote sensing problems; together with important considerations in the selection of proper instrumentation and the conduction of data collection missions.

3.1. THE SENSOR-USER CHAIN*

Since World War II, many techniques for producing imagery at wavelengths outside the photographic region have been developed. Imagery can now be generated over the entire optical range from 0.3 to 14.0 micrometers, thus spanning the ultraviolet-visible, and infrared wavelengths. At microwave wavelengths, passive and active systems have been developed to produce imagery under all weather conditions. While some effort has been spent to develop and use the sensing instrumentation, less effort has been directed toward gaining an understanding of the relationships between what the remote sensor sees, the dynamic parameters of the vegetation, the soil at the surface, and the illuminating conditions available for passive observations. Further more, much of the experimentation has been carried out with sensors and camera systems confined to one or another relatively broad spectral bands, in or near the visible region and often there has been relatively little "ground truth" collected at the time the images were made. In addition, interpretation of the imagery is always subject to the limitations of the photographic process and the skill and training of available interpreters.

At The University of Michigan, a unique passive technique is under development centered around the use of a truly simultaneously multispectral optical sensor. This sensor detects more of the information about an object than has been used previously and permits a new form of interpretation by providing data not available to the human eye with the added potential for

* This section taken from a lecture by F. Polcyn.

automatic recognition of objects from among large numbers of observations on the basis of their spectral properties.

For the user who wants to understand the capability and potential of multispectral sensing, an appreciation for the types of information that this new sensor provides and its problem areas is a needed first step. Consider the sensor-user chain suggested in figure 3-1. Sensor designs can provide outputs from which quantitative measurements can be derived and inferences can be drawn or deductions made to answer specific questions about parameters of interest. Figures 3-2 a, b, and c give specific examples of the types of information that are potentially derivable from a multispectral sensor for one specific example of the use of remote sensing: an investigation of plant productivity. The important point is that one should not expect something of the sensor which it cannot provide. A given set of sensor requirements, e.g., good signal-to-noise ratio, synchronous multichannel registration, voltage references, and absolute radiance or reflectance calibration, will provide a set of sensor outputs which give information about a specific parameter. The user must bring to the problem an a priori knowledge as to the meaning of this parameter in a given environmental context. An example is the generation of a spectrum of a particular vegetation which led to an identification of a plant species which in turn permitted an inference of subsurface soil conditions.

One can conjecture on the capabilities of multispectral sensing techniques for the solution of a great many problems, and suggest some potential applications. Multispectral sensing may be used as a tool for the inventory of ecosystem components (e.g., vegetation, large animals, soils, landforms, and hydrologic features) and for monitoring changes in physical and physiological processes.

Plant species can be identified and mapped by means of their multispectral signatures. To illustrate, discrimination of major crop species in the agricultural regions of Ann Arbor, Michigan; Lafayette, Indiana; Weslaco, Texas; and Davis, California has been achieved by analysis of voltage spectra of the crops and image tonal differences in 18 spectral channels. Recognition of algal mats growing at shallow water depths in the Florida Everglades by means of their unique spectral characteristics is a possibility.

The distribution and abundance of livestock and large wildlife species may be mapped chiefly on the basis of emission due to their higher temperature. Accurate counts of livestock and deer in enclosures have been achieved by means of nighttime thermal infrared sensing.

Soil types derived from glacial sands and gravels, moraines and till plains, and bog, swamp, or river bottom sediments might be differentiated and mapped due to differences in

surface color or heat and water absorption properties. Where the soil surface itself is obscured by vegetation, physiographic and vegetative indicators might be used to infer soil type. Mapping of soil moisture differences after rainfall or irrigation reveals soil permeability differences which may be used as an aid to their description and classification. Studies of physical and chemical properties of different soil types within controlled sites are needed to determine the influence of moisture, inorganic and organic matter, and structural composition on the spectral radiance characteristics of soil types before soil type mapping can be conducted reliably within larger sites or regions. Once signature and decision criteria for the identification of soil types within a site or region are established, multispectral data may be used to monitor changes in composition, moisture, or nutrient levels as a function of time.

Geomorphic features of ecosystems might be identified and mapped by means of their heat and water absorption properties and reflectance characteristics. A capability of differentiating lava flows of different ages at Mono Lake, California has been demonstrated using the ultraviolet portion of the electromagnetic spectrum. For landforms obscured by vegetation, vegetative indicators may be used to differentiate different landform types. For example, tree islands in the Florida Everglades can be differentiated from other landform types because of the spectral contrast of the mangrove species from other plant species and from the surrounding water surfaces. Inferences concerning rates of consolidation of the tree islands may be made by periodic mapping of their distribution.

Hydrological features of ecosystems such as distribution, size, and depth of surface water bodies (where water is transparent) and distribution and size of snowpack can be mapped by means of their unique spectral and thermal properties. The mapping of the distribution of wildfowl breeding sites near Jamestown, North Dakota has been conducted to supply more rapid and accurate data concerning the number and size of open water bodies suitable for breeding.

There are some limitations on the amount of identification that can be performed. For example, for a given sample date and time, some plant species of similar gross geometry, pigmentation, and cellular structure cannot be differentiated by their spectral characteristics alone. In such cases, different sample dates during the growing seasons or different times of day can be selected to take advantage of any temporal differences which might exist between the species.

From species identification several other types of information concerning plant resources might be inferred. First, the presence of the dominant plant species in an area might let one deduce the plant species which make up the remainder of the plant community. Second, if the geographical distributions and seasonal variations of plant communities can be mapped

in areas having strong environmental gradients of light, water or temperature, information concerning floristic composition, structure, phenological status, successional status, and variation in standing crop biomass as a function of time might be derived or inferred from such maps.

An example of the selection of remote sensing techniques and appropriate instrumentation for the detection and measurement of shallow water details is given in appendix D, which is a report prepared at The University of Michigan for NAVOCEANO, under contract No. N62306-67-C-0243.

As a somewhat different example illustrative of the versatility of multispectral data collection in remote sensing its potential in the study of ecosystems is discussed in appendix E. Additional examples are discussed in references 6, 7, and 8.

3.2. INFORMATION IN THE SPECTRAL CHARACTERISTICS OF OBJECTS*

The preceding sections and the discussions contained in the referenced appendices illustrate one of the more promising features of multispectral sensing—i.e., aspects of the phenomena of interest can manifest themselves as changes in the spectral reflectance of vegetation, soil or objects in the scene. These changes come about in a variety of ways. Different plant species tend to have characteristically different spectral reflectance which provides the basis for multispectral sensing in the agricultural crop censusing problem. Figure 3-3 is an example of spectral radiance differences among several crops. These data were obtained from airborne surveys using The University of Michigan multispectral scanner and demonstrate the feasibility of recognition of crop type on a purely spectral basis. Examples of tree leaves with different spectral curves have been given by Gates (see ref. 9). Even the different states of leaf maturity are associated with a changing spectral reflectance (fig. 2, ref. 9). In addition, the top and bottom of the leaf may also show spectral difference. Soil types may have different reflectances but caution should be exercised since in the optical region of the spectrum only the surface is observed and sometimes surface conditions are not truly representative of a practical depth of material. Moisture and organic content and size of particles can all influence the percent reflectance of soils. See, for an example, the results of studies by Bowers and Hanks (ref. 10). Similar spectral differences occur at the long infrared wavelengths where many common materials exhibit emissivity variations as a function of wavelength. Calcite, quartz, and dunite are examples of materials which show characteristic emittance minima between 7 and 14 microns. See result of measurements by Lyon and Hunt (refs. 11 and 12).

* This section taken from a lecture by F. Polcyn.

There are also more subtle ways in which a composite object such as a tree can exhibit a spectral difference such that it produces a tonal contrast with its background. An example is shown in figure 3-4, an ultraviolet image of an apple orchard. The arrow points to two trees which appear lighter in tone. Spectral measurements at the time showed that the bark of the tree has a higher reflection in the UV than the leaf. The two trees which showed the lighter tone were found to contain a smaller leaf area and to have been damaged by mouse girdling. Consequently, the loss in vigor of the tree was detected through the process brought about by the relative ratio of bark to leaf area, while the normal trees with larger leaves had a lower overall tone.

The loss of vigor of plants or infestation of trees by insects may also produce spectral differences in the individual leaves especially in the red region of the spectrum (0.62 to 0.66 microns), and in the near infrared (0.7 to 2.3 microns). Consider figure 3-5, which shows characteristic spectral reflectance curves for green corn and dry brown leaves. Absorption by the pigments (mostly chlorophyll) account for its shape in the ultraviolet and red region, while scattering in the leaf and water absorption account for its shape in the near infrared. Loss of vigor in an individual leaf appears to be associated with a reflectance change between 0.7 and 0.9 microns but may not be detected since one actually observes a stack of leaves which can produce an apparent increase in reflectance due to the addition of reflections from multiple layers (ref. 13). In addition, it is well known that the application of fertilizers produces a darker green color to vegetation and it has also been observed that moisture deficient grass exhibits a lower reflection when viewing forward scattered sunlight whereas, the same area viewed by back scattered sunlight produces no discernible contrast between dry and moist areas. Other examples of angle dependent reflectance properties are illustrated in figures 2-24 and 3-6. To illustrate the importance of angle effects, consider the scan-angle and scan-direction dependencies that have been noted (at about 0.5 μ) in some of the data which were obtained from low altitude aircraft. As shown in figure 2-24, corn shows little scan-angle dependence when it is scanned in a plane perpendicular to the sun's direction (and here coincidentally the row's direction as well). When it is scanned in a plane parallel to the sun's direction and toward the sun, significant changes in radiation levels are observed; an intermediate variation is observed for scans at 45° to the sun's direction. A larger backscatter signal is evident when the scan is away from the sun. The data for corn in figure 2-24 represent the averages of more than 70 scan lines of data for each angle plotted; also, each point of each scan line was adjusted to correct for scan-angle-dependent effects due solely to the instrumentation and aircraft roll. It has been found that the scan-angle dependence in this example, channel 5 is different from those of other channels. For example, it is as much as +60% of the nadir (0°) value in channel 8

(0.58 to 0.62 μ) and as little as +5% to +10% in the infrared channels 11 (0.72 to 0.80 μ) and 12 (0.80 to 1.00 μ).

Variations such as those above can mask differences between various crops unless the geometry of illumination and observation is accounted for. To illustrate this effect and also to show the strong similarities between signals from corn and soybeans in July, figure 2-24, shows a dashed line which represents the signal obtained from a soybean field (that also had east-west rows) on a pass with scans perpendicular to the sun's direction. Similar results were obtained in the other eleven channels. The discrimination between corn and soybeans in July then becomes a function of angles of view for a given sun illumination angle and scan direction.

Other examples between corn and wheat are shown in figure 3-6. These data were obtained from photographic film. Note the better discrimination capability for these two crops using IR film.

Investigations concerned with satellite applications must consider whether or not the spectral differences discussed here can still be observed after the radiation has undergone absorption and scattering along the much longer atmospheric path. The University of Michigan has conducted a preliminary test of the effects of altitude on multispectral data. Results are described in appendix F.

A major usefulness of the multispectral characteristics of objects lies in the potential of automatic recognition and contrast enhancement. These subjects are discussed in detail in other sections. However, it is instructive to examine multispectral imagery and to observe the variety of tonal patterns exhibited in various bands. Two examples, figures 3-7 and 3-8, are given here, together with a short explanation of highlights in each.

3.3. GROUND MEASUREMENTS

In any remote sensing mission the ground measurements to be made before, during and after an airborne or spaceborne overflight are of considerable significance. All too often, however, this important aspect of remote sensing is underemphasized or ignored entirely. Since the reader should expect, beyond a discussion of the basic mechanics of scanners, some comment on important considerations in their use, it seems appropriate to conclude with a general discussion of "ground truth". Similarly, in consideration of scanners from an operational viewpoint, appendix G presents a general discussion of various data processing techniques and a short tabulation of important scanner and mission operational procedures.

3.3.1. GROUND TRUTH.* Ground truth can generally be defined as any information about a target area or specific target which aids in the interpretation, analysis, or evaluation of airborne remote sensor data. Specifically, this information can be divided into seven classes: 1) general classification of area; 2) specific properties of area; 3) general classification of targets; 4) specific properties of individual targets; 5) general environment in area; 6) specific environment of individual targets; and 7) distribution of targets within the area and relative to each other. Items 1), 3), 5) and 7) are often adequately observable in a qualitative manner, while the rest of the items are more likely to require quantitative observation.

Optical-mechanical scanners and radiometers sense directly the apparent spectral radiance of a target. This quantity is not, unfortunately, an inherent target property, but rather is a function of the target's environment as well. In particular, the apparent spectral radiance observed for a given target is determined by the reflectivity or emissivity of the target, the environmental illumination upon the target, the atmospheric path effects between target and sensor, and for thermal wavelengths, the temperature of the target. Since it is usually the purpose of remote sensing to determine the specific type and condition of areas or targets, a means must be available for interpreting the remotely sensed radiance in terms of specific target attributes.

Of the four parameters which determine the apparent spectral radiance of a target, reflectivity or emissivity and temperature are of the most interest since these relate directly to the type and condition of the target as shown in figure 3-9. The arrows in figure 3-9, indicate the direction of the logical process in using remote sensor data, i.e., starting with the remotely sensed radiance and inferring backward through reflectivity/emissivity and temperature to desired target characteristics. The purpose, then for acquiring ground truth is to provide the information necessary to complete this logical process.

The parameters of target type, target conditions, and environment can be defined by specific measurables. Examples of these are shown in figure 3-10. The mechanics of ground truth involves determining some or all of these measurable either quantitatively or qualitatively, depending upon the specific requirements of a given program or investigation.

The sources of ground truth data can be divided into three main categories: 1) accumulated knowledge and experience; 2) laboratory measurements; and 3) field measurements.

Accumulated knowledge and experience is a source of ground truth sometimes overlooked and often underused. Reference books can supply such information as a general

* This section was taken from a lecture by R. Horvath.

description of an area and its weather as often found in an encyclopedia, or a comparison of the physical properties of various materials in a chemistry and physics handbook. Maps and surveys can be of great usefulness in obtaining information about topographic features, river drainage systems, soil distributions, etc. General scientific literature abounds with the results of experiments and investigations which can provide significant data (either directly or by analogy) for the particular problem at hand.

Laboratory measurements and tests are extremely useful sources of ground truth data, especially when very accurate and very detailed information is required. It must be realized, however, that the transferring of a specimen or sample from its natural environment into the laboratory may cause a change in the property to be measured. As a consequence, laboratory measurements should be restricted to inherent characteristics which do not significantly depend upon the specific environment, or whose dependence upon environment can be compensated for. As an example, the mechanical composition of a particular soil can probably be measured better and more comfortably in the laboratory than in the field, while the temperature or moisture content cannot, since the latter are tightly related to environment. On the other hand, environment related parameters can often be measured under a wide variety of controlled conditions in the laboratory, and the results applied to an in situ sample under specific environmental conditions. For example, extensive laboratory measurements of the thermal conductivity of a particular soil type under various moisture and compaction conditions can be used to specify the thermal conductivity of a particular soil in situ if its compaction and moisture content are known.

Field measurements on-site are required for most target characteristics which are environmentally dependent. Such quantities as temperature, vigor, orientation, and of course, meteorological conditions require direct field observation. This is especially true if the measurement must correspond to some particular point in time, such as when there is an airborne remote sensor overflight.

There is a certain amount of ground truth effort necessary prior to any remote sensing mission. In general, one must have enough information available about an area to insure that the flight paths pass over the areas or targets of interest, and so that flight altitudes are commensurate with both resolution requirements and safety. As a minimum this includes maps and a general description of the area. In addition, general weather conditions must be known, not only for aircraft operation, but also to insure that ground instrumentation and personnel can be properly prepared and provisioned for the ground environment.

For many types of ground truth data, it is necessary to perform measurements coincident with the airborne sensor overflight. All those characteristics of the targets and the environment which can change significantly with time fall into this category. For some measurements, especially involving meteorological conditions, it is often desirable to make continuous measurements starting sometime prior to the mission.

There are certain types of data which can be acquired almost anytime, either before, during, or after a mission. Information such as vegetation type, geologic structure, physical dimensions and geometry are often essentially independent of any reasonable time scale and can thus be acquired as is convenient.

The specific remote sensing application is the prime factor which determines the type of ground truth necessary. For instance, the same ground area can be studied for geological, hydrological, or biological reasons, and in each case the required ground truth would be different. A geologist, for example, sees the world as being composed of a wide variety of identifiable and interesting rock and mineral types and formations covered by an innocuous mantle of vegetation of only secondary or inferential interest. A botanist, on the other hand, sees the world as being composed of various vegetation types with subtle and interesting variations and distributions, while perhaps ignoring the geologic aspects involved. Thus, depending upon application, the same area can be "ground-truthed" in entirely different ways, providing entirely different data.

The capabilities of the remote sensor system can, and should, have a strong role in determining the type of ground truth data acquired. A thermal infrared scanner, for instance, has a definable sensitivity expressible as a noise equivalent temperature difference sensitivity (NEΔT). If the NEΔT of a particular system is, say, $1^{o}C$, then it is wasteful of time and money to require a ground truth team to measure target temperatures to an accuracy of $0.01^{o}C$. Another limiting factor for an imaging sensor is its spatial resolution. The coarser the resolution, the less desirable are point sample ground measurements, and the more desirable are larger area-averaged measurements.

Unfortunately, the quantity and quality of ground truth acquired for any program is quite often limited by time and money available for that task. Historically, the emphasis for money expenditure has been directed primarily toward sophistication of aircraft instrumentation, while the data analysis and ground truth programs have received much less emphasis. This unbalance has begun to be appreciated recently as more users from natural science disciplines have become involved in remote sensing.

3.3.2. INSTRUMENTATION FOR GROUND MEASUREMENTS. The instrumentation necessary for successful execution of a particular ground truth effort will depend upon the type of information desired, and the necessary accuracy and quantity of such information. As previously indicated, the determining factors will be application, remote sensor characteristics, time, and money. In general, the application and the physical phenomena producing the remotely sensed radiation will qualitatively define the type and quantity of data necessary. The remote sensor characteristics will define the accuracy with which the measurements need be made. Finally, considerations of time and money will force streamlining of the ground program to maximize as far as possible its cost-effectiveness. Due to these variable factors, it is not appropriate to attempt any explicit description of the particular type and quality of instrumentation desirable. The references at the end of this section provide some information in this respect. References 14 and 15 give a good description of how a target is coupled to its environment, and thus indicate what environmental and target parameters would be important in a given situation. References 16 and 17 describe, in general terms, various devices and techniques available for measuring given environmental and target parameters, indicating both good and bad characteristics of each.

A short listing of the sources for various equipments are given in Table I. The environmental and target parameters listed in Table Ia are followed by numbers referring to the respective manufacturers or distributors listed in Table Ib. Although the list is not at all comprehensive, it provides a good starting point for specific information.

TABLE Ia

Measurement Type	Equipment Source
Environmental Radiation	3, 6, 12, 16, 18
Target Radiation (Spectral)	2, 4, 6, 7
Radiometric Temperature	2, 14, 19
Surface or Bulk Temperature	5, 8, 12, 14, 16, 17, 18, 20
Air Temperature	1, 3, 5, 8, 12, 14, 16, 17, 18, 20
Relative Humidity/Dew Point	1, 5, 9, 12, 14, 16, 17, 18
Barometric Pressure	1, 5, 12, 14, 16, 17
Wind Speed/Direction	3, 11, 12, 14, 16, 17, 18
Evaporation/Transpiration	12, 16
Soil Moisture	12, 13, 14, 16, 17
Rainfall	11, 12, 16, 17
Soil Heat Flux	3, 17, 18
Thermal Properties	17
Chemical and Mechanical Composition	13, 15, 17
Mechanical Properties	17
Vegetation Physiology and Chemistry	17
Water Clarity/Turbidity/Impurities	9, 12, 17
Water Flow	11, 12, 14, 16
Atmospheric Gases/Particles	1, 4, 10, 20
Data System/Recorders	3, 11, 14, 15, 16, 18

TABLE Ib

Equipment Manufacturers

1. Bacharach Instrument Co., 625 Alpha Drive, RIDC Industrial Park, Pittsburgh, Pa. 15238
2. Barnes Engineering Co., 30 Commerce Road, Stamford, Conn.
3. Packard Bell, Meteorological Instruments Div., Newbury Park, Calif.
4. Block Engineering, Inc., 19 Blackstone St., Cambridge, Mass. 02139
5. The Dickson Co., 7420 Woodlawn Ave., Chicago, Ill. 60619
6. E. G. & G., Inc., Products Div., 160 Brookline Ave., Boston, Mass. 02215
7. Exotech, Inc., Instrumentation Div., 12721 Twinbrook Parkway, Rockville, Md. 20852
8. Fenwal Electronics, Inc., 63 Fountain St., Framingham, Mass.
9. The Foxboro Co., Neponset Ave., Foxboro, Mass.
10. Gelman Instrument Co., 600 S. Wagner Rd., Ann Arbor, Michigan 48106
11. W. & L. E. Gurley, Troy, New York
12. Kahl Scientific Instrument Corp., P. O Box 1166, El Cajon, Calif. 92022
13. Kel Instruments, Inc., P. O. Box 446, New Brunswick, N. J.
14. Leeds & Northrup Co., 20820 Greenfield Road, Oak Park, Mich. 48237
15. Photovolt Corp., 1115 Broadway, New York, N. Y. 10010
16. Science Associates, 230 Nassau St., Box 230, Princeton, N. J. 08540
17. TESTlab Corp., 216 N. Clinton St., Chicago, Ill. 60606
18. C. W. Thornthwaite Associates, Laboratory of Climatology, Route 1, Elmer, N. J. 08318
19. Williamson Development Co., 317 Main St., West Concord, Mass.
20. Yellow Springs Instrument Co., Yellow Springs, Ohio 45387

Appendix A
SIGNAL HANDLING, IMAGE STABILIZATION AND CALIBRATION
FOR THE OPTICAL-MECHANICAL SCANNER

L. M. Larsen

Signal handling can be defined as the functional step of translating the irradiated power focused on the detector element(s) by the collecting and dispersing optics, to a recording medium such as magnetic tape or photographic film.

The prime goal of this function is to obtain high fidelity or to minimize the distortion and noise that is added to the signal during this functional step. The general philosophy is to use a detector element that has a high responsivity and a high detectivity for the spectral region and electronic frequency range of interest. This translates the irradiated power on the detector element into a signal voltage that is as high and as noise free as the detector state of the art will allow; however, the signal level and detector impedance are usually such that conventional amplifiers will add too much noise or distortion to the signal. The distortion is caused by both limited frequency response and nonlinearities in the amplification process. Special amplifiers, called preamplifiers, must be used in the initial stages of amplification. The signal out of the preamplifiers is then amplified with conventional postamplifiers. Normalization of the signal amplitude is also accomplished in the postamplifier stages. This normalization is required in order to match the signal variation to the dynamic range of the recorder. If the signal into the recorder is too high, saturation is reached and the resulting distortion is intolerable; on the other hand, if the level is too low, noise in the recording medium tends to overshadow the signals. This condition requires that the gain of the postamplifiers be adjusted to the proper value. The level or DC offset of the signal must also be adjusted to match the recorder. Level and gain adjustments in the postamplifiers are used to normalize the signal. The signal handling equipment must have adequate frequency response in all stages from the detector element to the recorder. This requirement includes low as well as high frequency response. The system must respond fast enough to pass the rapidly varying signals generated as the optics scan the scene. The minimum high cutoff frequency is inversely proportional to the time required to scan one resolution element on the ground. Adequate low frequency response is set by the tolerable droop in the signal waveform.

Signal saturation or clipping in the region of interest is unacceptable at any stage in the signal handling chain, however, clipping in the region slightly outside the region of interest is often desirable. This eliminates overdriving stages, such as the recorder, which follow the clipper.

As can be seen from the previous discussion, signal handling equipment is nothing more than a Hi-Fi transducer followed by an adjustable Hi-Fi amplifier. The transducer converts the irradiated power to a fluctuating voltage and the amplifier increases the signal to an appropriate magnitude for recording. Stabilization of the imagery is the removal of image distortion caused by platform motion, such as roll and sometimes pitch and yaw.

The first and most obvious way this effect can be removed is to mechanically stabilize the scanner so that it doesn't roll, pitch or yaw with the aircraft. Although such stabilization has been effected in some instrument design it is expensive and complicated. Generally the pitch and yaw motion is less objectionable than roll and therefore, only roll effects are corrected. This can be accomplished in at least three ways.

First, if the recorder is mechanically scanned photographic film, only the film needs to be roll stabilized with respect to the scanning shaft. This can be done by stabilizing either the film or the shaft.

A second way is to delay the video signals the proper amount to compensate for aircraft roll. This becomes quite complicated if many video channels are used.

The third way, and the one usually used, is to generate a synchronization trigger that is delay compensated to correct for aircraft roll. The sweep waveform of the image playback system is then initiated by this synchronization trigger.

Stabilization is desirable, but not always necessary, and in many systems no effort is made to stabilize the imagery.

QUANTITATIVE IMAGERY FROM SCANNERS; CALIBRATION

Most airborne line scanners have been developed with the goal of obtaining high contrast imagery for direct human interpretation of spatial pattern. Image classification as well as contrast enhancement in the electronics is desired to aid the interpreter. All humans tend to automatically compensate for the overall brightness and contrast when viewing an image and therefore tend to use only relative brightness and contrast for classification. In practice the overall brightness and contrast has been dropped in the electronics in order to make good pictures for the interpreter. In the electronics the normal approach has been to AC couple the video signal which maintains the average brightness or offset at a certain film density and to adjust the peak to peak amplitude of the video signal to the full dynamic range of the film densities. Using this approach has produced good pictures but discards information that is of value in identifying the target.

If two identical targets surrounded by different backgrounds are to be spectrally compared for enhancement or classification the absolute radiation from each target must be measured.

It is not sufficient to know the radiation relative to its adjacent background radiation. Therefore, enhancement and classification should be done by comparing the absolute and not the relative radiation received from the target and background in the various spectral bands. This absolute signal must be carried through the system.

Some of the common scanner problems which cause errors in quantitative imagery or in more familiar terms, errors in system responsivity, are outlined below:

I. Angular dependent parameters, usually present in scanners which change optical paths for different scan angles are caused by:
   A. changes in aberrations
   B. changes in aperture masking
   C. changes in stray radiation.

II. Temporal parameters causing errors along a line, line to line or day to day:
   A. electronic characteristics
      1. signal saturation
      2. gain (contrast) variations
      3. high and low cutoff frequency
      4. dc offset (brightness) variations or ac coupling without clamping
   B. dust, condensation and other surface degradation of the optical surfaces
   C. detector temperature and bias changes as well as general aging.

III. Nonlinearities in the transfer function from electronics output voltage to cathode ray tube light output or glow tube light output and from the light output to film density.

The first problem doesn't exist if the optical path in the scanner doesn't change as a function of scan angle. Most scanners, however, are built with a changing optical path and experience this problem to some extent. This problem shows up as a hump or ramp on the "A scope" presentation. These problems can sometimes be overcome by compensation masking of the aperture and by reducing stray radiation.

It is general practice that the dynamic range must be linear and wide enough to include the signals from all radiation levels of interest and that extreme signals should however be purposefully clipped or limited and not allowed to enter the saturated or nonlinear range of the electronics. This clipping will then define the linear dynamic range in the data and the nonlinearities which are usually not well controlled will not cause errors in the measured radiance. Beyond the linear range there will be no change in density and this density will be known to be clipped. This clipping also eliminates the usual recovery time after saturation is reached in the electronics.

The gain variations and high cutoff frequency or rise time requirement are well understood, however the low cutoff frequency is not. Basically this cutoff must be low enough to keep the signal droop tolerable (see fig. A-1). This figure shows the original signal above and an AC coupled signal without clamping below. On the left end of the lower waveform 1.6% droop is illustrated for each scan line. Up to this point only responsivity problems along a scan line, which take place in a small fraction of a second, have been discussed.

For longer periods, as for the time it takes to go from line to line or even from day to day other parameter changes must be considered. These parameters, such as optical and detector responsivities, are hard to stabilize directly but can usually be tolerated if the scanner is used as a comparator. If not only the scene radiation, but also a known reference radiation, is viewed during each scan line the responsivity changes over longer periods can be compensated.

Referring again to figure A-1, the lower the cutoff frequency the less droop will be experienced; however, the average signal will eventually be zero unless clamping (DC restoration) is used. This signal should be clamped once each scan line to a fixed DC voltage at the time when a known reference source is viewed. This clamping won't remove the droop during one scan line, however it automatically corrects for the droop once each scan at the clamping point. Reference radiation is therefore needed for responsivity calibration and clamping AC coupled signals.

From a mathematical approach, a straight line can be defined by knowing two points on that line. The optics, detectors and electronics are usually linear and therefore the output of the electronics is a straight line function of the scanner input radiance. If the radiation from two known reference sources is viewed by the scanner once each scan line along with the scene, two points on that straight line are known and the transfer function from the radiance into the scanner to the electronic output is known. In fact, the transfer function all the way to the film density is established except for the nonlinearity from the electronic output to the film density. This function can be determined by periodically inserting a series of different known voltages (grey scale) within the dynamic range of the signal and noting the film density resulting from each voltage. The transfer function from the scanner input radiance to film density is now established.

In the reflected region of the electromagnetic spectrum (to about 3.5μ) both the reflected radiance and source illumination must be known to determine scene reflectance. Scene illumination can be measured at the aircraft if it is assumed to be the same as that received at the ground. This basic reasoning was followed in modifying two AN/AAS-5(XE-2) scanners at The University of Michigan. Figure A2 is an example of imagery in the solar reflected region with

reference sources and a grey scale reference also being presented. The 1st, 3rd and 5th band to the right of the image represent the dark reference (no radiation) while the 2nd, 4th, and 6th are the 0.4 to 2.6 μ reference lamp, the solar illumination reference and the UV reference lamp respectively. The grey scale reference is above the image and represents 16 equally separated voltages at the electronics output.

The modified AN/AAS-5(XE-2) scanners are used to give a combined output of 17 video channels by using four detector assemblies. One detector assembly is installed in each of the two double-ended scanners. A choice of detector configurations is available (see fig. A3) but the usual grouping is as follows:

Scanner	End	Detector Assemblies
1	A	GeHg detector filtered for one 8.0-14.0 μ channel.
1	B	InSb detector with three channels filtered for 1.0-1.4 μ, 2.0-2.6 μ and 4.5-5.5 μ.
2	A	Photomultiplier spectrometer with ten channels between 0.4 and 1.0 μ (ten become twelve when a UV detector is used instead of the InAs detector in end B).
2	B	InAs detector with three channels filtered for 1.0-1.4 μ, 1.5-1.8 μ, and 2.0-2.6 μ.

The basic optical configuration of the scanner is shown in figure A4. The only rotating part of the scanner is the axe blade scanning mirror that rotates about the center line shown in figure A5. This mirror has four flat mirror surfaces at 45° from the rotational axis. The received energy from the ground is focused on the detector by the parabolic mirror as shown in figure A4. The scanner is positioned in the aircraft so that the rotating axe blade mirror provides the scanning perpendicular to the aircraft path while the aircraft motion advances the scan lines along the aircraft path. The scanning is, therefore, similar to an endless television raster. The sync for the sweep circuitry and video gates are stabilized about the roll axis to reduce distortion in the imagery. No corrections are made for aircraft pitch and yaw motion during the scan.

Each radiation detector in scanner #2 (at wavelengths between 0.3 and 2.6 μ) sequentially scans the scene below the aircraft, two quartz-iodine reference lamp sources in the scanner, an opal glass plate in the top of the aircraft which integrates the sky illumination, and finally the dark interior of the scanner. Figure A6(a) illustrates the configuration used for scanner #2.

The radiance from the reference lamp is controlled by adjusting the filament current from a regulated supply. The current is monitored by measuring the voltage across a 0.4 ohm resistor with a digital voltmeter. The radiation from one lamp is filtered for 0.3 to 0.4 μ and is

used to calibrate the UV detector. and the other lamp is filtered to match the solar spectrum from 0.4 to 2.6 μ.

The nominal field of view (FOV) of the scanners is 80 degrees across the flight line. The lamp reference sources and solar illumination reference for scanner #2 are registered during the period of internal scan so that the FOV for this scanner is unaffected by the calibration sources. However, when the externally mounted thermal reference plates are used with scanner #1, its unvignetted external FOV is reduced to 37 degrees. Each detector in scanner #1 (at wavelengths between 1.0 and 13.5 μ) sequentially scan a lamp, a thermal blackbody plate, the scene below the aircraft, and the other plate. Figure A6(b) illustrates the configuration used for scanner #1. Two blackbody plates with controlled and monitored temperatures that fill the field of view are used for calibration in the thermal spectral region (4.0-15.0 μ).

A quartz-iodine lamp is used for calibration in the reflective region (1.0-4.0-μ) of scanner #1. The thermal radiation in the scanner field of view surrounding the lamp filament cannot be considered negligible in this region. The lamp envelope radiation is largely excluded from the FOV by a baffle close to the envelope and the radiation from this baffle is excluded by a second baffle formed by a hole in the blackbody plate. The radiance from the lamp filament is controlled by controlling the current from a regulated supply. The current is monitored and can be related to radiance through a calibration procedure.

Each plate is controlled (automatically heated or cooled as required) to a set temperature. One plate is normally set to a temperature which results in a radiance comparable to the lowest ground radiance of interest and the other to the highest. The plates are grooved and painted dull black (3M black paint) to provide a non-reflective, high-emissivity surface. A uniform temperature is maintained by using a thick layer of copper (1/4 inch thick) for the plate. Heat is conducted to or away from the copper plate by thermoelectric modules coupled to a liquid cooled heat sink. The liquid is cooled by a radiator in the aircraft air stream. The temperature of the plate is maintained constant by a closed loop servo using a thermistor embedded in the plate as a sensor. The maximum heating or cooling rate on the blackbody plate is greater than $0.2^{o}C$/second. Measurements show the plate temperature to be uniform to $0.1^{o}C$ over the surface and controllable to $\pm 0.1^{o}C$ over the range of $0^{o}C$ to $60^{o}C$. If, however, the plate is cooled below the dew point temperature, either water or frost forms on the surface which results in an error in the apparent radiance and it therefore should not be used under these conditions.

As an example of results obtained from a calibrated scanner the spectral reflectances of the three color panels shown in figure A7 have been measured from the aircraft with the twelve

channel spectrometer. This was done by comparing the grey panels with the color panels in each channel. The resulting spectral reflectance for the light green panel is shown in figure A8. Further verification of the calibration in both the reflective and thermal calibration is being done at the present time. More data on this system is documented in reference 18.

Appendix B
THE THERMAL MODEL

R. Horvath and D. Bornemeier

The Heat Diffusion Equation

The model is based on the assumption that for a large class of targets the most significant heat fluxes are vertical, and that transverse heat flow in the object is negligible. This is surely the case for targets which are uniformly illuminated, whose observable surface normal is essentially vertical, and whose spatial extent is at least comparable to several resolution elements (truck tops, building roofs, road way surfaces, etc.). Furthermore, the target's properties are considered to be homogeneous in its transverse dimensions. Under these assumptions the temperature T(z, t) is given by the solution of the heat diffusion equation

$$\frac{\partial^2 T(z, t)}{\partial^2 z} = \frac{\rho(z)c(z)}{k(z)} \frac{\partial T(z, t)}{\partial t}$$

subject to the boundary conditions

$$\sum_{i=1}^{n} q_{oi} = 0 \text{ at } z = 0$$

and

$$\sum_{i=1}^{m} q_{di} = 0 \text{ at } z = d$$

The observable surface is z = 0, the lower surface is z = d, and $\rho(z)$, c(z), and k(z) are the density, specific heat capacity, and conductivity respectively. The q_i represent heat fluxes at the surfaces of the slab target.

The realistic accuracy of this model depends to some extent on the degree to which the thermal characteristics can be approximated by constant values and the physical configuration by an infinite planar slab. More importantly, however, it depends on the quantitative accuracy of the approximations q_i which represent the actual time-dependent heat transfer processes which take place at the surface exposed to the natural meteorological driving functions.

Boundary Conditions

1. Upper Boundary (z = 0)

For non-vegetated surfaces there are six essential heat transfer processes which must be taken into account. These are (a) body conduction, (b) solar absorption,

(c) net thermal radiative transfer, (d) convection, (e) rain, and (f) evaporation. An additional transpiration term must be considered if the surface sustains living vegetation. The detailed derivation of fluxes appropriate for (c), (d), (e) and (f) have been given elsewhere [1] and here we will describe only briefly their individual dependences on the measurable meteorological parameters (ambient air temperature, horizontal wind velocity, relative humidity, cloud cover, cloud type, etc.) which are in general time-dependent input data for the model.

a. Body conduction. This term accounts for the conduction of heat away from the upper surface into the interior of the target. It depends simply on the thermal conductivity of the surface material and the temperature gradient evaluated at the surface.

b. Solar absorption. In order to allow for the many different illumination conditions which are possible in a diurnal cycle this term is an arbitrarily specifiable tabular form which is just the total direct and diffuse solar irradiance multiplied by the total solar absorptivity of the surface.

c. Net thermal radiation. Net thermal radiation is specified by the difference between the instantaneous total greybody radiation emitted by the surface and that absorbed by the surface from a radiating atmosphere whose mean total emissivity is an analytical function of the time varying relative humidity, cloud cover and cloud type.

d. Convection. The relative contribution of the convection process can vary over several orders of magnitude from a very small rate (molecular conduction) for stable no-wind conditions to a very large rate (forced convection) for unstable atmospheric conditions and large wind velocity. The analytical expression which we have used in the model was deduced by the present authors using theoretical and empirical formulas and experimental data available in the literature. A complete discussion of the origin of this expression and the determination of the coefficients in it is given in reference 19. The convection term is written in the form of a heat transfer coefficient multiplied by the temperature difference between the air at 160 cm and the surface temperature. The transfer coefficient is a function of the horizontal wind velocity at 160 cm above the surface, a surface roughness parameter, a coefficient which depends on the above mentioned temperature difference, and a so-called atmospheric stability coefficient. Basically this convection term is capable of describing quantitatively the heat transfer which can take place under the various conditions of stable, neutral, and unstable

temperature and wind profile combinations which encompass the conditions from laminar flow to turbulent mixing.

e. Rain. The intensity (depth/time) and temperature of rain falling on the target and background are specifiable as arbitrary (tabular form) function of time. The immediate thermal effect is accounted for by a temperature equilibriating heat exchange between the water and the surface.

f. Rain evaporation. A certain portion of the rainfall (that remaining after runoff) is susceptible to evaporation. In computing the rate of heat transfer due to evaporation, a form of Dalton's formula is used which accounts for the effects of wind turbulent mixing, and diffusion in the immediate atmosphere. This term depends on the wind speed, the relative humidity and temperature of the air, and the surface (water) temperature.

2. Internal Material Configuration for the Model

The model allows the target or background to be described by consecutive layers (maximum of 6) of arbitrary thermal properties and arbitrary thickness. Hence, to configure the target or background the density, specific heat capacity and conductivity, (constants) of each layer must be specified. In addition, the total solar absorptivity, thermal emissivity, and surface roughness of the top layer must be specified.

3. Bottom Boundary Conditions

The boundary conditions to be imposed on the lower side of the target will depend mainly on what type of target is being modeled, whereas essentially the same form of boundary conditions at the upper surface apply to almost all targets. For example, if one is modeling a roadway or runway the convenient condition would be to choose a constant temperature on the bottom of the target which is chosen to be, say, six feet in the ground. The road could be structured as follows; 12 inches of concrete, over 6 inches of sand, over 6 inches of crushed rock, over 4 feet of moist clay. In the case of a motor vehicle, the target (e.g., engine hood area) could be structured in the following way; one thirty-second of an inch of paint, over one eighth inch of steel, over one half inch of insulation. The bottom boundary condition in this case would be a heat flux specified by the net radiative transfer of energy between the engine and the bottom surface of the insulation. This in turn will depend on the assumed time-dependent temperature of the engine and the effective mean total emissivity of the engine and insulation surfaces. A term representing the cooling effect of the air stream from the motor fan on the insulation would also be necessary.

The bottom boundary condition, however, may be essentially meteorological in nature. If, for example, a bridge is being modeled, the under side of the bridge would be subjected to convection by a time dependent air temperature and wind velocity, and also a radiative exchange with the water in the river. Some solar wavelength energy would also impinge on this lower side.

As can be inferred from this discussion, the flexibility of this model for handling various targets is largely dependent on the user's ability to describe accurately the heat transfer processes at the lower boundary. The current computer program is capable of handling an arbitrary temperature or heat flux boundary condition on the bottom boundary.

Initial Conditions

The solution of the partial differential equation requires that the spatial temperature distribution be specified at sometime $t - t_o$. This distribution is the cumulative result of the effect of the thermal environments which prevailed prior to t_o. The response of the target to the environments occurring after t_o will be influenced by the spatial temperature distribution at t_o. However, this dependence decreases as time increases and finally becomes negligible after a time comparable to the "time constant" of the system. The model (computer program) uses as input data the time-dependent solar insolation, horizontal wind velocity, ambient air temperature, relative humidity, percent cloud cover, and cloud type, for, say, six diurnal cycles. Five of these cycles are used to obtain the initial temperature distribution for the start of the sixth cycle. Thus the response of the target to the input environment on the sixth cycle accurately reflects the influence of the past thermal history. Experience with the model has shown that 4-5 cycles is sufficient to establish the thermal history effects on most targets.

Appendix C
THE GENERALIZED MULTISPECTRAL DISPERSIVE SCANNER

J. Braithwaite

The performance of a generalized multispectral dispersive scanner (fig. C1) may be described in terms of selected system parameters through the set of parametric equations developed in section C.2. Because of the relations which exist among various parameters, the equations are not the only ones which may be used. However, these particular parametric equations use the sets of parameters which are judged to be the most useful in the sense that any one parameter in any equation can be varied, at any rate, over some range of feasible values without necessitating changes in any of the other members of the selected set.

The S/N equations developed in the following section may be used to calculate the sensitivity of a proposed system. This can be done either in terms of S/N for a given L_λ or ΔL_λ or by finding the ΔL_λ which will give the S/N needed to allow a required detection probability. Then if the performance turns out to be inadequate it is possible by inspection of the equations to find changes in the original parameters which will provide the necessary increase in performance.

For easy reference, the most important of the relations are given below, and are followed by an explanation of the symbols.

$$\dot{\alpha} = \frac{2\pi(V/H)}{n\beta} \tag{C1}$$

$$\tau = \frac{n\beta^2}{2\pi(V/H)} \tag{C2}$$

$$\delta\lambda = \frac{D_1\beta}{D_2(d\theta/d\lambda)} \tag{C3}$$

$$d_2 = \beta D_1 F_2 \tag{C4}$$

$$d'' = m\beta D_1 F_d \tag{C5}$$

$$\ell = M\beta D_1 F_2 \tag{C6}$$

$$\Delta f = \frac{\pi(V/H)}{\beta^2 n} \tag{C7}$$

$$\left(\frac{S}{N}\right)_{I\lambda} = \frac{L_\lambda \delta\lambda'}{4}\sqrt{\frac{\pi n}{m^p}}\cdot\frac{D_1 D^*\beta^2}{F_d\sqrt{V/H}}\ (\text{o.e.} \times \text{s.e.}) \tag{C8}$$

$$\left(\frac{S}{N}\right)_{S\lambda} = 7.4 \times 10^8 D_1 \beta^2 \Delta\rho(\lambda) \sqrt{\frac{nR_c E_\lambda \delta\lambda(o.e.)}{\pi\rho(V/H)}} \cdot (s.e.) \qquad (C9)$$

D_1 = scanner telescope aperture

D_2 = aperture of dispersing element

D^* = specific detectivity of solid-state detectors

d'' = width of one detector

d_2 = width of exit slit

$d\theta/d\lambda$ = angular dispersion of dispersing element

F_2 = f-number of spectrograph telescope

F_d = f-number of beam at the detector

H = vehicle altitude

E_λ = irradiance at target

M = number of spectral channels

$m = \delta\lambda'/\delta\lambda$

L_λ = target radiance

n = number of faces of scan mirror assembly

o.e. = optical efficiency

p = 2 if field lenses are used but 1 if they are not

R_c = cathode radiant sensitivity of photoemissive detector

$(S/N)_{I\lambda}$ = spectral signal-to-noise ratio when noise level is independent of signal level

$(S/N)_{S\lambda}$ = spectral signal-to-noise when signal-dependent shot noise predominates

s.e. = signal efficiency

V = vehicle velocity

α = instantaneous scan angle relative to nadir

β = instantaneous angular field of view of scanner

Δf = electronic noise bandwidth

$\Delta\rho(\lambda)$ = reflectance variation

$\delta\lambda$ = instrument-limited resolved wavelength interval

$\delta\lambda'$ = resolved wavelength interval used

λ = wavelength

$\rho(\lambda)$ = target spectral reflectance (generally a function of angle of view and angle of illumination)

τ = dwell time

For instance, suppose that our first selection of parameters for use in equation C8 led to an S/N of 0.5, whereas 5 is needed to give adequate detection probability. Further suppose we have used the best practical values for n, m, D*, and F, and operational considerations fix V/H and ΔL_λ. Then it is clear that the product of the remaining parameters $\delta\lambda D_1\beta^2$ must be increased by a factor of 10 or the experiment abandoned as impracticable.

It must be noticed, however, that, in any selection of the parameters occurring in the S/N equations, values for other parameters such as the detector and prism or grating sizes are implied. Thus at all times care must be taken to ensure that impractical values of parameters are neither used nor implied in the S/N equations. The remainder of this section deals with these limitations. The symbols discussed are defined above.

n. The effects of the dead time (during which parts of more than one mirror face cover the entrance aperture) will cause the size of the overall scan mirror assembly to increase faster than n. Also, as S/N varies with D but only with the square root of n, it may be more practical to increase D. Alternatively, it may be more practical to obtain an effective increase in n by using multiple synchronized units, of which the double-ended scanners are examples.

m. This is at best a poor way of increasing sensitivity and is best thought of as a factor to introduce the effects of increasing the effective bandwidth given by a particular spectrograph at various wavelengths.

D_1. Notice that, if D_1 is changed but not F_1, the dispersion of the prism or grating or any of the other parameters, then the whole instrument will scale in size accordingly.

F. Theoretically the minimum possible value of F is 1/2, and the smallest practical value will be somewhat greater, say between 0.7 and 1.0. However, it should be noticed that for radiation-noise-limited detectors in which the f-number is increased by means of a cooled shield, D* will increase to offset the change in F.

d". Semiconductor detectors can be made in sizes ranging from about 5 mm down to about 0.05 mm. The smaller sizes are obtained by the use of cooled apertures in front of larger detector flakes which, as a result, may have higher impedances than can be properly matched to the amplifier chain, resulting in lowered effective D^*. The dark currents of photomultipliers are negligible for our purposes, so that the photocathode area does not appear in the equations and very small photocathode areas are not required. The largest photocathode areas are about 1 sq. in., which is adequate for our purposes.

$\dot{\alpha}$. The rotational rate of the scan mirror must not exceed appropriate engineering limits. However, because for both earth and lunar orbits V/H ~ 0.025 and β is of the order of 10^{-3} rad, a typical value for $\dot{\alpha}$ is 1500 rpm, which is quite practical.

τ. The dwell time should not be less than the detector and associated electronic time constant, or the effective value of D^* will fall off. Again with $\beta^2 \sim 10^{-6}$ and V/H ~ 0.025, we find $\tau \sim 6$ μsec. Thus most detectors likely to be used have adequate time constants provided they are used correctly. Exceptions are some of the film-type detectors such as those made with lead sulfide.

$\delta\lambda$ and D_2. Equation C3 relates these to the prism or grating dispersion and the scanner parameters D and β. This relation can be thought of as a way to determine the size of dispersing element (D_2) needed if a particular prism material or grating spacing is used. It quickly becomes apparent that prisms are not practical for long wavelengths or large β.

Detector array length is given by equation C6. This is perhaps one of the most important equations because, if the array is more than a few inches long, the optical system and, for cooled detectors, the cryogenic system, become very large and complicated. Thus in practice M, the number of spectral channels, is likely to be limited. For example, for practical values of β and D and for $F_2 = 2.5$, we find an array length on the order of M/4 cm. Thus even for 20 optimum spectral intervals the array length is 5 cm, or 2 in. Note that M, the number of spectral intervals, and not the spectral resolution or the spectral range individually is limited. However, these too will have restrictions imposed by other practical limitations such as the limited wavelength range over which any grating or prism material can be used efficiently.

The limitations of dispersing mediums are illustrated in figure C2. On this figure, the spectral intervals predicted by use of equation C3 for a variety of prism materials and of grating line spacings are given as a function of wavelength. It must be remembered that this example refers directly to a system with the following specifications only:

(1) Angular field of view, β: 0.0067 rad
(2) Collector diameter, D_1: 15 cm
(3) Prism or grating aperture, D_2: 5 cm
(4) Prism apex angle = 60^o

However, it is a simple matter to obtain corresponding results for other systems by use of the relation

$$\delta\lambda \propto D_1\beta/D_2$$

to change the vertical scale as appropriate. In the example chosen it will be seen that the resolving power, $R = d\lambda/\lambda$ is less than unity over wide ranges of wavelengths for all the prism materials except dense flint. However, if the field angle β can be reduced to 1 mrad and a 10-cm prism can be used, then it will be seen that resolving powers of 10 are readily obtained.

It will also be seen that gratings give uniform wavelength resolution, (in practice this is only approximately true because of angular effects) whereas that of prisms varies considerably with wavelength. Resolving power can always be reduced at a given wavelength by the use of larger detectors; but this degrades the sensitivity of the system.

It should also be remembered that the efficiency of a currently blazed grating falls off rapidly outside about a 1-octave wavelength range, though with careful design and the acceptance of some losses at the extreme wavelengths a range of 1 to 3 in wavelength can be covered.

A sketch of a 12-channel spectrometer designed to replace the original single detector in an existing scanner is shown in figure C3. The spectrometer entrance slit is placed in the position occupied by the original detector. The two lenses and the prism used are "off-the-shelf" items. The lenses are, in fact, 35 mm camera lenses and the prism is a commercial monochromator prism. The spectrum is sliced and relayed to 12 photomultipliers by means of a fiber optics image slicer. Figure C4 is a photograph of this spectrometer with the cover removed. Figure C5 shows the fiber optics image slicer.

The use of these components limits the spectral coverage to 0.4 to 1.0μ. Figure C6 is a sketch for the design of a scanner-spectrograph which would cover the range from .33 to 14μ.

C.2. DERIVATION OF THE PARAMETRIC EQUATIONS

Equations are developed below which describe the performance of a generalized multispectral dispersive scanner in terms of selected system parameters. Because of the relations which exist among many sets of parameters there is no unique way of doing this. However, the final equations are given in terms of a set of independent parameters which are judged to be the most useful in the sense that any one can be varied, at any rate, over some range of feasible values without necessitating changes in any of the other members of the selected set.

Consider replacing the single detector in the focal plane of the scanner telescope by a spectrograph whose entrance slit is in the telescope focal plane as shown schematically in figure C7. The D's are the diameters and the F's the focal ratios of the various optical elements. We assume that D_2, D_2', and the aperture of the dispersing element are all equal (in practice the smallest of these elements will act as the effective stop for all of them).

The field of view of the scanner is fixed by the spectrograph entrance aperture d_1, such that if the aperture is square and defines the field of a solid angle $\Omega = \beta^2$ steradians, then

$$\beta = \frac{d_1}{F_1 D_1} \tag{C10}$$

If the angular dispersion of the prism or grating is $d\theta/d\lambda$, then to obtain a resolved wavelength interval of $\delta\lambda$ in the most efficient way the entrance and exit slits must both subtend $\frac{d\theta}{d\lambda}\delta\lambda$. Since also the aperture ratio of the scanner telescope and the entrance optics of the spectrograph must be matched,

$$\frac{d_1}{F_1 D_2} = \frac{d\theta}{d\lambda}\delta\lambda$$

or, substituting for d_1/F_1 in equation C10

$$\frac{\beta D_1}{D_2} = \frac{d\theta}{d\lambda}\delta\lambda$$

Thus,

$$\delta\lambda = \frac{D_1 \beta}{D_2 (d\theta/d\lambda)} \tag{C11}$$

Next, because the individual detectors of the array act as exit slits, the detector size, d_2, is given by

$$\frac{d_2}{D_2 F_2} = \frac{d\theta}{d\lambda}\delta\lambda = \frac{\beta D_1}{D_2}$$

or

$$d_2 = \beta D_1 F_2 \tag{C12}$$

The detector can, of course, be made larger to increase the effective $\delta\lambda$, but this would, in general, be less efficient than increasing $\frac{d\theta}{d\lambda}$ by altering the prism apex angle or grating spacing as appropriate, if this is feasible. However, it may be desirable to vary the bandwidth across the spectrum by using exit slits (detector widths) wider than optimum by a factor m, which is at the disposal of the designer and may well vary across the array. Thus the new waveband is m times wider than that given by C11, or

$$\delta\lambda' = \frac{mD_1\beta}{D_2(d\theta/d\lambda)} \tag{C13}$$

and

$$d' = m\beta D_1 F_2 \tag{C14}$$

where d' = the width of the enlarged detector. The detector height will, however, still be given by equation C12. Further, if the detector array is to cover a wavelength range $\Delta\lambda$, then ℓ, the total length of the array, must be equal to the length of a detector as defined by equation C12 multiplied by M, the number of resolved wavelength intervals (as defined in equation C11) contained in $\Delta\lambda$. (Note: $\delta\lambda$ will vary, because $d\theta/d\lambda$ varies with λ; therefore it will not, in general, be correct to put $m = \Delta\lambda/\delta\lambda$.) Thus

$$\begin{aligned} \ell &= Md_2 \\ &= M\beta D_1 F_2 \end{aligned} \tag{C15}$$

We can regard equation C11 as the basis for the design of the spectrograph, using it to choose the size of the prism or grating, D_2, in terms of other parameters. However, because $d\theta/d\lambda$ is, in general, a function of wavelength, some compromise will have to be made in the choice of the wavelength used as a basis for the choice of D_2. Thereafter, the wavelength resolution and sensitivity can be controlled to some extent by increasing individual detector sizes as explained. Equation C13 then gives the wavelength resolution in terms of the other parameters.

Performance Calculations

Two situations must be distinguished and treated separately. First, we must consider the case in which the noise level is independent of the signal level. The various solid-state infrared detectors will fall in this class. Even those which are radiation noise limited can be treated as belonging to this class because generally the signals can be thought of as small changes in a much larger and essentially constant radiation level. Thus the noise is essentially independent of the signal level and the detector can be characterized by a specific detectivity, D^*. The photomultipliers, however, are in a second class. They are shot noise limited, and this noise is proportional to the square root of the cathode photocurrent, which is itself proportional to the incoming radiation level which may vary considerably at the wavelengths at which such detectors are used.

Noise Independent of Signal Level

If the detector noise is independent of signal level, it is desirable to make the f-number of the cone of radiation falling onto the detectors as small as possible to optimize the NEP (noise equivalent power). To reduce the aperture ratio below that practical in the design of the spectrograph, an array of square or rectangular field lenses could be used in the focal plane of the spectrograph in place of the detector array. Individual detectors would then be placed in the focal planes of each field lens. All the preceding equations still apply except that d and d' will now represent the sizes of the field lenses rather than of the detectors and ℓ is the length of the field lens array. It can easily be shown that the optimum size for each detector d'' is then given by replacing d' by d'' and F_2 by the f-number of the individual field lens F_f in equation C14, which then becomes

$$d'' = m\beta D_1 F_f \tag{C16}$$

The detectors now have the effective area of the prism or grating imaged onto their surfaces; therefore they will in general be squares with sides given by d'' of equation C16. The field lenses, however, have heights given by equation C12 and widths by equation C14 and so will only be square if n = 1.

The sensitivity of a given detector channel can be computed in the usual way, by writing down the S/N as the ratio of the power falling onto the detector and the NEP of the detector.

Assuming a uniform field of view with spectral radiance L_λ the power reaching a given detector is

$$P_\lambda = L_\lambda \delta\lambda' \times \frac{\pi D_1^2}{4} \times \beta^2 \times (\text{o.e.}) \tag{C17}$$

where $\delta\lambda'$ is the resolved wavelength defined by equation C13. We also use the standard formula

$$(\text{NEP})_\lambda = \frac{(\text{Detector area} \times \text{Electronic bandwidth})^{1/2}}{D^*_\lambda \times (\text{s.e.})} \tag{C18}$$

Here "signal efficiency" takes into account the extent to which the amplifier noise factor and the signal shaping due to the finite amplifier bandwidth degrades the S/N of the system.

We can use equation C14 or C16, as appropriate, to obtain the detector area in terms of system parameters. It will be seen that we can combine both cases by putting

$$\text{Detector area} = m^p \beta^2 D_1^2 F_d^2 \tag{C19}$$

where p = 1 if field lenses are not used, p = 2 if field lenses are used, and F_d is the f-number of the beam at the detector, that is, of the field lens if a field lens is used or of the spectrograph output beam if a field lens is not used.

If the electronic bandpass of the amplifier is chosen to optimize the probability of detecting a point target, then the bandwidth will be given by

$$\Delta f \approx 1/2\tau$$

or if we use the simple line scanner equation $\tau = \frac{n\beta^2}{2\pi(V/H)}$:

$$\Delta f = \pi(V/H)/\beta^2 n \tag{C20}$$

Thus we can write

$$(NEP)_\lambda = \frac{D_1 F\sqrt{\pi m^p (V/H)}}{D^*_\lambda \sqrt{n}(s.e.)}$$

Thus from C17 we finally obtain

$$\left(\frac{S}{N}\right)_{I\lambda} = L_\lambda \delta\lambda' \frac{1}{4}\sqrt{\frac{\pi n}{m^p}} \cdot \frac{D_1 D^* \beta^2}{F_d \sqrt{V/H}} (o.e. \times s.e.) \tag{C21}$$

Because $\delta\lambda'$ is related to m by equation C13, it is sometimes more convenient to rewrite equation C21 in terms of $\delta\lambda$, defined in equation C11 which is independent of m. Then we have

$$\left(\frac{S}{N}\right)_\lambda = \frac{1}{4} L_\lambda \delta\lambda \sqrt{\frac{\pi n}{m^{p-2}}} \cdot \frac{D_1 D^* \beta^2}{F_d \sqrt{V/H}} (o.e. \times s.e.) \tag{C22}$$

However, if this form is used, it must be remembered that the spectral interval actually resolved is not $\delta\lambda$ but m $d\lambda$.

Though the detector size and electronics bandwidth have disappeared from these relations, it must always be remembered that equations C19 and C20 are implicit, so that values of parameters which lead to impracticable detector sizes and bandwidths given by equations C19 and C20 must be avoided.

It should also be remembered that effects of path transmission may have to be taken into account, either by including this factor in the overall efficiency or by regarding L_λ as an "apparent radiance." In the signal-to-noise-ratio equations, C21 and C22, the whole of the radiant power falling on each detector is regarded as being "signal." However, more often the signal

in which we shall be interested will be a differential due to a small change in the temperature, emissivity, or reflectivity from one elementary field (resolution element) to the next. Equations C21 and C22 can readily be put into differential form by replacing L_λ by ΔL, $\Delta\epsilon L_\lambda$, $\int_{T_1}^{T_2} \frac{\partial L_\lambda}{\partial T}\, dT$, or whatever term is appropriate.

SHOT-NOISE-LIMITED DETECTORS

For a shot-noise-limited detector, such as a photomultiplier, the equivalent noise current at the photocathode is given by the well-known formula

$$I_N = k\sqrt{2ei\Delta f} \tag{C23}$$

where e = the electronic charge
i = the total photocathode current
Δf = the noise bandwidth
k = the noise factor of the multiplication process as in a photomultiplier tube

The corresponding signal current is

$$I_S = P_\lambda R_c \times (\text{s.e.})$$

where P_λ = the effective spectral radiant power impinging on the photocathode, in watts, and R_c = the photocathode radiant sensitivity, in amperes per watt. Thus

$$\frac{S}{N} = \frac{P_\lambda R_c}{k\sqrt{2ei\Delta f}} \times (\text{s.e.}) \tag{C24}$$

Now

$$i = I_S + I_D$$

where I_D = the dark current; but in practice except at very low light levels we can write $I_D \ll I_S$, so that $i = I_S = P_\lambda R_c$ and

$$\left(\frac{S}{N}\right)_\lambda = \frac{1}{k}\sqrt{\frac{P_\lambda R_c}{2e\Delta f}} \times (\text{s.e.})$$

Using equation C17 for P_λ and equation C20 for Δf, we obtain

$$\frac{S}{N} = \frac{\beta D_1}{2k}\sqrt{\frac{nR_c L_\lambda \delta\lambda(o.e.)}{2e(V/H)}} \times (s.e.)$$

$$= 7.4\beta \times 10^8 D_1 \sqrt{nR_c L_\lambda \delta\lambda(o.e.)/(V/H)} \times (s.e.) \qquad (C25)$$

putting k = 1.2 and e = 1.6×10^{-19} coulomb.

As before, path transmission should be included in the optical efficiency, or alternatively L_λ should be regarded as the apparent radiance as measured at the instrument.

DIFFERENTIAL SIGNALS WITH SHOT-NOISE-LIMITED DETECTORS

At the wavelengths at which shot-noise-limited detectors can be used in orbital surveying experiments, the sun is the only significant illuminating source. Thus if E_λ is the solar spectral irradiance at the target we can put

$$L_\lambda = \frac{1}{\pi}\rho(\lambda)\cos\alpha$$

where $\rho(\lambda)$ is the appropriate reflection coefficient for the directions of illumination and observation, both having two components: azimuth and elevation. Now because E_λ and the angle of illumination will be constant over a period of several scans, we can put

$$L_\lambda = \frac{1}{\pi}\rho(\lambda)E_\lambda$$

and

$$\Delta L_\lambda = \frac{1}{\pi}\Delta\rho(\lambda)E_\lambda$$

Substituting in equation C17 we find

$$P_\lambda = \frac{1}{\pi}\rho(\lambda)E_\lambda \frac{\pi}{4}D_1^2\beta^2\delta\lambda$$

and

$$\Delta P_\lambda = \frac{1}{\pi}\Delta\rho(\lambda)E_\lambda \frac{\pi}{4}D_1^2\beta^2\delta\lambda$$

so that

$$i = P_\lambda R_c$$

$$= \rho(\lambda)E_\lambda R_c \frac{1}{4}D_1^2\beta^2\delta\lambda$$

Then substituting in equation C24, in which we first replace P_λ by ΔP_λ, and using equation C20 for Δf, we obtain

$$\left(\frac{S}{N}\right)_{S\lambda} = 7.4 \times 10^8 D_1 \beta^2 \Delta\rho(\lambda) \sqrt{\frac{nR_c E_\lambda \delta\lambda(o.e.)}{\pi\rho(\lambda)(V/H)}} \times (s.e.) \qquad (C26)$$

Appendix D
REMOTE SENSING TECHNIQUES FOR THE DETECTION OF DOUBTFUL SHOALS

F. Polcyn and V. Noble

The need for updating navigation charts to remove doubtful hydrographic data was expressed as a major concern at the Fourth Session of the International Oceanographic Committee meeting in 1965. The use of satellites with remote sensors is being considered under this contract* as a methodology which may aid in the location of shallow waters or other obstructions to navigation over large areas in a relatively short time.

To give an indication of the magnitude of the problem, consider the number and the distribution of doubtful soundings for the North Atlantic Ocean as reported in Special Publication 20 of the International Hydrographic Bureau. Figure D1 is adapted from that report. Note that in Marsden square 043 there are 49 listed doubtful soundings and ten other soundings for which doubt has been removed.

Many doubtful soundings have been due to errors brought about in the use of the lead line in collecting the data or by the use of echo sounders by small fishing boats. The errors with the lead line usually are one of placing the bottom too deep because of the curving of the line while errors with the echo sounder are usually too shallow because of soundings from intermediate scattering layers and from the fact that small boat operators do not use their sounders continuously so that the proper range scale can be in doubt (ref 20). Many times the error is in location and is due to the inexactness of the knowledge of the ships' position. If satellite technology is to be of any help, accurate knowledge of the position of any shoals detected must be obtained from space. This places a burden on the accuracies of knowing the satellite position and attitude in orbit.

In order for remote sensors to be employed in the detection and measurements of shallow depth, certain consistent observables must be present and be detectable with reasonable systems (in cost and size). Figure D2 shows the relationship of several factors of importance to the detection of a particular water depth.

The four observables of color, thermal change, wave refraction and direct measurement of a reflected laser signal from the bottom are the parameters on which a remote sensor will operate. As in all detection problems, the final classification of a particular depth will depend on

---

*NAVOCEANO Contract No. N62306-C-0243, The University of Michigan, Willow Run Laboratories.

a "convergence of evidence" from several sensors because each one by itself may not be 100% reliable; false detections will usually be encountered from time to time.

The reliability of color anomalies is influenced as an indicator by the spectral quality of the sunlight, by suspended material, by bottom material type, and by benthonic species present. However, it is expected that a shallow area will not only have a greener color associated with it but also under certain conditions produce a water wave refraction. In addition, if feasible laser systems can be built, the direct measurement of water depth by laser ranging methods is a possibility. The thermal anomaly observable tends to hold the least promise because of its non-unique relationship to a particular water depth. It is well known that thermal differences can be observed at the surface, but it is difficult to reliably interpret the differences as related to shallow water and not some other factor.

Figure D3 summarizes the types of sensors that may be employed from space or aircraft to detect remotely the effects associated with shallow water. Consideration of cost, size, ease of processing, all weather capability, etc. must be considered in the final selection for operational use. As an example of the type of data that can be obtained showing bottom details, consider the imagery shown in Figure D4. The area shown is off the Florida Keys near Carysfort Reef and Pacific Reef. Data was taken with The University of Michigan 18 channel multispectral optical-mechanical scanner. Data in the visible region where water transmits is obtained simultaneously in 12 different regions between 0.4 and 1 micron. The figures show only a few of the channels in order to illustrate the wavelength effect on the imagery quality. Note that the best detail is seen in the 0.55 to 0.58 micron, while only land above the water surface or white foam, etc., is shown in the 0.8 to 1 micron [fig. D4(c)]. In the region of 0.62 to 0.66 microns, only the shallowest points are observed while in the blue region, 0.40 to 0.44 micron, scattering of light tends to lower the contrast of objects beneath the water. Much work is being done to try to relate density to depths. One note of caution is made in any attempt to relate density with depth of water. The amount of light reflected back to a sensor is the product of several factors including the transparency of the water and the percent reflection of the bottom material. Thus, one must be careful to distinguish a given density on film as the result of light coming from a higher reflector at a deeper depth or a poorer reflector at a shallower depth, both of which may give the same density. Only in the areas where bottom type and water clarity is uniform, may there be a reliable correlation between density and depth of water.

The water depth for which useful data can be obtained with an optical mechanical scanner is shown in figures D5(a) and (b). The results of a system trade-off analysis is given for two different altitudes, 1000 ft and 200 n miles based on certain assumptions shown on the graph.

The analysis permits one to estimate the size of the primary collecting aperture (a measure of the cost and complexity of the sensor) for a given water transmission value in order to obtain a recognition of a small percent change in the light reflection from the bottom at a given depth with a reasonable signal to noise ratio. For example, from a 200 n mile orbit, an optical-mechanical sensor with 1 mr instantaneous field of view would require a 5 cm diameter in order to see a 10% change in bottom reflection (for S/N = 3 at 0.55 microns) at a depth of 7.5 meters in coastal waters or a theoretical maximum of 30 meters considering distilled water type clarity. However, the ground resolution size would be approximately 1200 feet (400 meters). In order to map smaller shoal areas, a larger diameter system is needed. For example, for a 0.2 mr instantaneous field of view, (approximately 240 feet or 80 meter ground resolution) a diameter of 125 cm is needed to see the same depths.

Another of the observables under investigation is wave refraction. By using an optical processor (see fig. D6), a fourier transform can be obtained of a photograph and a search made for a change in the spatial wavelength content of the water wave pattern in successive frames (see fig. D7).

Several researchers such as Stilwell (NRL)* are attempting to use the optical processor to form a fourier transform of the water wave patterns in order to obtain statistics on wave characteristics. However, they are concerned with measuring both the wavelengths present as well as their relative amplitude based on the density of the film. This is a very difficult task since one has to deal with the sensitometry of film. Fortunately, the application of wave refraction related to shallow water detection may only require one to measure whether a particular wavelength is present in one frame and disappears in another frame. In the optical processor, the fourier transform plane is photographed, and a water wave swell of a give longer wavelength and direction appears as a point near the center of the transform on the film. The shorter wavelengths appear farthest from the center. In a preliminary test, a change in the fourier transform was observed in that a set of waves disappeared in the transformed plane since the wavelength composition changed due to the presence of an island (see fig. D8). What is needed now is a series of photographs of wave refraction taken near shallow depths and then optically processed to measure the refracted wavelength. Previous theoretical studies can be used to estimate water depth from the knowledge of deep water wavelength and the refracted wavelength.

One of the more direct methods of measuring water depths remotely is by the use of a laser beam and attempts to measure the signal return from the water surface and a second

*Private Communication.

return from the bottom. A high powered pulsed laser operating at a carefully chosen wavelength is needed. A preliminary system trade-off analysis was performed for a laser system of 10^5 watts (peak power). Comparison of the size of the receiver collector required for different altitudes of 1000 ft and 200 n miles is shown in figure D9(a) and (b). The quantity noise equivalent water depth, is the depth for which the laser signal return will just equal the system noise. In a practical case, the detected depth will be less since this parameter represents an upper bound. Thus, under the assumption made, a laser system operating in an aircraft at 1000 ft with a receiver collecting area of $0.1 m^2$ will measure a noise equivalent depth of 40 meters in coastal waters (220 meters in distilled water) while at 200 n miles, the same system would only measure a noise equivalent depth of 1.5 meters in coastal waters (8 meters in distilled water). Consequently, large diameter receiver collecting systems will be required or higher power lasers will need to be developed along with cooling systems. Intermittent pulsed laser operations from a satellite is expected to be necessary.

Additional sensor trade-off studies are continuing as well as surveys of shoal characteristics to determine the levels of percent reflectance of underwater surfaces to be encountered and the geographic distribution of water characteristics such as total attenuation coefficients.

Appendix E
POTENTIAL OF REMOTE SENSING TECHNIQUES IN THE STUDY OF ECOSYSTEMS

Two major types of contributions which remote sensing offers to the analysis of ecosystems are: (1) the inventory of ecosystem units (e.g., identification and assessment of spatial and temporal distribution of biotic and abiotic components; and (2) the assessment of those energy relations governing primary productivity which are made manifest by characteristic patterns of reflected and emitted radiant energy.

Inventory of Producers. Plant species of sufficiently different spectral reflectance and emission characteristics may be differentiated by means of their multispectral signatures and by changes in these with time. Mapping of the spatial distributions and their temporal variations will permit information concerning floristic composition, structure, and successional stages of plant communities to be derived.

Herbage cover by species, clumping of individuals of a species and community stratification may be detected within the limitations of resolution of the sensor used. When inventoried seasonally, plant community composition and structure data yield information concerning phenological cycles, and annual variations in primary productivity. The gross geometry of a plant community together with illumination, altitude, and angle of view parameters have been shown to control reflectance characteristics of the remotely sensed vegetation (ref. 21). The study of plant community geometry (ground and aerial based) would provide further information for the identification of plant communities, and for the assessment of the influence of morphological features of vegetation upon primary productivity. Distribution maps may be established not only for plant species but also for different plant communities, plant formations, and for standing crop biomass.

Correlation of data collected by ground and aerial based sensors with data obtained by orbiting satellites may permit the extension of the vegetation inventory to other areas of each biome not intensively studied, and serve as a basis for evaluating rates of biomass change within and between different biomes.

Inventory of Consumers. The determination of species identifications, distributions, numbers, and biomass of selected conspicuous consumers is more difficult but may be made for each biome under optimum conditions. Nighttime thermal infrared sensing has been used successfully for large mammal detection. Numbers and kinds of consumers may also be assessed

by detection of the effects of their activities (e.g., trampling, overgrazing or browsing, etc.) upon vegetation.

Inventory of Abiotic Components. Geomorphological features of ecosystems such as landform, drainage patterns, and erosion patterns may be mapped by means of their shape and contrast characteristics due to thermal and reflective differences. Soil type identification and mapping may be accomplished by aerial based studies of surface color, surface roughness, temperature, and moisture content. Characteristics of soils within experimental areas are studied and by extension of the derived spectral signatures, unknown soil type in other areas are assessed. Geomophological, hydrological and vegetational indicators would also be used to infer soil type. Eventually, if operational systems are developed, the spatial distribution of hydrological characteristics of ecosystems (volume and pattern of surface runoff, size and distribution of snowpack, distribution of thunderstorm activity, surface waters, etc.) may be inventoried and mapped diurnally and seasonally.

Assessment of Energy Relationships within Ecosystems. Some of the types of energy relationships existing between vegetation and environmental factors which may be possible to investigate by remote sensing techniques in conjunction with localized ground based measurements are listed below:

A. Vegetation—Temperature Relationships
   1. Mapping of plant canopy temperatures and correlation with the total energy budget of the plant community.
   2. Identification and mapping of phenological, morphological, and physiological differences of species and ecotypes due to different temperature environments.
   3. Diurnal mapping of heating and cooling rates of different plant communities as an aid to their identification.
   4. Determination of correlation of plant thermal responses with physiological states (e.g., photosynthesis and evapotranspiration rates).
   5. Identification and mapping of temperature-induced plant stress and determination of frost or heat tolerance of species and ecotypes.

B. Vegetation—Light Relationships
   1. Determination of the effects of leaf area density and chlorophyll levels upon spectral reflectance and emission characteristics of species and ecotypes in different light environments (e.g., changes in quality and quantity of light due to plant canopy position, slope aspect, latitudinal and altitudinal differences).

2. Determination of the effects of leaf angle, arrangement and depth of canopy upon spectral reflectance and emissions characteristics of given species and ecotypes.

C. Vegetation—Moisture Relationships

1. Detection and mapping of plant moisture stress and determination of drought hardiness of species and ecotypes.
2. Determination of the correlation between soil and moisture distribution patterns (mapped periodically during the growing season and during the wettest and driest parts of the year) and plant distribution patterns (floristic composition, abundance and vitality) and derivation of plant indicators of soil moisture conditions.
3. Detection and mapping of spectral reflectance and emission changes due to different rates of transpiration, soil moisture conditions, and internal plant water conditions.

D. Vegetation—Soil Relationships

1. Detection and mapping of nutrient stressed vegetation (due to excess or deficient amounts of specific nutrients, or groups of nutrients).
2. Determination of the influence of different edaphic conditions upon the spectral reflectance and emission characteristics of species and ecotypes.

E. Vegetation—Abiotic Relationships

1. Detection and mapping of spectral reflectance and emission changes of plants due to changes in evapotranspiration and photosynthesis rates under given conditions of wind speed, air temperature, solar radiation, carbon dioxide exchange, vapor pressure, and given manipulative treatments (e.g., irrigation, fertilization).
2. Detection and mapping of variations in phenological events of given species due to different regional and local environmental conditions (e.g., changes in quality and quantity of light and range of temperature due to latitudinal and altitudinal differences, micrometerological, edaphic and topographical differences).

F. Vegetation—Biotic Relationships

1. Detection and mapping of vegetation damage due to insect infestation and disease.
2. Determination of the influences of grazing intensity and duration on spectral reflectance and emission characteristics of vegetation.

Appendix F
EFFECTS OF ALTITUDE ON MULTISPECTRAL IMAGERY

F. Polcyn

F.1. INTRODUCTION

As a part of a research program conducted in 1966, multispectral data were collected at several altitudes in order to conduct a study of the effects of atmospheric transmission and scattering on the uniqueness of the vegetation's spectral signature. The analysis was preliminary, but it was done to find the general trend of effects and to provide guidelines for defining future experimental studies.

The initial analysis of the data dealt primarily with the 0.4-1 μ region of the visible spectrum. Three questions were raised:

(1) What are the general effects of altitude upon the apparent radiance of the object?

(2) Are the effects spectrally dependent?

(3) What meteorological parameters correlate with these effects?

F.2. THEORETICAL CONSIDERATIONS

The effect of sensor altitude upon the appearance of a target which fills the sensor's instantaneous field of view can be ascribed to two simultaneous processes. The matter in the path between the target and the sensor (1) attenuates the radiation emanating (by reflection of sunlight or self-emission) from the target by absorption and scattering, and (2) scatters and emits unwanted radiation into the field of view of the sensor. If L(0) is the actual radiance of the target in a small spectral band $\Delta\lambda$, then the apparent radiance, L(h), sensed from altitude h is given by

$$L(h) = \tau(\lambda)L(0) + L(a) \qquad \text{(F1)}$$

where $\tau(\lambda)$ is the path transmittance, and L(a) represents the extraneous radiation emitted by or scattered into the beam collected by the remote sensor. $\tau(\lambda)$ and L(a) are functions of the atmospheric conditions as well as of $\Delta\lambda$ and h. They are, however, independent of the properties of the target.

If equation F1 is differentiated with respect to altitude, then

$$\frac{\partial}{\partial h}L(h) = L(0)\frac{\partial}{\partial h}\tau(\lambda) + \frac{\partial}{\partial h}L(a) \qquad \text{(F2)}$$

Now $\frac{\partial \tau(\lambda)}{\partial h}$ is negative while $\frac{\partial L(a)}{\partial h}$ is positive. Consequently the direction of change with altitude for the apparent target radiance will depend upon their relative magnitudes and upon the magnitude of the actual target radiance. Qualitatively, it can be seen that, if the actual target radiance is quite small, then the apparent radiance may increase with altitude. Conversely, if the actual target radiance is large, and if, in addition, its attenuation is predominately due to absorption rather than to scattering, then the apparent target radiance might very well decrease with altitude.

F.3. DATA ANALYZED

The data chosen for analysis were acquired in July and September 1966 near Lafayette, Indiana. On the morning of July 26, 1966, multispectral imagery was acquired in consecutive flights at altitudes of 2000, 4000, and 6000 ft. On the morning of September 15, 1966, similar imagery was acquired at six different altitudes ranging from 700 to 10,000 ft.

The data analyzed were acquired by the 12-channel spectrometer which senses in twelve contiguous narrow bands over the wavelength range from 0.4 to 1 μ. Table FI shows the spectral range of each of the twelve channels. Once during each complete revolution of the scanning mirror, the spectrometer field of view scanned each of two Fiberfax reflective panels illuminated by calibrated lamp sources, thus providing two calibrated signals in each of the twelve channels. The voltage signals from the twelve channels, including the calibration signals, were recorded in parallel on a 14-channel tape recorder.

TABLE FI. SPECTRAL BANDWIDTHS OF THE 12-CHANNEL SPECTROMETER
(defined for 50% Response Level)

Spectrometer Channel	Spectral Band (μ)
1	0.404-0.437
2	0.437-0.464
3	0.464-0.482
4	0.482-0.502
5	0.502-0.524
6	0.524-0.549
7	0.549-0.580
8	0.580-0.617
9	0.617-0.659
10	0.659-0.719
11	0.719-0.799
12	0.799-1.000

In the laboratory, this recorded data was transformed by electronic sampling techniques into a form suitable for analysis. Electronic gating was used to retrieve samples of the signal from a given ground target simultaneously from each of the twelve channels, for various altitudes. These voltages were then averaged, channel by channel, to produce twelve voltages, V(1), V(2), . . . , V(12), which are proportional to the average apparent radiance of the target in each channel. Similar sampling of the calibration signals produced twelve reference voltages, V(r, 1), V(r, 2), . . . , V(r, 12), proportional to the known radiances of each of the calibrated lamps. The absolute apparent radiance of the target L(h) was then computed as

$$L(h, \lambda_1) = \frac{V(1)}{V(r, 1)} L(r, 1)$$

By this method, values were obtained of the absolute apparent radiances as a function of spectrometer channel and altitude for wheat stubble and soybeans on July 26, and for corn and soybeans on September 15.

F.4. RESULTS

Figure F1 is a typical example of the type of data obtained. It presents the apparent radiance of soybeans versus spectrometer channels for three altitudes. This is not a true spectral representation since each channel measures the average radiance in a wide (by spectrometry standards) band, and the straight lines connecting points are only for convenience. There is, however, a recognizable spectral shape as seen by radiance peaks in the green (channel 7) and near infrared (channels 11 and 12), separated by several channels of reduced response which include the chlorophyll absorption region.

Figures F2 and F3 present the data in terms of a normalized radiance (radiance ratio) versus altitude. The normalization is relative to the radiance at 2000 feet; vis.:

$$\frac{L(h)}{L(2{,}000)} = \frac{\tau(\lambda, h) L(0) + L(a, h)}{\tau(\lambda, 2000) L(0) + L(a, 2000)} \qquad \text{(F3)}$$

Such a ratio provides a means of comparing the relative change with altitude in apparent radiance among the several spectral bands.

In general, the ratio, and thus the apparent radiance, is seen to increase or remain constant with altitude in all spectral channels. This indicates that the scattering of radiation into the beam (the term L(a, h) in eq. F3) is greater than or equal to the loss due to the attenuation of the actual target radiance by the path.

The exceptions occur at the longer wavelengths for the soybean data on September 15. These anomalies may not be real, however. Because of the many altitudes flown on September 15, the intensity of solar illumination increased considerably between the 10,000-ft flight (0850 hours) and the 700-ft flight (1024 hours). An attempt was made to account for the illumination change for each altitude by use of ground-based solar illumination measurements made at the time. The correction procedure used was necessarily approximate since the ground illumination measurements were made by a very broadband instrument, and the integrated values had to be divided into contributions from each of the twelve channels.

A very definite spectral trend is evident in figures F2 and F3. The ratio of apparent radiances increases much more with altitude in the blue wavelengths (channels 1 and 2) than in the longer wavelengths. This spectral variation results from the fact that the scattering effect of the atmosphere varies inversely with wavelength. The flatness of some of the longer wavelength curves also indicates that the tendency of scattered radiation to increase the apparent radiance can be offset by the tendency of attenuation to decrease the apparent radiance.

The effect of atmospheric conditions is very evident in a comparison of the July 26 data and the September 15 data. On July 26, the air was quite hazy; on September 15, it was quite clear. As a result, the radiance ratios for the July data increase much more rapidly with altitude, so much so that the radiance ratio of 6000 ft on July 27 exceeds that at 10,000 ft on September 15.

The foregoing data show the combined effect of attenuation and scattering into the beam. The effect of attenuation alone may be studied by looking at the apparent radiance difference between two targets as a function of altitude. This difference is defined as

$$L(h, 1) - L(h, 2) = \tau(\lambda)(L(0, 1) - L(0, 2))$$

Since $\tau(\lambda)$ cannot increase with altitude, it is evident that the apparent radiance difference cannot. Figure F4 shows the difference in apparent radiance between soybeans and wheat stubble on July 27 versus spectrometer channel for the three altitudes flown. An increase of 0.1 mw-sr^{-1}-cm^{-2}-μ^{-1} in the apparent radiance difference between 2000 and 6000 ft in channel 1 of the figure is indicative of the relative uncertainty in the calculated differences, since such an increase is not possible. The magnitude of the uncertainty is even greater for channels 11 and 12 because of excessive noise in the instrumentation.

Figure F4 indicates that any change in $\tau(\lambda)$ between the several altitudes was no more than the uncertainty in our measurement. If, for example, the uncertainty is taken to be

0.1 mw-sr^{-1} cm^{-2} μ^{-1}, then the change in $\tau(\lambda)$ between 2000 and 6000 ft could be no more than $100 \times 0.1/1.0 = 10\%$ in channel 9 and $100 \times 0.1/0.2 = 50\%$ in channel 1.

In section F2 it was stated that, according to theory, the rate of change with altitude of apparent radiance depended upon the magnitude of the actual target radiance in such a way that, the larger the actual target radiance, the less positive the rate of change. This was shown in equation F2 because $\partial\tau/\partial h$ was negative. Similarly, if the radiance ratio (eq. F3) were differentiated with respect to altitude, the same relationship would hold. Thus, according to theory, if for two targets (1 and 2)

$$L(0, 1) > L(0, 2)$$

it must follow that

$$\frac{L(h, 1)}{L(2000, 1)} < \frac{L(h, 2)}{L(2000, 2)}$$

Table FII shows the results of such a comparison for the data of 27 July 1966. The first column indicates the sign of the difference between the actual radiances of wheat stubble and soybeans as implied by figure F4. The second column indicates the sign of the differences between the apparent radiance ratios at 6000 ft as presented in figure F2. It will be noted that, for every spectral channel, the two differences are of opposite sign, as predicted. Table FIII presents the same comparison for the September 15 data on corn and soybeans. The apparent radiance ratio differences were taken from figure F3 for the 10,000-ft data. Again, the differences are of opposite sign in each spectral channel, as predicted by theory.

On the basis of this exploratory study in the visible region of the spectrum, the following preliminary conclusions can be drawn.

(1) Increase in altitude increases apparent target radiance.

(2) The magnitude of the increase in apparent radiance with altitude is inversely related to wavelength.

(3) The increase in apparent radiance with altitude is directly related to the amount of particulate matter (e.g., haze) in the path.

(4) Attenuation of the spectral radiance differences between objects is small compared to the increase in individual radiances with altitude.

(5) The increase in apparent radiance with altitude can be lessened, or even reversed, if the actual target radiance is large enough.

TABLE FII. DIFFERENCE IN ACTUAL TARGET RADIANCES COMPARED WITH DIFFERENCES IN APPARENT RADIANCE RATIOS, WHEAT STUBBLE AND SOYBEANS, JULY 27, 1966

Spectrometer Channel	Sign of Radiance Difference (wheat minus soybeans)	Sign of Apparent Radiance Ratio Difference (2000 to 10,000 feet)
1	+	-
2	+	-
3	+	-
4	+	-
5	+	-
6	+	-
7	+	-
8	+	-
9	+	-
10	+	-
11	-	+
12	-	+

TABLE FIII. DIFFERENCE IN ACTUAL TARGET RADIANCES COMPARED WITH DIFFERENCE IN APPARENT RADIANCE RATIOS, CORN AND SOYBEANS, SEPTEMBER 15, 1966

Spectrometer Channel	Sign of Radiance Difference (corn minus soybeans)	Sign of Apparent Radiance Ratio Difference (2000 to 10,000 feet)
1	-	+
2	-	+
3	-	+
4	-	+
5	-	+
6	-	+
7	-	+
8	-	+
9	-	+
10	-	+

Because of the very tentative nature of the above conclusions, it is recommended that further, more intensive investigations be conducted, especially regarding (4) above. The most important technical problem to be overcome in such an investigation is concerned with the necessity of reducing to a minimum the elapsed time between the first and last run over the target, in order to insure fairly constant illumination conditions. This may best be realized by starting with the highest altitude run first, and limiting the target area to a very small size in order to minimize the time taken for each run.

Appendix G
DATA PROCESSING AND OPERATIONAL PROCEDURES

P. Hasell

G.1. SIGNAL PROCESSING

Single band imagery. The tape recorded airborne scanner data channels can be reproduced on film strips in post flight playback. The resulting continuous film strip imagery can have a calibrated relationship between the video signal voltage from the magnetic tape and film tone through a multi-step linear voltage scale introduced periodically in place of the video signal. The video voltage gain and DC offset can be adjusted in playback to place any portion of the tape signal over a preselected grey scale range. The choices in single band film strip reproduction are described.

Normal tape voltage to film tone print. In the normal reproduction of film strip imagery, the lightest film grey scale tone is clamped electrically to the signal representing the darkest or coldest reference source in the video signal and the voltage gain is adjusted to place the signals representing the brightest or hottest objects of interest at the level of the darkest film grey scale tone. This provision takes full advantage of the film dynamic range but it does place the signal extremes in the nonlinear portion of the film transfer characteristics.

Special tape voltage to film tone print. The video signal level can be clamped electrically to the signal representing the darkest or coldest reference source for calibration purposes. Then the video voltage can be shifted in DC level and adjusted in gain to produce any preselected range of film grey scale tone. Calibration reference levels may be lost in the film print because they are usually set to match the expected signal extremes. However, the signal levels of interest (including calibration sources) can be placed within the linear region of the film grey scale or small signal variations can be expanded over a wide grey scale range for ease of data analysis.

Amplitude gating of tape voltage to film tone print. As an aid in data analysis, either or both an upper and lower voltage threshold level can be set in playback after electrical clamping such that signals outside these limits will not be printed on film. Again, some calibration may be lost since the reference sources are usually adjusted to match the signal extremes.

Quantized tape voltage to film tone print. As a further aid in imagery analysis, the tape signal may be quantized to a preselected number of bits before printing on film. In effect, the signals are forced to adjust to the nearest quantized level before printing and incremental contours are therefore produced in the imagery. This effect is very pronounced in a 2 bit (4 level)

quantization and is hardly apparent in a 4 bit (16 level) one. Of course data accuracy is normally degraded in the quantization process.

Multiband imagery or data. The airborne scanner data channels can be recorded as electrical signals on magnetic tape and thus can be easily subjected to electronic processing for extraction of desired information during data playback. The extent of multiband processing is limited only by the number of scanner channels that are in exact registration. The University of Michigan airborne system in its current configuration gathers data in four groups ((1) UV, (2) Visible and NIR, (3) Mid IR and (4) FIR and only the data within a group have common registration. The data in all groups are gathered at the same time but each group views a different spot on the ground at any instant of time. In general, electronic processing of multiband scanner signals collected by The University of Michigan is limited to groups (2) (0.4 to 1.0 μ) and (3) (1.0 to 5.5 μ).

Simple addition and subtraction of bands. The simplest type of electronic multiband processing is addition and/or subtraction of selected wavelength bands in a data group. In this manner, the contrasts between say foliage and bare ground can be emphasized, but the technique tends to be limited to contrast enhancement of gross object classes.

Two band discrimination. A somewhat more powerful discrimination technique is the area or position gating of two selected scanner data channels whose amplitudes are displayed orthogonally in an x-y plot, one channel assigned to each axis. Objects in a scene tend to have a unique relationship in the relative signal level between channels for a given set of environmental conditions. The electronic gating of this unique relationship can provide a film strip printout of only those objects in a scene which match the relationship.

Multiband discrimination. If we carry the two band discrimination technique mentioned above to n dimensional space (n limited by the number of data channels in a group), then we have described the most sophisticated class of discrimination processing that can be provided. Within this class, there are many selections of processing techniques which essentially provide different electronic gating boundaries and different decision logic. The two processing techniques which we have used the most are (1) absolute value distance measuring and (2) maximum likelihood ratio determination. The techniques and equipment at The University of Michigan provide for the accomplishment of this sophisticated multiband processing in real or slowed time. The results are normally displayed on an oscilloscope to viewers and recorded on 70mm film strips for interpretation.

Statistical analysis of multispectral data. The signals received from various objects and materials on the ground exhibit variations for a variety of reasons. In many applications, it is desirable for analysis to obtain statistical descriptions of these multispectral signals and their variations. Facilities exist at The University of Michigan for obtaining such descriptions from these objects or areas in the scene that are selected for detailed study. The following types of information can be computed:

(a) mean signal in each channel and the corresponding variances for:
  (1) each selected scan line that contains the material of interest
  (2) each scan angle at which the material is observed and sampled
  (3) all samples considered as a group
(b) covariances and correlations between signals in the various channels
(c) eigenvalues and eigenvectors of each variance/covariance matrix or distribution
(d) weighted combinations of statistics from several objects or areas
(e) measures of the amount of separation between such distributions for any set of spectral channels

The data computed have the following features:

(a) corrections are first made for scan-angle-dependent variations that are introduced into the data by the scanner
(b) calibration information is simultaneously computed
(c) the results are available in one or more of the following forms: computer printouts, punched cards, magnetic tapes and computer-generated plots

Generation and analysis of spectral signatures. The statistical parameters discussed above are initially given in terms of the voltages that are present on the raw data tapes. In order to relate them to physical parameters and provide a basis for comparison with other data, they must be calibrated in terms of average equivalent spectral radiance, equivalent reflectance, or some related quantity. The calibration of the data requires knowledge of both the physics involved and the specific characteristics of the data collection system. We have developed the procedures necessary for calibrating the data so that broad band spectral signatures can be retrieved through machine processing of the multichannel airborne scanner signals.

G.2. IMAGERY INTERPRETATION AND ANALYSIS

Single band imagery. The interpretation and analysis of single band imagery is a function of the form of the imagery and what is to be determined from it. The imagery forms were

previously discussed. Reviewed here are some of the things that can be done with the imagery. Usually, extensive ground truth information is needed for meaningful imagery analysis for exploratory applications.

Prepare displays of imagery. A very useful display for a quick review of imagery information content is a chart board display of a selected scene as it appears in multiple wavelength bands. The original single band imagery of a scene is arranged in a selected order on a large chart board which can be used directly or rephotographed for convenient reproduction in 8 1/2 × 11 size prints. Such charts could be made up to show a time history of a scene in selected wavelength bands, or whatever imagery displays may help in a visual interpretation of an observed phenomenon. Normal and false color views of the imagery can be made by printing appropriate scanner wavelength bands a particular color and superimposing these colors in final printing.

Analyze imagery to relate observables to user interests. The scanner imagery quantitatively displays reflectance and/or thermal contrasts in a scene. These contrasts must be related to the physical characteristics of objects in the scene if the data is to have useful applications. The associated areas of technology to make these relations are available at The University of Michigan.

Analyze scanner data to determine spectral signatures of selected objects. The electrical signals representing the radiation of an object in multiple wavelength bands can be displayed on an oscilloscope or chart record for determination of the object's spectral signature. The extraction of raw spectral signatures in this manner is much more accurate and less time consuming than densitometer measurements from the imagery. The restriction in this technique is the size of the object. It must be several resolution elements in size because of necessary electronic gating guided by visual recognition of the object in the scene of an oscilliscope display.

Analyze imagery to determine thermal contours or object patterns. Experienced personnel and the supporting equipment must be available for imagery analysis to establish the apparent temperature of objects or the pattern of temperature variations in a scene. Likewise, the reflectance characteristics of a scene can be determined when sufficient illumination information is known.

Multiband processed imagery. The processed imagery is normally an identification of objects or areas of common spectral characteristics. The interpretation or analysis of the processed results amounts to correlation with other data to determine how well known objects

were recognized or to use the identifications as an input to other studies. Aside from our interests and capabilities to pursue studies using the processed data, we can assist others in their studies by providing the following:

Recognition success. Using other object and terrain identification data usually provided by the data collection sponsor, the processed results can be analyzed to determine the probability of detection and false alarms in the recognition process.

Display of recognition results. The recognition results are usually printed as film strip imagery preserving the shape pattern of the original scanner imagery. The recognition of several different targets in a scene can be displayed as: (1) each recognized target as a selected film tone or color on a separate negative transparent film strip which may overlay the original imagery, (2) all recognized targets as different selected film tones or colors on a single negative transparency, and (3) all recognized targets as different tones or colors on a single positive film print.

G.3. SCANNER OPERATIONAL PROCEDURES

Types of equipment performance checks

Periodic—thorough ones performed several times per year

Preflight—functional checks before a flight mission

Inflight—monitoring of signal appearance during mission

Postflight—functional checks to note changes in performance during mission

Periodic checks

Radiation detector characteristics—i.e., noise level, sensitivity, spectral response

Reference source characteristics—i.e., spectral response, output compared to standard, uniformity

Optical alignment, focus and surface conditions

Video signal gain, linearity and frequency response

Data record and retrieval characteristics—i.e., signal to noise, frequency bandwidth, data synchronization

Inflight checks

Visual appearance of signals

Proper gain and offset

Proper relation between signal and references

Registration of known ground targets

Postflight checks

Operation with static inputs of known characteristics

Complete identification of equipment and data collection conditions in flight records

Operational procedures

Equipment warmup, detector cooling

Functional checks

Make equipment settings for desired performance

Monitor signals to note proper settings, record settings

Operate data recorder over selected targets

Secure equipment after operation

Complete data records, label stored data

Transfer data to laboratory

Review data in laboratory to confirm correct coverage and satisfactory equipment performance

Rules for successful scanner operation

Know limitations and idiosyncrasies of your equipment

Plan your mission well and within the limitations of your equipment and environment

Thoroughly brief all associated with mission

Follow your plan, make changes only when necessary to fit new circumstances

Have operators use a check list to avoid human error

Concentrate on data quality rather than quantity

Keep thorough record of data collection conditions such as weather, equipment description, malfunctions, equipment settings

G.4. PLANNING AND CONDUCTING A FIELD MEASUREMENT OPERATION

Scope of Planning

Objective—what are we to accomplish?

Logistics—how do we get people and supplies to right place?

Supporting functions—what help do we need?

Limitation of successful operations—time, weather, funds, equipment performance

Coordination—timeliness of functions

Thorough Definitions of Objectives

Specifically what is to be accomplished?

How can it best be achieved within limitations?

Establish priorities on multiple objectives

Logistics

Establish base of operations

Arrange for transport and housing of personnel and supplies in a timely manner

Supporting Functions

Communications and navigation

Marking of target areas

Ground truth provisions

Operating Restrictions

Tolerable weather conditions

Equipment performance limitations

Available time at the site

Coordination

Reliable communications

Delegated responsibilities and authority from a single coordinator

Adherence to a published plan

Execution of Typical Missions

General description of typical missions

(1) Agricultural data from Purdue Agronomy Farm, summer 1966

(2) Water polution data from TVA rivers and reservoirs, summer 1967

(3) Geological and hydrological data from Florida, summer 1967

(4) Thermal balance and ecology at Barrow, Alaska, fall 1967

Problem Areas

(1) Nonavailability of liquid coolants (particularly liquid helium) in vicinity of base of operations

(2) Unreliability of freight shipments of liquid coolants

(3) Unsatisfactory weather conditions

(4) Detector failures

(5) Equipment performance limitations

(6) Inadequate ground truth data

(7) Improper operator setting of controls

(8) Inadequate planning of data utilization

Mission Parameters

Altitude above terrain. High as possible for large area coverage, low as possible for registration of small objects, constant within run for ease of data scaling. For reliable multispectral recognition, object needs to appear in 4 or 5 spatial resolution

elements. For registration of a statistically reliable multispectral signature, object should appear in ten times that many resolution elements

Ground speed. High for large area coverage, low for maximum sensitivity and signal to noise, constant for ease of scaling data.

Scan rate. Should be adjusted for desired scan overlap at selected ground speed and altitude (v/h). A scan overlap of 50% is usually optimum for uniform data display.

Ground track. Coverage obtained in straight line segments to minimize data distortion and signal obscuration caused by aircraft maneuvers. Scanner usually provides roll stabilization for small bank angles caused by turbulence, but FOV is obscured by viewport for large bank angles. Normally no pitch or yaw (directional) corrections are made to the data so that these motions distort the data display.

REFERENCES

1. Proceedings of the IRE, "Special Issue on Infrared Physics and Technology," The Institute of Radio Engineers, September 1959.
2. Holter, M. R., et al., Fundamentals of Infrared Technology, MacMillan, 1962.
3. Handbook of Military Infrared Technology, Editor William L. Wolfe, The University of Michigan, Ann Arbor, 1965.
4. Manual of Photographic Interpretation, American Society of Photogrammetry, George Banta Co., Menasha, Wisconsin, 1960.
5. Lowe, D. S., Braithwaite, J. G., and Larrowe, V. L., "An Investigative Study of a Spectrum-Matching Imaging System," The University of Michigan Report on Contract NAS8-21000, October 1966.
6. Hemphill, W. R., "Application of Ultraviolet Reflectance and Stimulated Luminescence to the Remote Detection of Natural Minerals," Presented at the 1968 American Society of Photogrammetry Convention, 11-15 March, Washington, D. C.
7. Sabins, F. F., Jr., "Infrared Imagery and Geologic Aspects," Photogrammetric Engineering, July 1967, pp. 743-750.
8. Nordberg, W., "Inference of Properties of the Earth from Satellite Measurements of Infrared Emission," pp. 567-584, Progress in Astronautics and Aeronautics, Vol. 20, 1967, Academic Press, Inc., New York.
9. Gates, D. M., Keegan, H. J., Schleter, J. C., Weidner, V. R., "Spectral Properties of Plants," pp. 11-20, Applied Optics, Vol. 4, No. 1, January 1965.
10. Bowers, S. A. and Hanks, R. J., "Reflection of Radiant Energy from Soils," pp. 130-138, Soil Science, Vol. 100, No. 2, 1965.
11. Lyon, R. J. P., "Evaluation of Infrared Spectrophotometry for Compositional Analysis of Lunar and Planetary Soils: Rough and Powdered Surfaces," Final Report, NASA Contract NASr49(04), 1964.
12. Hunt, G. R., "Rapid Remote Sensing by a Spectrum Matching Technique," pp. 2919-2930, J. of Geophys. Research, Vol. 71, June 15, 1966.
13. Meyers, V., Wiegand, C. L. Heilman, M. D., Thomas, J. R., "Remote Sensing in Soil and Water Conservation Research," Proc. Fourth Symposium on Remote Sensing of Environment, The University of Michigan, Ann Arbor, Michigan, April 1966.
14. Physics of Plant Environment, Editor W. R. Van Wijk, North-Holland Publishing Company, Amsterdam, 1963.
15. Geiger, R., The Climate Near the Ground, Harvard University Press, Cambridge, Massachusetts, 1965.
16. Middleton, W. E. K., and Spilhaus, A. F., Meteorological Instruments, University of Toronto Press, 1953.
17. Platt, R. B., and Griffiths, J. F., Environmental Measurement and Interpretation, Reinhold Publishing Corp., 1965.

18. Hasell, P. G. and Larsen, L. M., "Calibration of Airborne Optical Sensors," The University of Michigan Report 6400-137-T, September 1968.

19. Horvath, R. and Bornemeier, D. D., "Infrared and Photo Record Analysis, Volume I: A Mathematical Predictive Model for Target Temperature as A Function of Environment," RADC TR66-117, Research and Technology Division, RADC, Griffiss AFB, New York, April 1967.

20. Gilg, J. G. and McConnel, J. J., Jr., "Non-Existent Seamounts—A Case Study," Information Manuscript No. IM 66-28, U. S. Naval Oceanographic Office, Washington, D. C., September 1966.

21. Polcyn, F. C. and Malila, W. A., "Investigation of Spectrum-Matching Sensing in Agriculture," The University of Michigan, Willow Run Laboratories, Semi-Annual Report on Contract NsG715/23-05-071, September 1967.

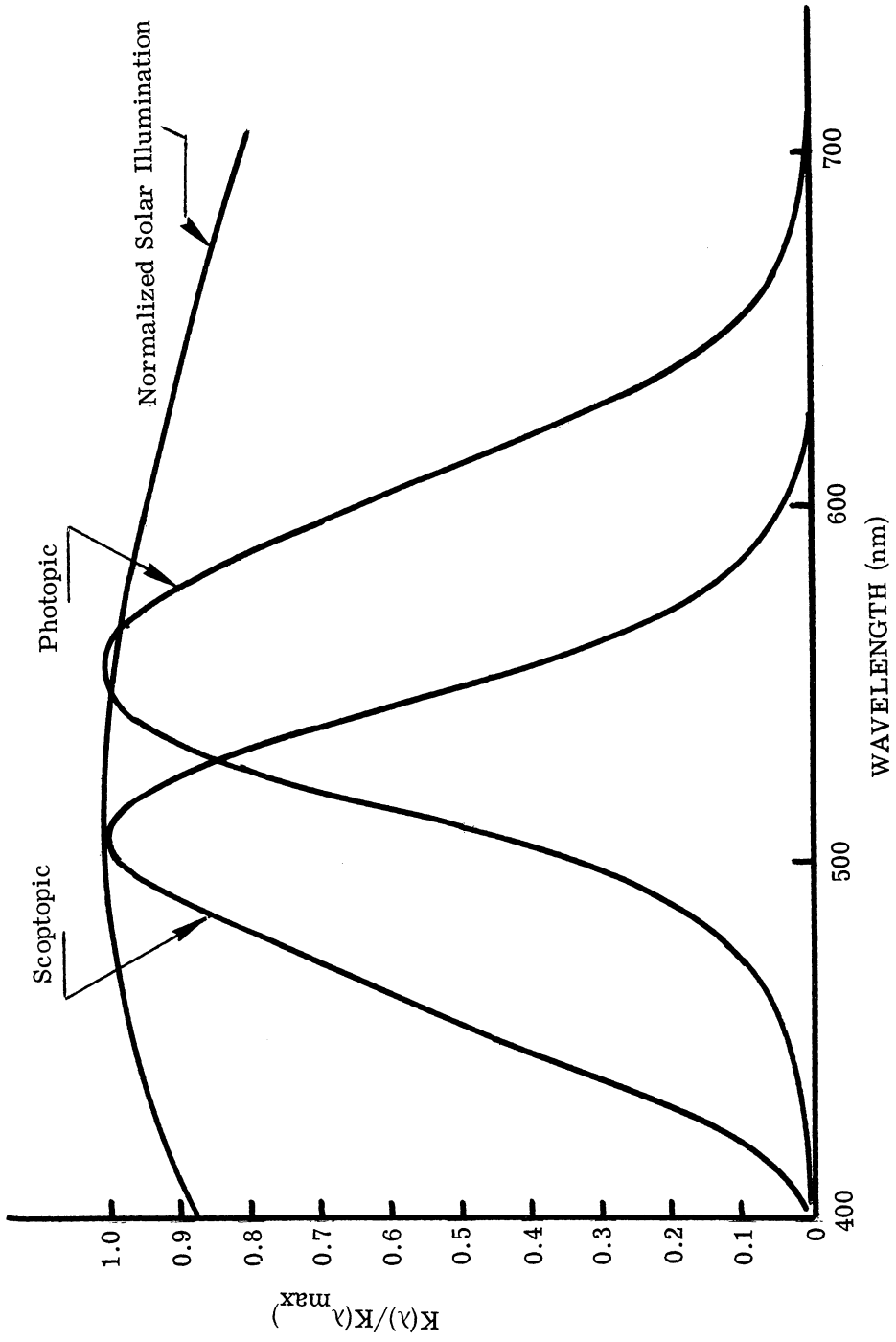


FIGURE 1-1. COMPARISON OF THE SPECTRAL CURVES OF HUMAN VISIBILITY AND SOLAR ILLUMINATION. Normalized luminosity versus wavelength for photopic (daytime) and scotopic (nighttime) vision as compared to solar illumination.

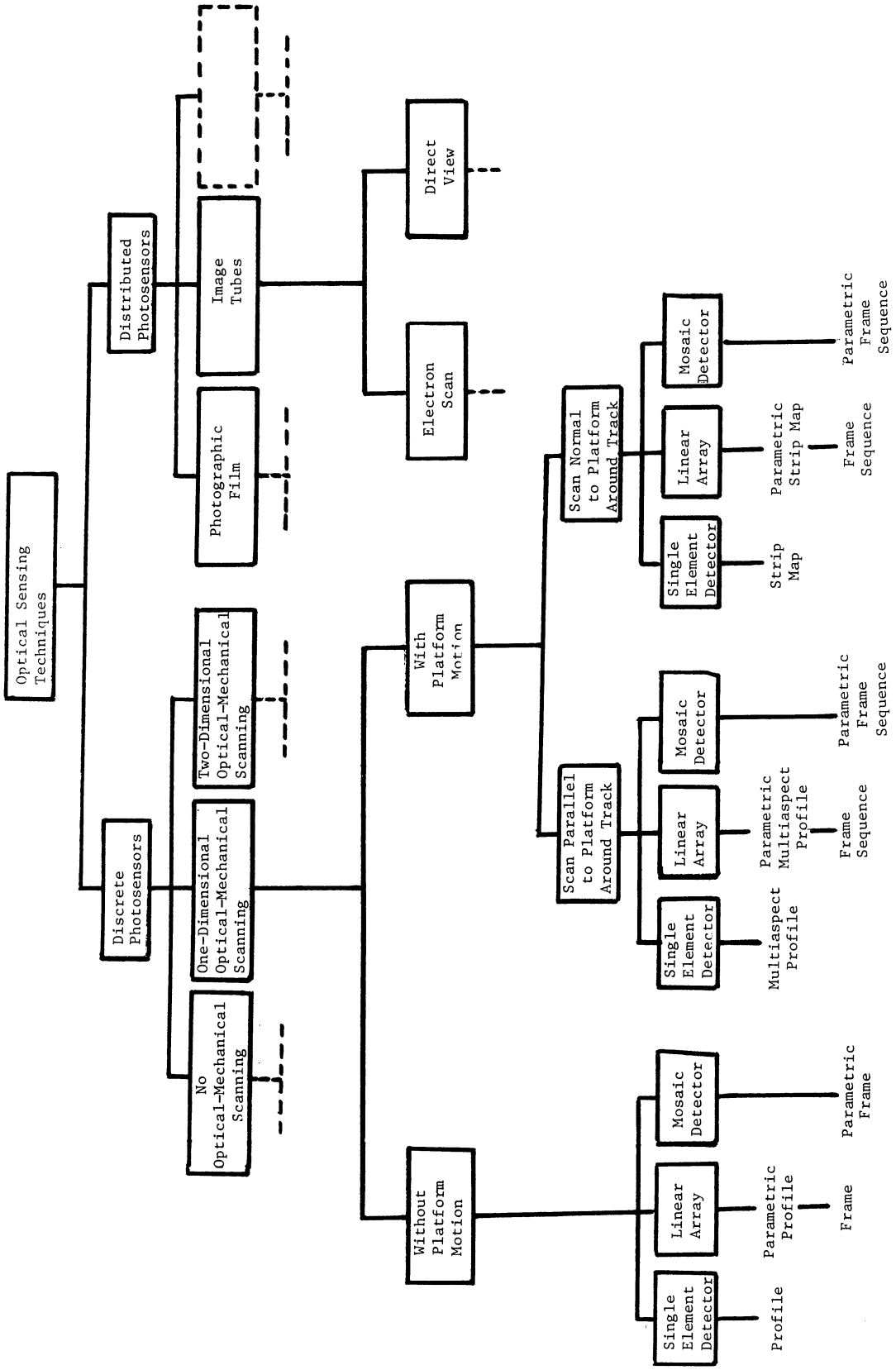


FIGURE 2-1. CLASSIFICATION OF OPTICAL SENSING INSTRUMENTATION

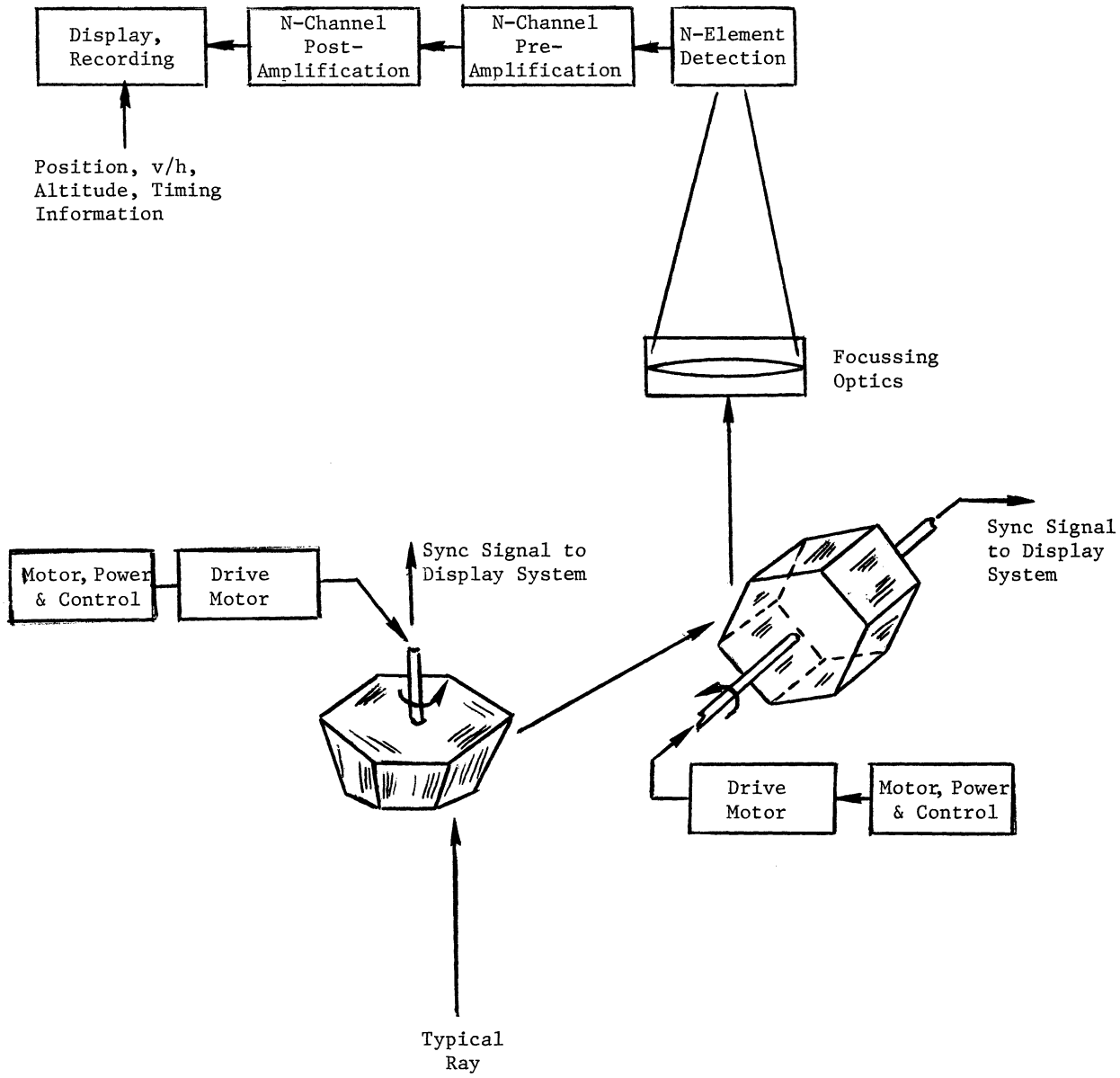


FIGURE 2-2. EVOLUTION OF A "FAMILY" OF OPTICAL INSTRUMENTATION FROM SINGLE BASIC INSTRUMENT. (a) With no scanning, a simple radiometer can observe temporal changes in single or multiple bands. By adding platform motion one can obtain a radiation profile or frame along ground track. Also see figure 2-1. (b) Addition of a one-dimensional scan motion to basic radiometer resulting in a multiaspect profile or framing sequence. (c) Addition of a two-dimensional scan motion to basic radiometer.

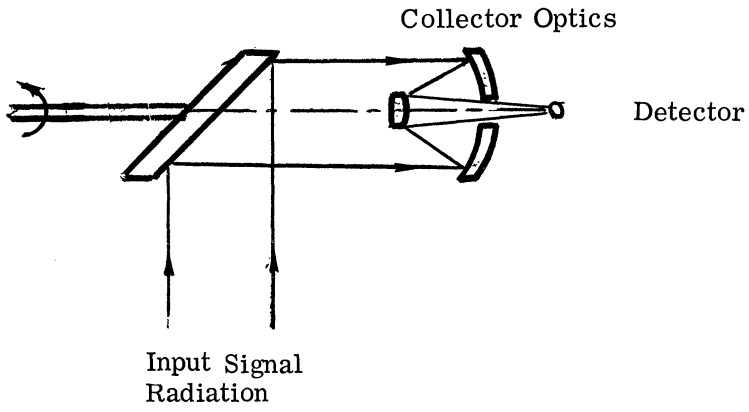
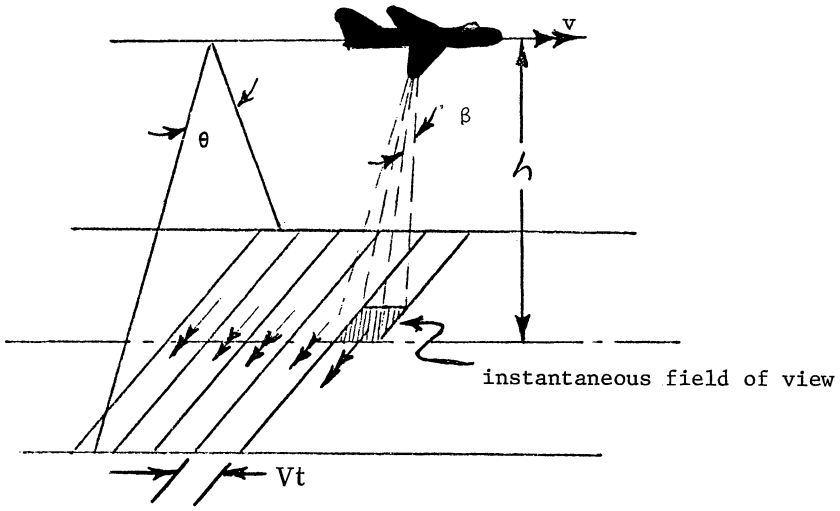


FIGURE 2-3. OPTICAL SCHEMATIC OF BASIC LINE SCANNER



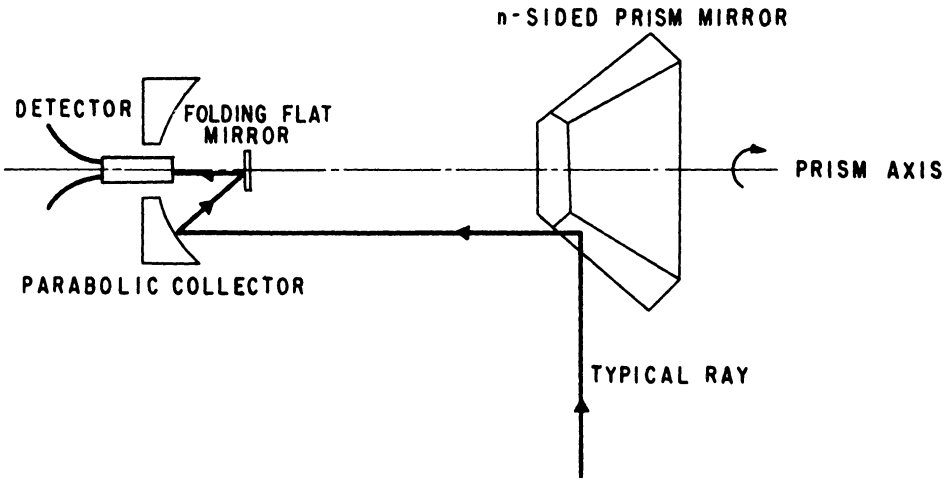
v = aircraft ground speed

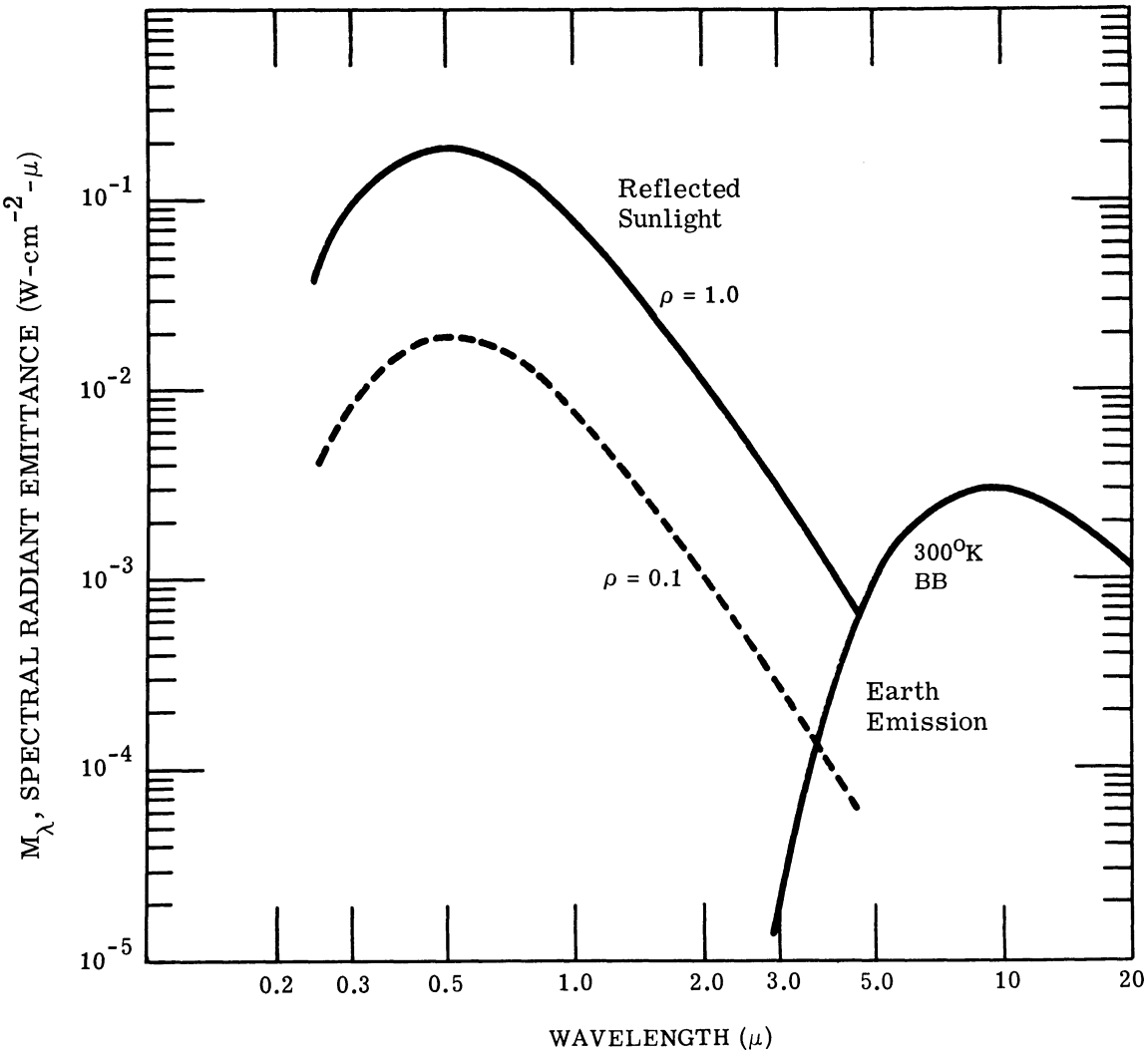
h = aircraft height above terrain

θ = total angle of scan

β = instantaneous angular field of view

FIGURE 2-4. AIRBORNE SCANNER-SCENE GEOMETRY


n-SIDED PRISM MIRROR
DETECTOR
FOLDING FLAT MIRROR
PRISM AXIS
PARABOLIC COLLECTOR
TYPICAL RAY



ENERGY AVAILABLE FOR REMOTE SENSING

FIGURE 2-6. ENERGY AVAILABLE FOR REMOTE SENSING

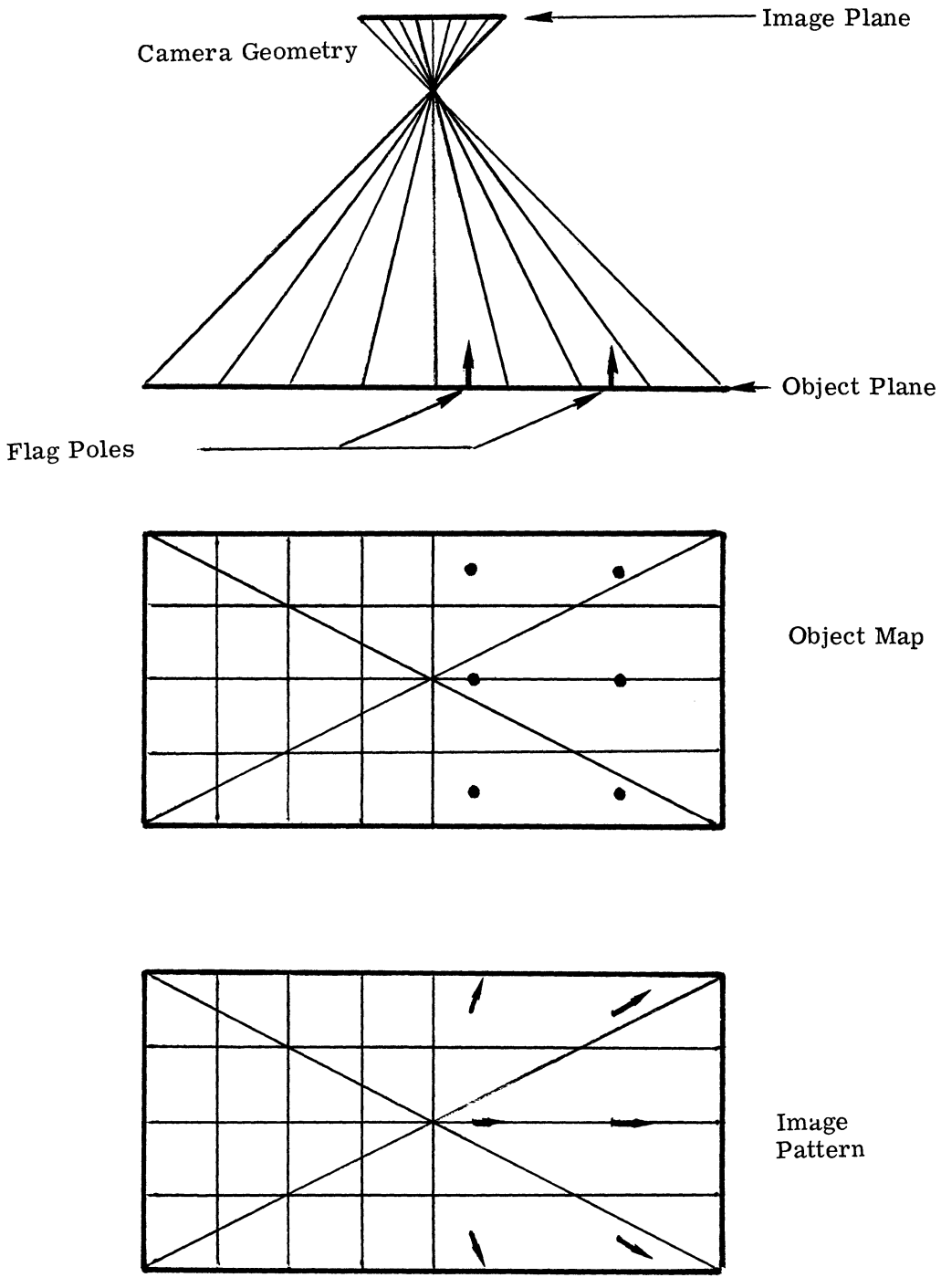


FIGURE 2-7a. IMAGE PLANE PRESENTATION OF MAPPING CAMERA FOR TYPICAL AIRBORNE OBSERVATION

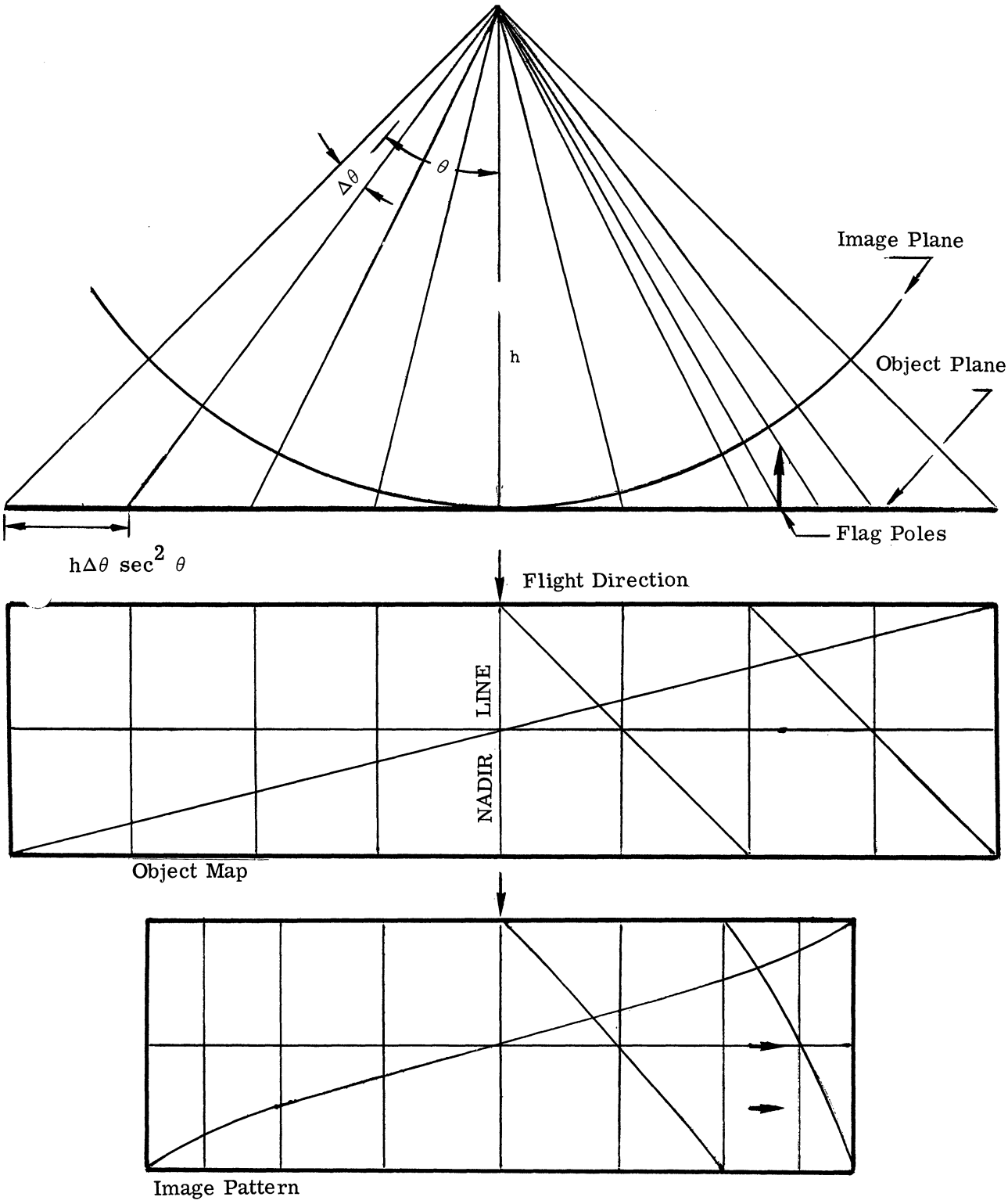


FIGURE 2-7b. SCANNER DISTORTIONS (CONSTANT ANGULAR SCAN RATE)

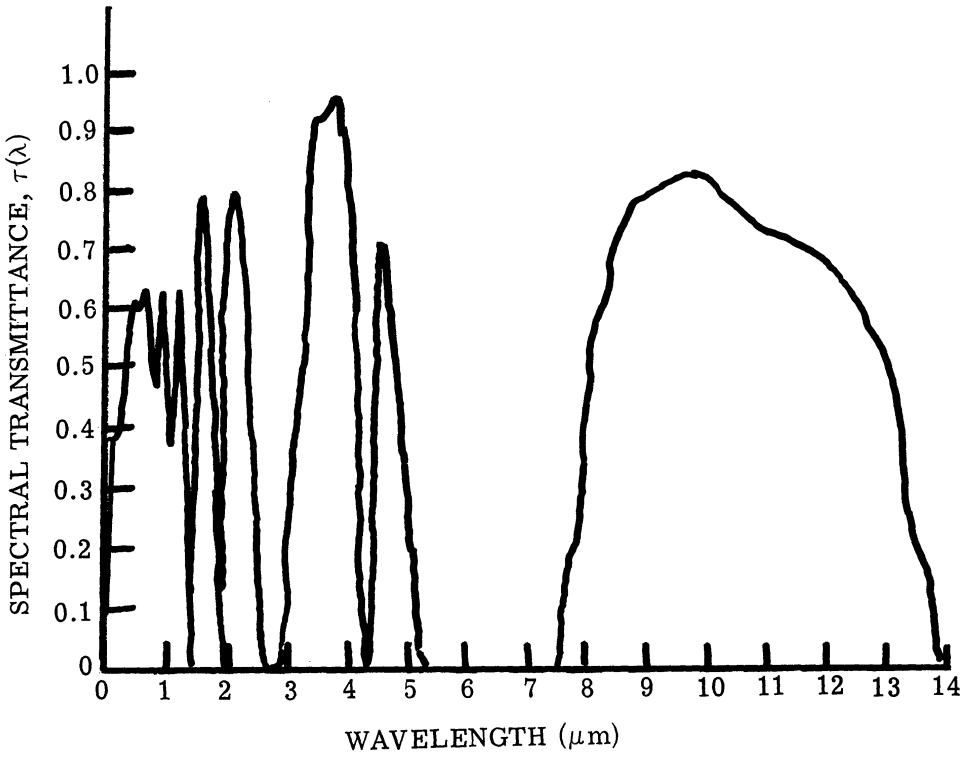


FIGURE 2-8. TYPICAL TRANSMISSION SPECTRUM OF THE ATMOSPHERE

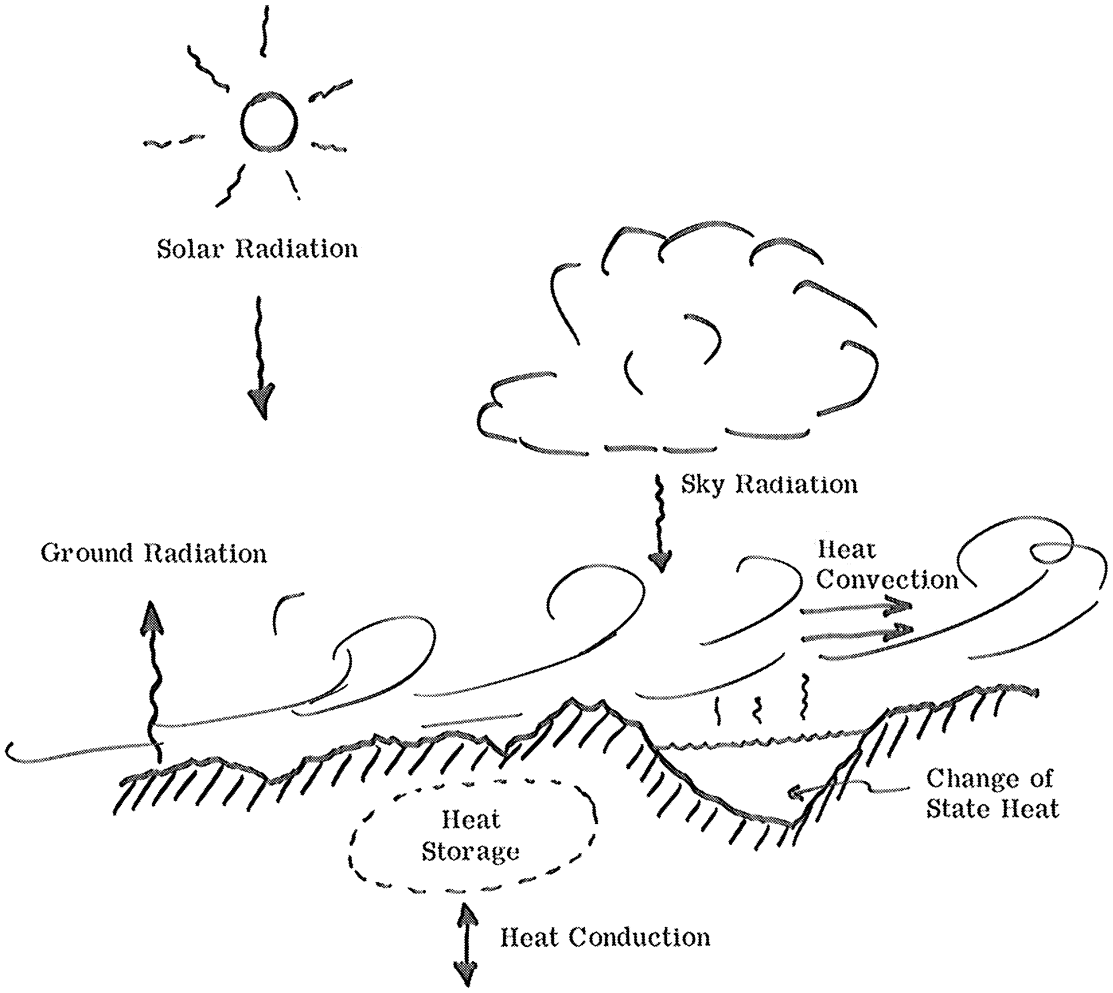


FIGURE 2-9. SCHEMATIC OF FACTORS IN RADIATION AND HEAT EXCHANGE

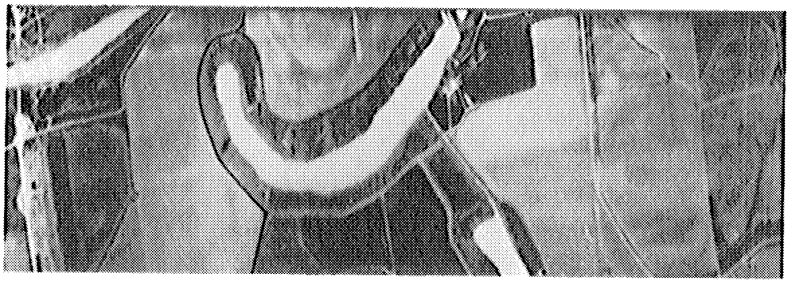


FIGURE 2-10. THERMAL IMAGE OF AN AGRICULTURAL SCENE, 8-13 MICRONS

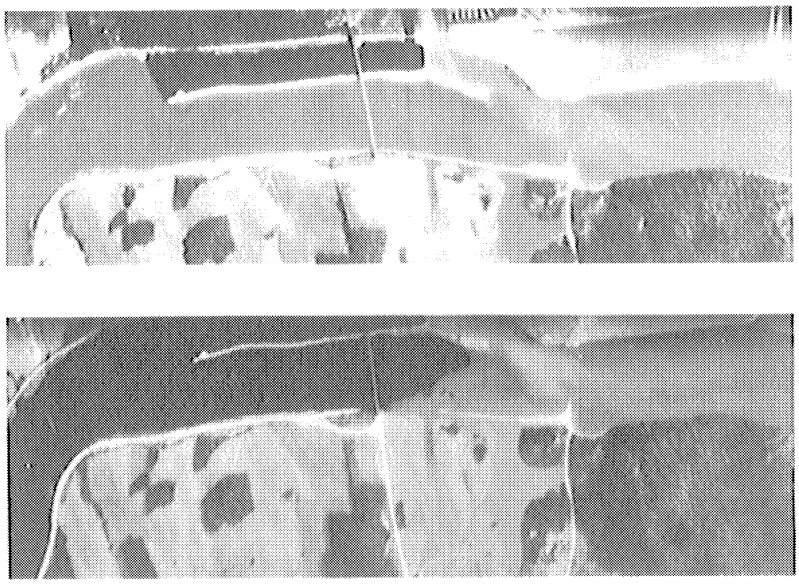


FIGURE 2-11. TIME-LAPSE THERMAL IMAGES OF RIVER SHOWING SLUG OF COLD WATER ARRIVING FROM A DAM 20 MILES UPSTREAM. Upper, at time of release; lower, 24 hours later as cold water arrived at this point. 8-13 microns.

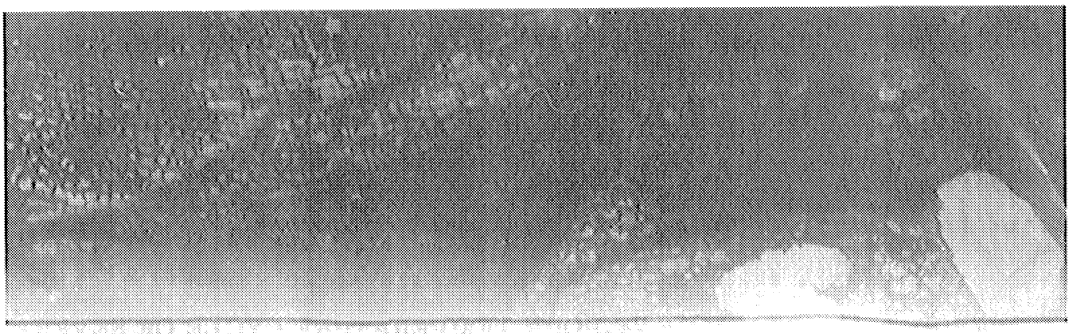


FIGURE 2-12. THERMAL IMAGE GENERATED AT PT. BARROW, ALASKA, OCTOBER, 1967. Note differentiation between snow-covered lakes and dried lake beds in the polygonal ground structure. 8-13 microns.

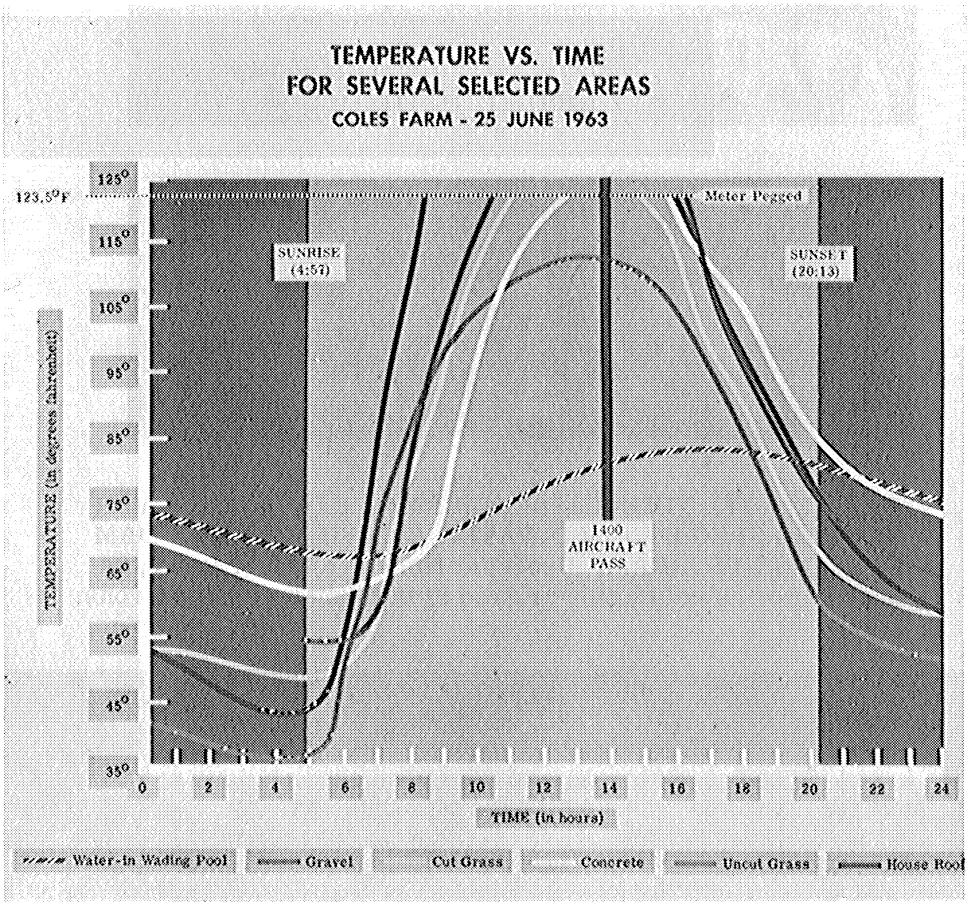


FIGURE 2-13. GRAPH OF 24-HOUR TEMPERATURE CYCLE OF SEVERAL SURFACES. Note temperature differentiation of different materials as a function of time of day. Water temperature is fairly uniform throughout the day, while roof, gravel, and grass surfaces undergo relatively rapid temperature changes at sunrise and sunset. 25 June 1963.

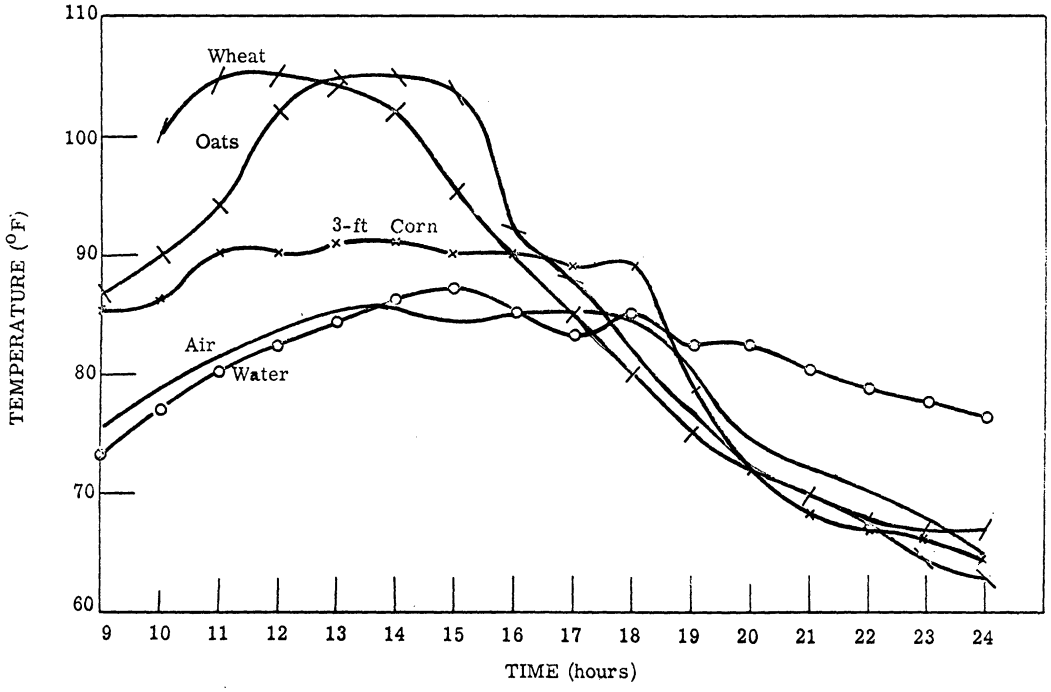


FIGURE 2-14. TEMPERATURE VARIATIONS OF CROPS

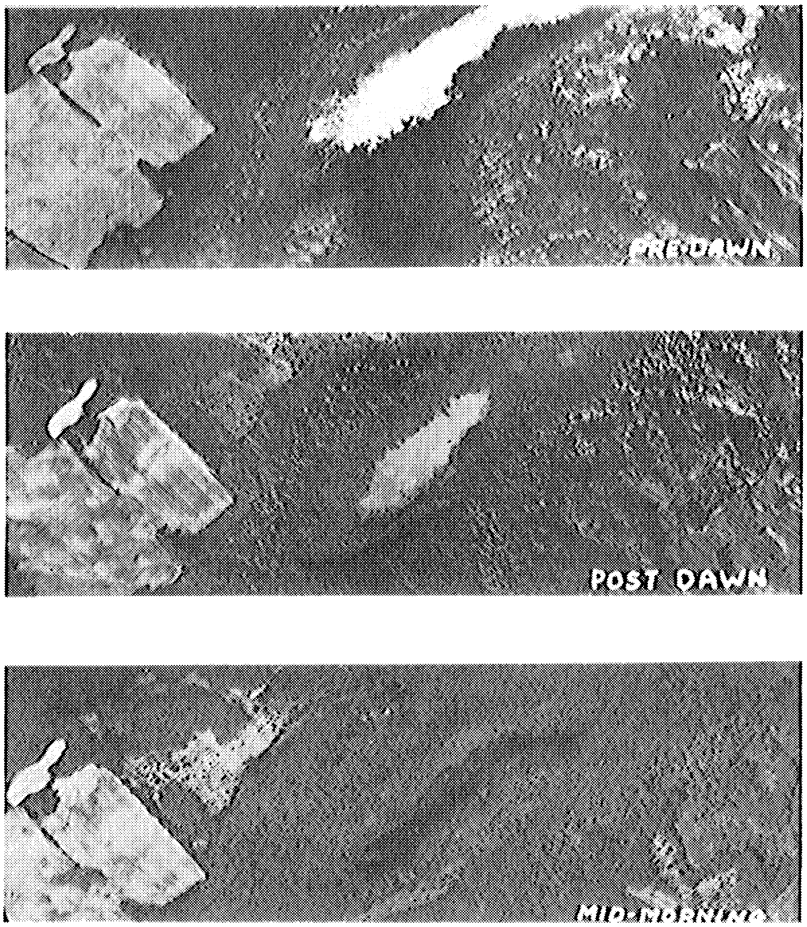


FIGURE 2-15. DAWN TRANSITION IN THERMAL RADIANCE LEVELS OF WATER AND VEGETATIONAL ENVIRONS

Condition	Band 1	Band 2	Band 3	d
1	0	0	0	0
2	0	0	1	1
3	0	1	0	1
4	0	1	1	2
5	1	0	0	1
6	1	0	1	2
7	1	1	0	2
8	1	1	1	3

FIGURE 2-16. COMPARATIVE INFORMATION CAPACITIES, SINGLE VS. MULTIBAND SENSOR

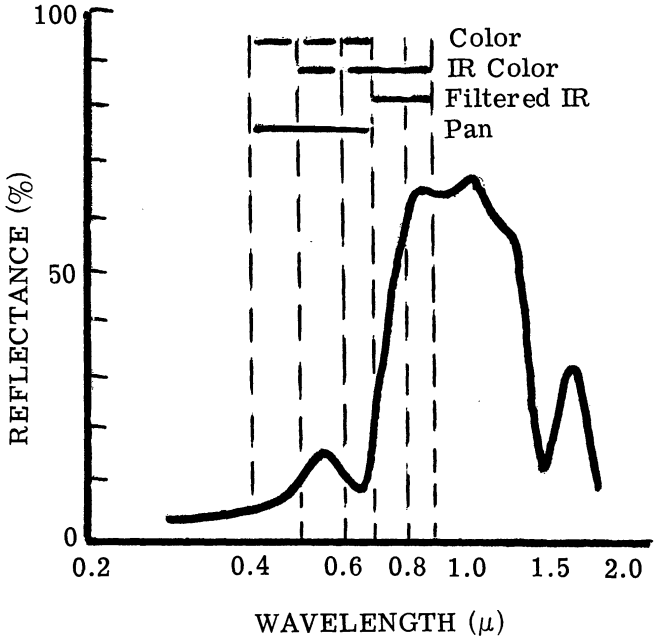


FIGURE 2-17. TYPICAL REFLECTANCE CURVE OF GREEN VEGETATION AND SPECTRAL RESPONSE OF VARIOUS FILMS

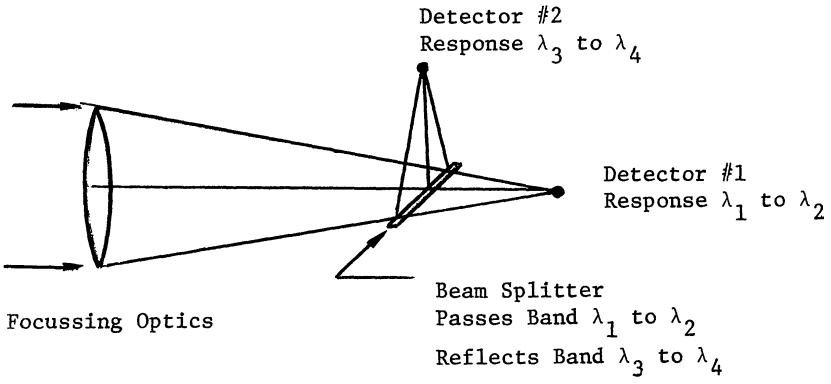


FIGURE 2-18. BEAM SPLITTER FOR MULTIBAND SENSING

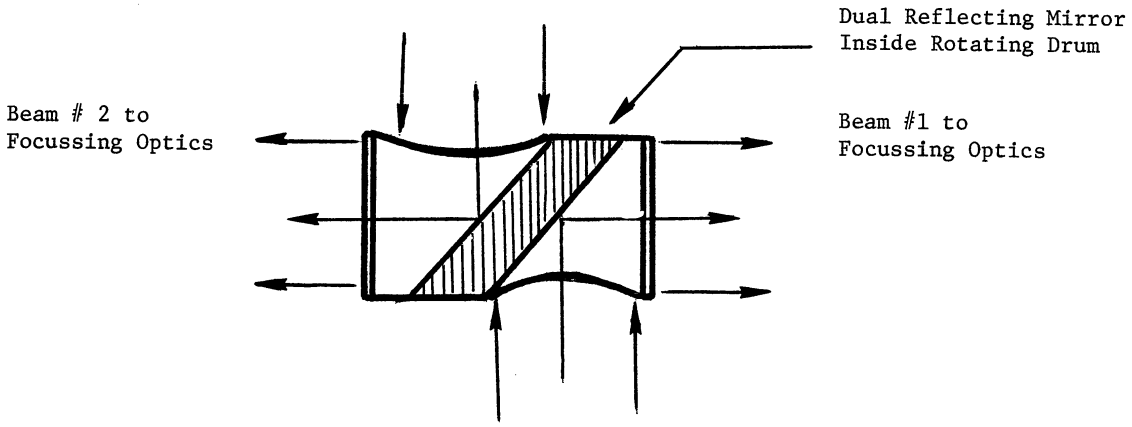


FIGURE 2-19. DUAL SCANNING "DRUM" FOR MULTIBAND OPERATION. As the drum rotates, beams 1 and 2 alternately "look" down and scan the terrain below. Calibration is accomplished in each channel during the other half cycle.

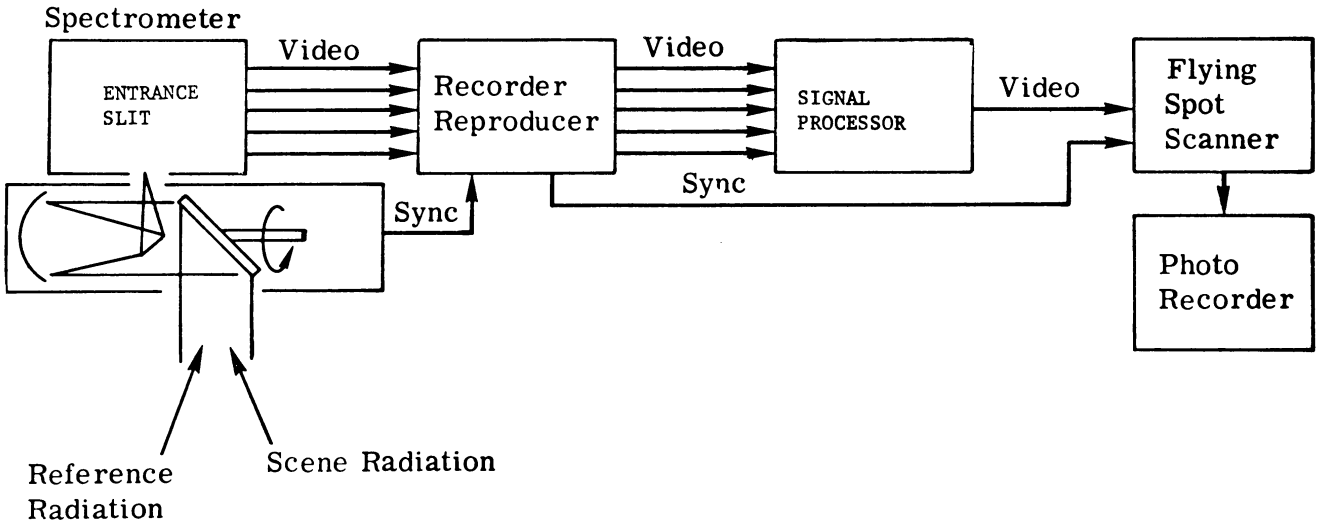


FIGURE 2-20. SCHEMATIC OF MULTISPECTRAL SCANNER AND DATA PROCESSOR

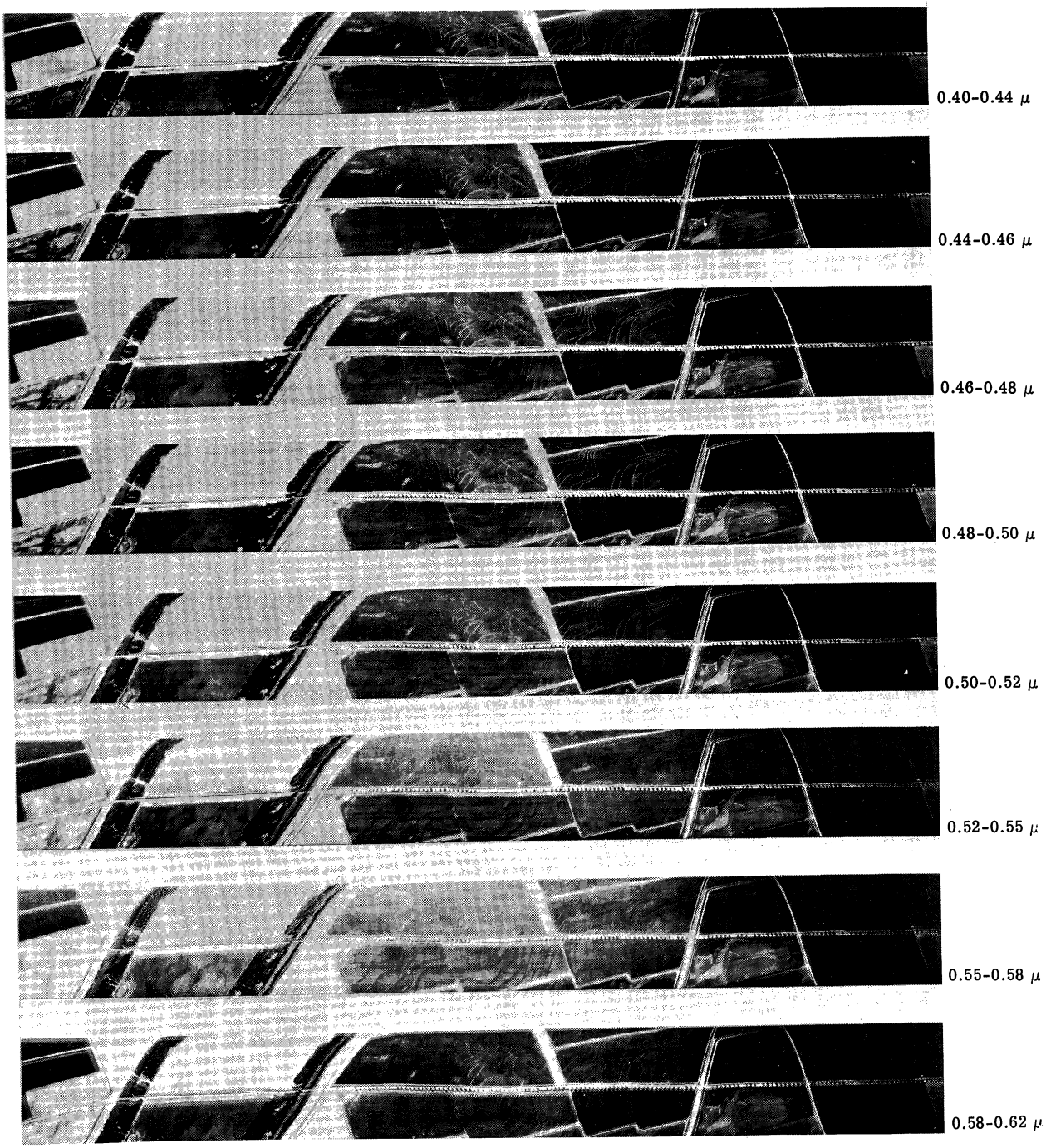


FIGURE 2-21. MULTIBAND IMAGERY OF RICE AND SAFFLOWER FIELDS

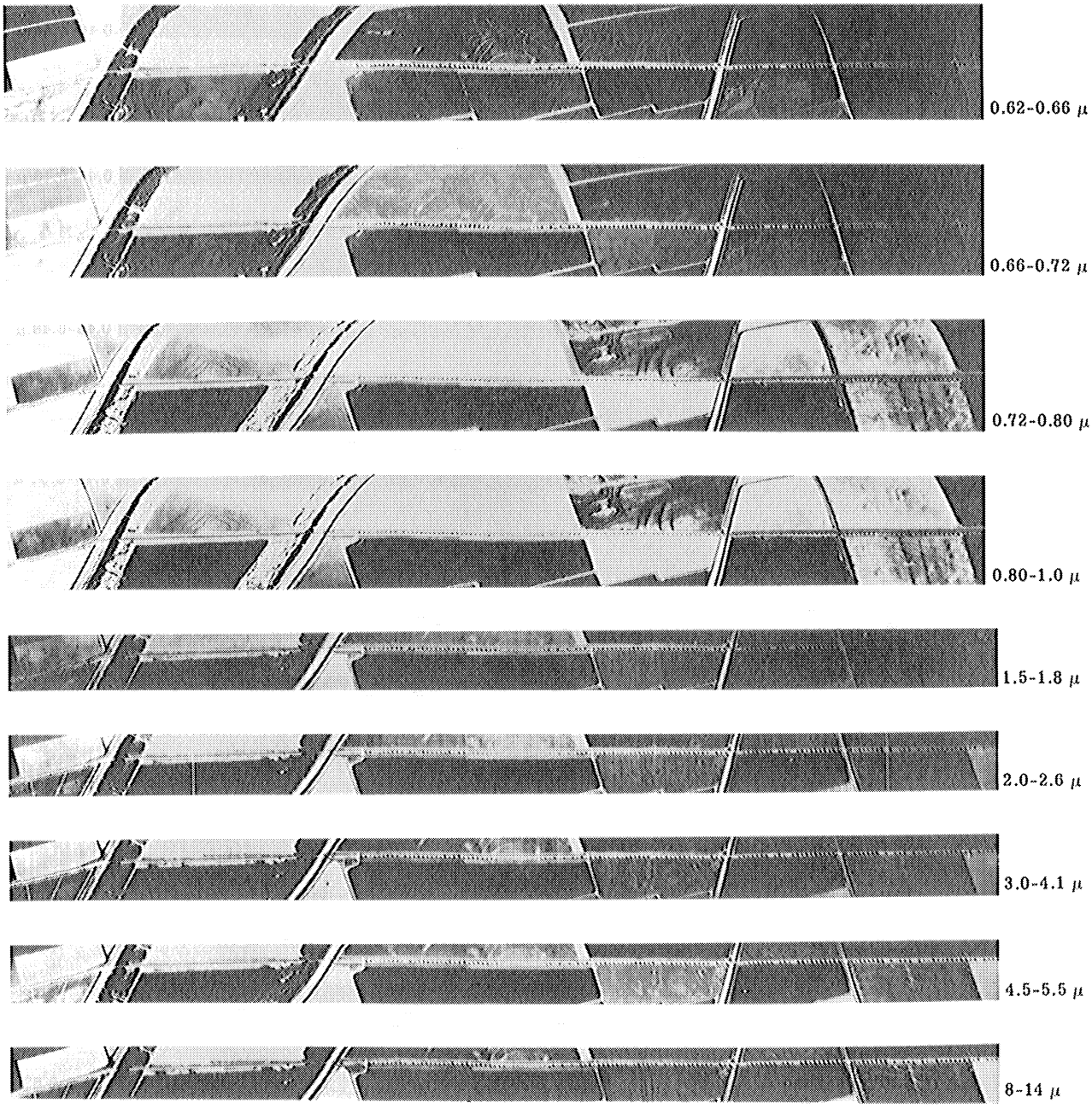


FIGURE 2-21. MULTIBAND IMAGERY OF RICE AND SAFFLOWER FIELDS (Continued)

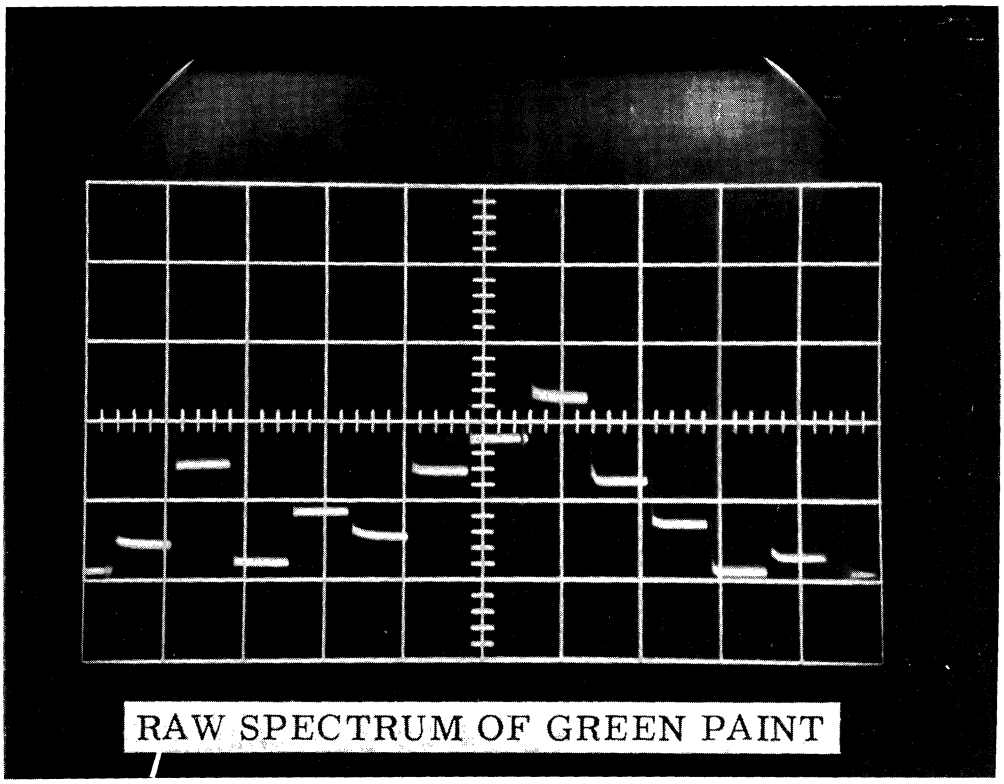


FIGURE 2-22. RAW SPECTRUM OF GREEN PAINT AS OBTAINED WITH A 12-CHANNEL AIRBORNE SYSTEM

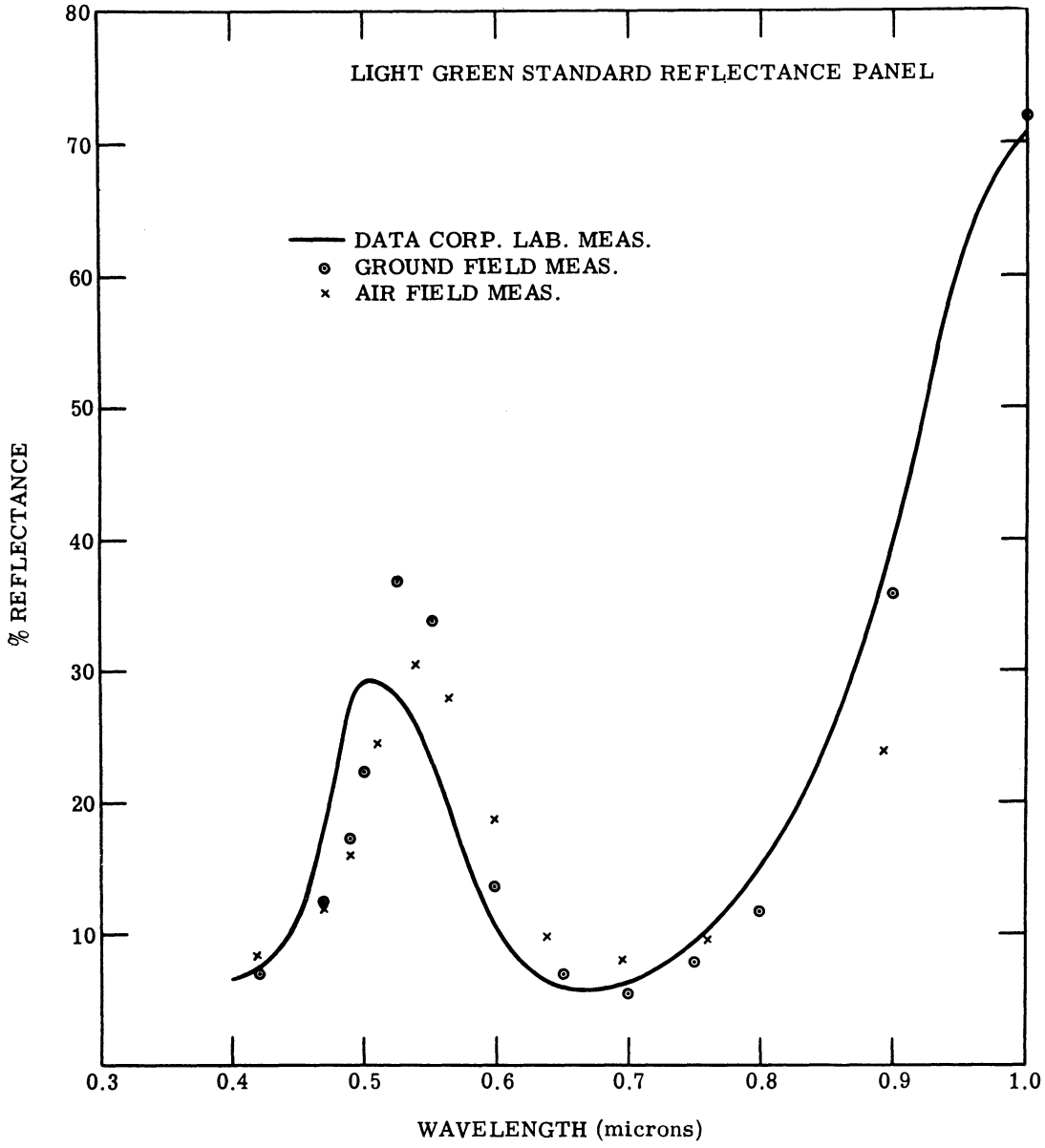


FIGURE 2-23. PLOT OF REFLECTANCE AS A FUNCTION OF WAVELENGTH OBTAINED WITH LABORATORY, FIELD, AND AIRBORNE MEASUREMENTS

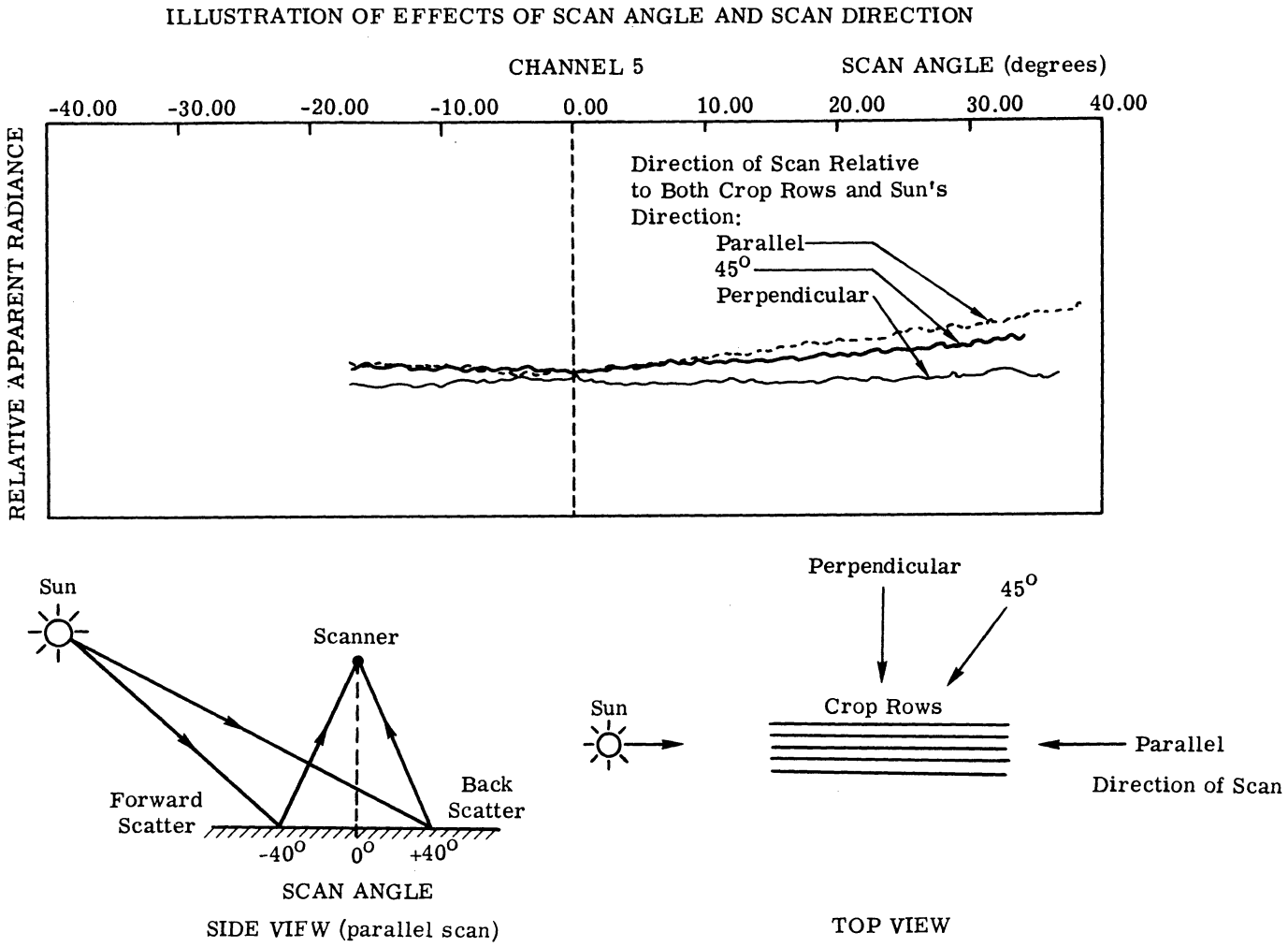
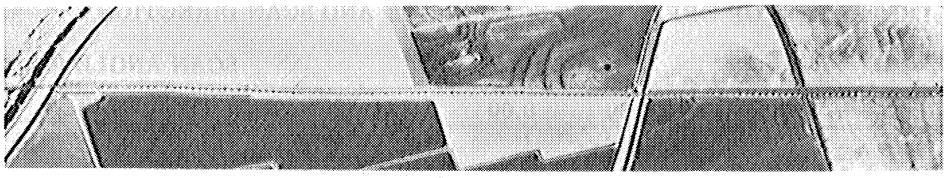
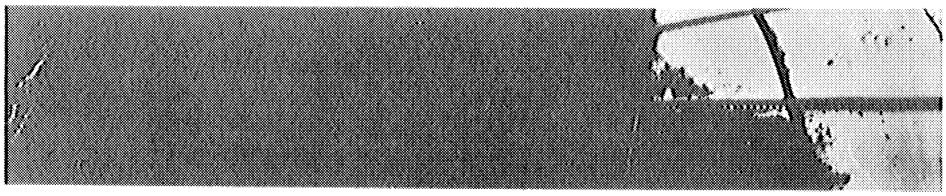


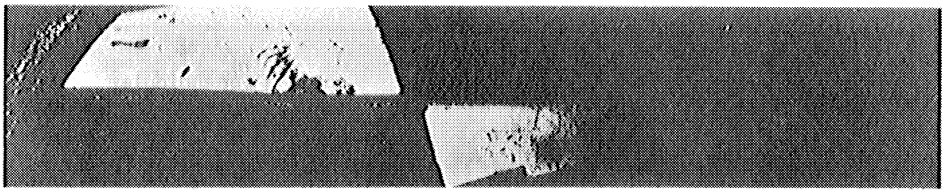
FIGURE 2-24. ILLUSTRATION OF EFFECTS OF SCAN ANGLE AND SCAN DIRECTION



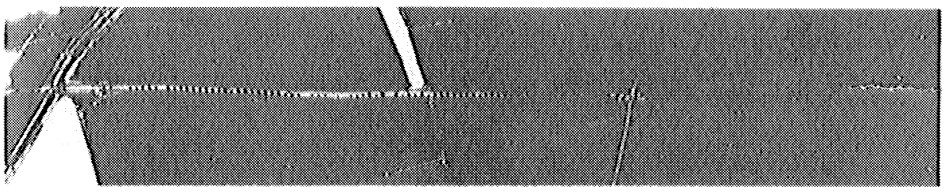
Video Data, 0.80 to 1.0 μ



Young Rice Recognition



Safflower Recognition



Bare Soil Recognition

FIGURE 2-25. RECOGNITION OF FARM CROPS IN MULTISPECTRAL IMAGERY. Recognition pictures and video data of Davis, Calif., agricultural area May 26, 1966, at 1400 hrs., 2000 ft.

SENSOR-USER CHAIN

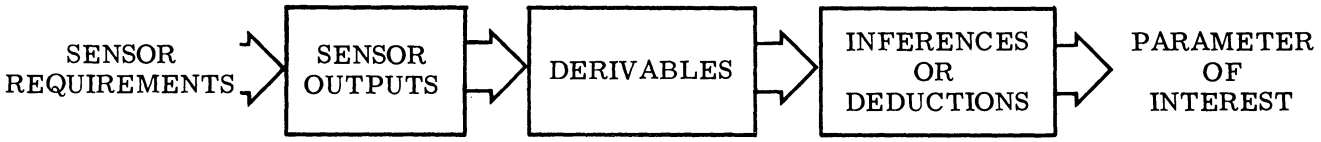


FIGURE 3-1. SENSOR-USER CHAIN

EXAMPLE

Parameter of Interest

Plant Productivity

i.e.,
1. Acreage of Species
2. State of Maturity
3. Losses Due to Disease, Moisture Deficiency
4. Areas of Poor Soil Conditions
5. Knowledge of Available Energy for Growth
6. Estimate of Total Yield When Given Average Yield Per Acre

FIGURE 3-2a. INFORMATION CONCERNING PLANT PRODUCTIVITY, DERIVABLE FROM THE MULTISPECTRAL SENSOR

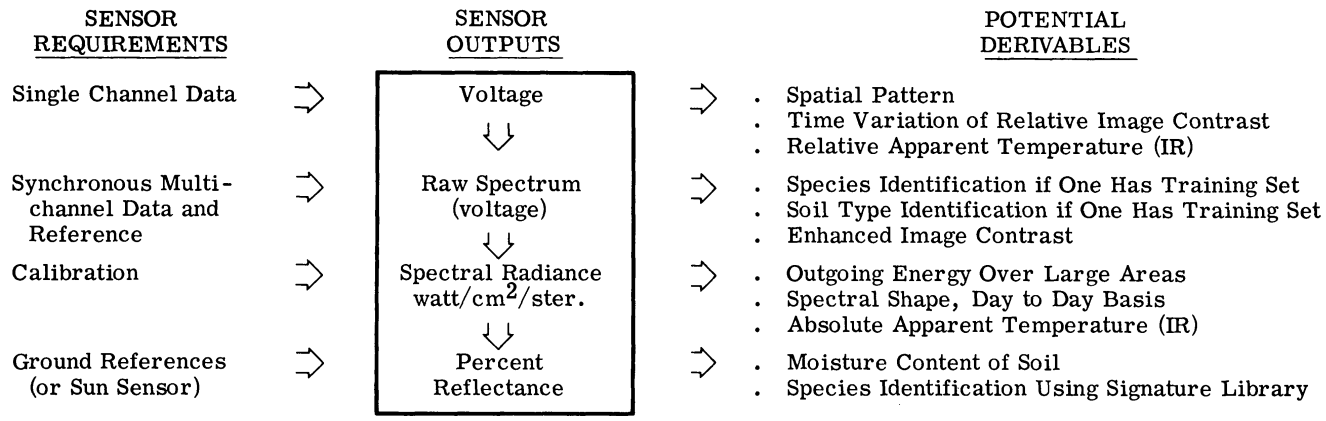


FIGURE 3-2b. INFORMATION CONCERNING PLANT PRODUCTIVITY, DERIVABLE FROM THE MULTISPECTRAL SENSOR

POTENTIAL DERIVABLES	POSSIBLE INFERENCES OR DEDUCTIONS
. Spatial Pattern	. Plant Community Structure and Stratification
. Time Variation of Relative Image Contrast	. Maturation or Wilting
. Relative Apparent Temperature (IR)	. Effects Due to Change in Energy Balance
. Species Identification if One Has Training Set	. Distribution of Plant Species over a Site
. Soil Type Identification if One Has Training Set	. Probable Soil Fertility Levels
. Enhanced Image Contrast	. Change in Species: Maturity, Disease, Nutrient Status, or Moisture Status
. Outgoing Energy over Large Areas	. Energy Budget Calculations
. Spectral Shape, Day to Day Basis	. Distribution of Plant Species or Soil Types, Day to Day Basis
. Absolute Apparent Temperature (IR)	. Quantitative Knowledge of Thermal Processes
. Moisture Content of Soil	. Possible Moisture Availability for Plant Growth
. Species Identification Using Signature Library	. Distribution of Plant Species or Soil Types over a Region with Minimum Ground Truth

FIGURE 3-2c. INFORMATION CONCERNING PLANT PRODUCTIVITY, DERIVABLE FROM THE MULTISPECTRAL SENSOR

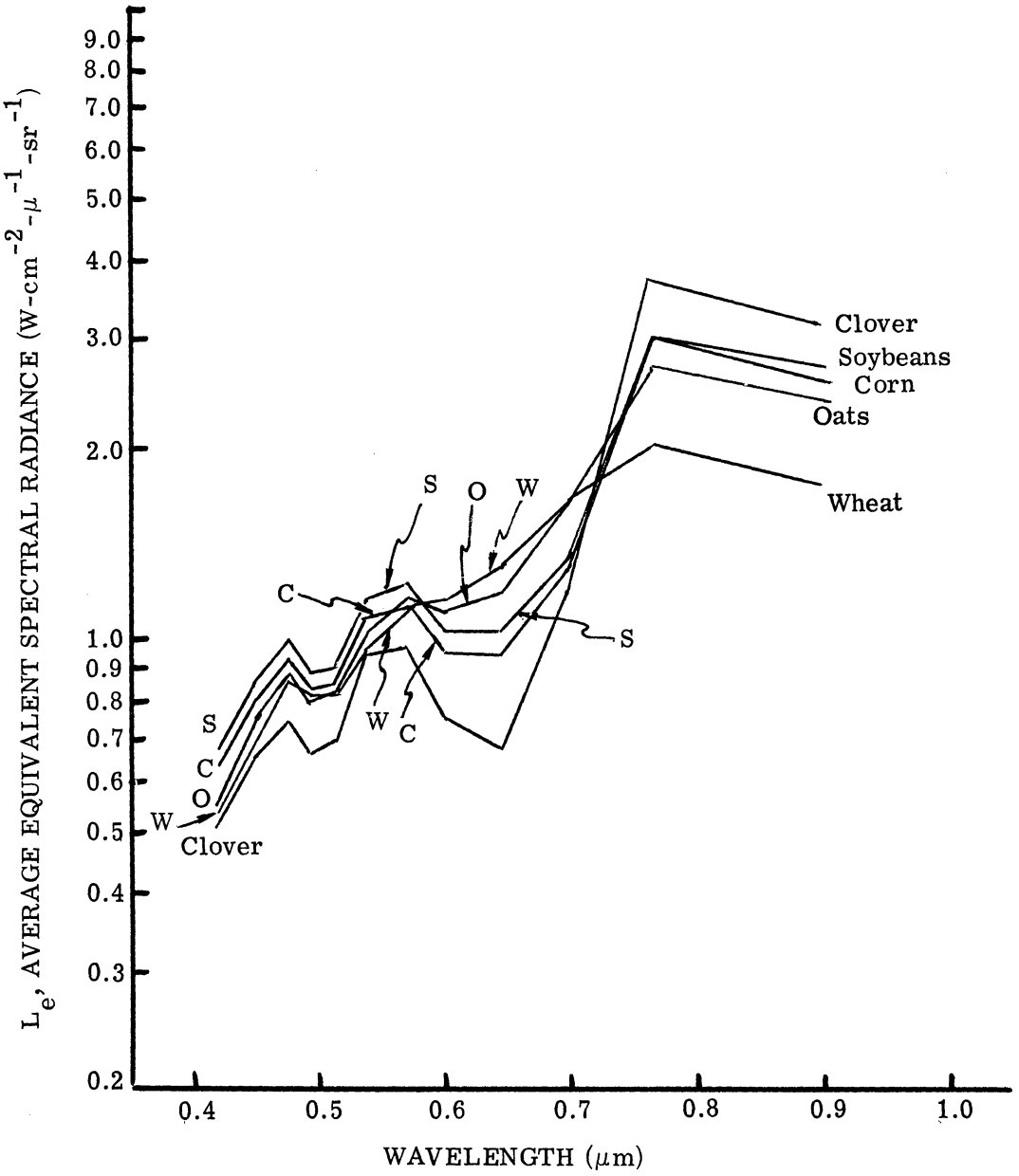


FIGURE 3-3. COMBINED SIGNATURES FOR FIVE CROPS. Data acquired during 1400 hrs. Flight, 6/30/66.

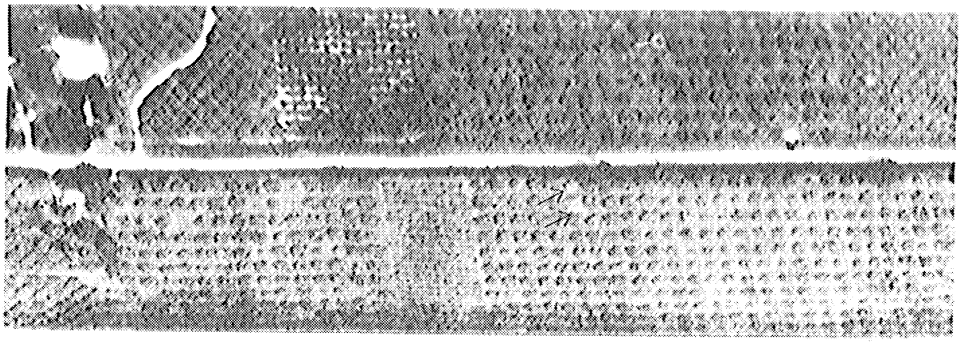
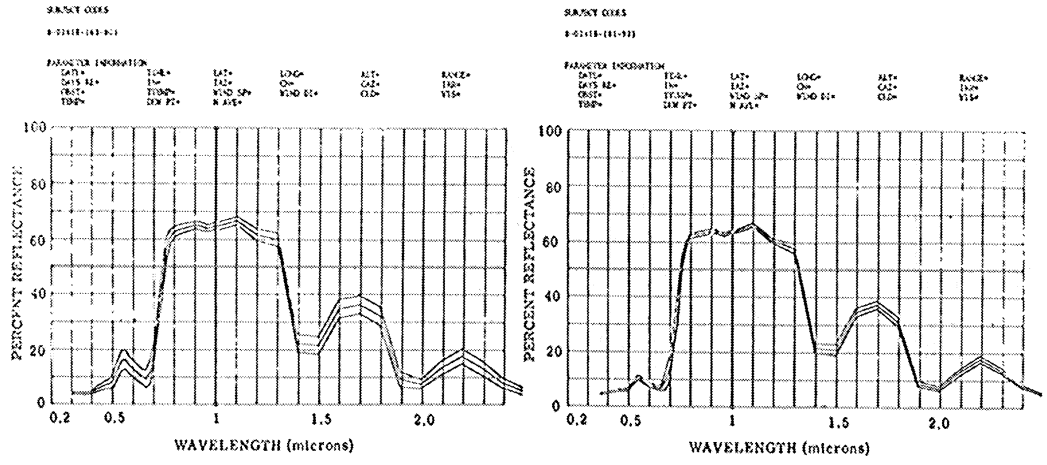


FIGURE 3-4. ULTRAVIOLET IMAGERY SHOWING EFFECT OF BARK-TO-LEAF RATIO IN APPLE TREES. 8/16/63, 1400 hrs., 500 ft.



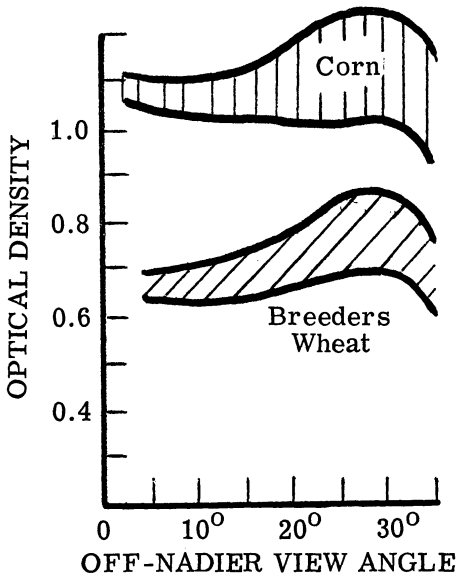
(a) Green Leaves from Top of Plant

(b) Green Leaves from Center and Base of Plant

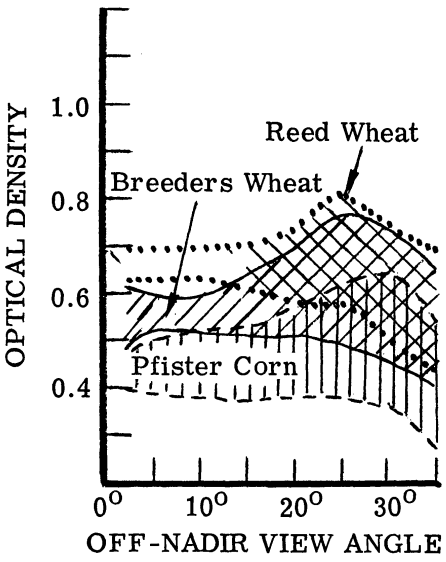
(c) Dry Brown Leaves

(d) Dry Tassle

FIGURE 3-5. SPECTRAL REFLECTANCE CURVES FOR CORN



(a) Infrared Film (0.7 to 0.9 μ)



(b) Visible Film (0.4 to 0.7 μ)

FIGURE 3-6. COMPOSITE PLOTS OF DENSITY VS. OFF-NADIR VIEW ANGLE, 1430 HRS

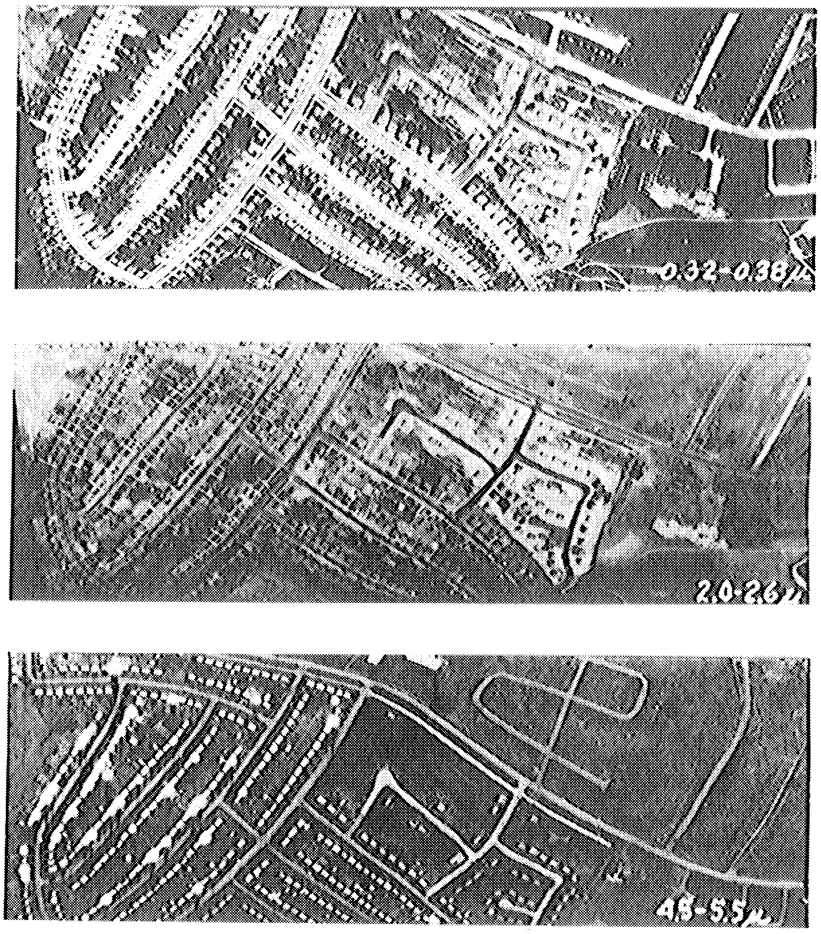


FIGURE 3-7. SIMULTANEOUS MULTISPECTRAL COMPARISON OF RESIDENTIAL AREA, ANN ARBOR, MICH. Note in particular the high contrast of sidewalks and driveways in the ultraviolet channel with respect to grass background. Compare tone changes of same items in IR channels. The reverse is true for tone patterns of roofing materials. The 2.0- to 2.6-micron channel shows small tone differences; while the 4.5- to 5.5-micron channel shows unexplained strong tonal differences between a number of roof tops, as well as good contrast of most roof tops with respect to grass background. Also note the comparative tone patterns of the area under construction (upper middle of each strip). 30 July 1963, about 1100 hrs., scale 1/11800.

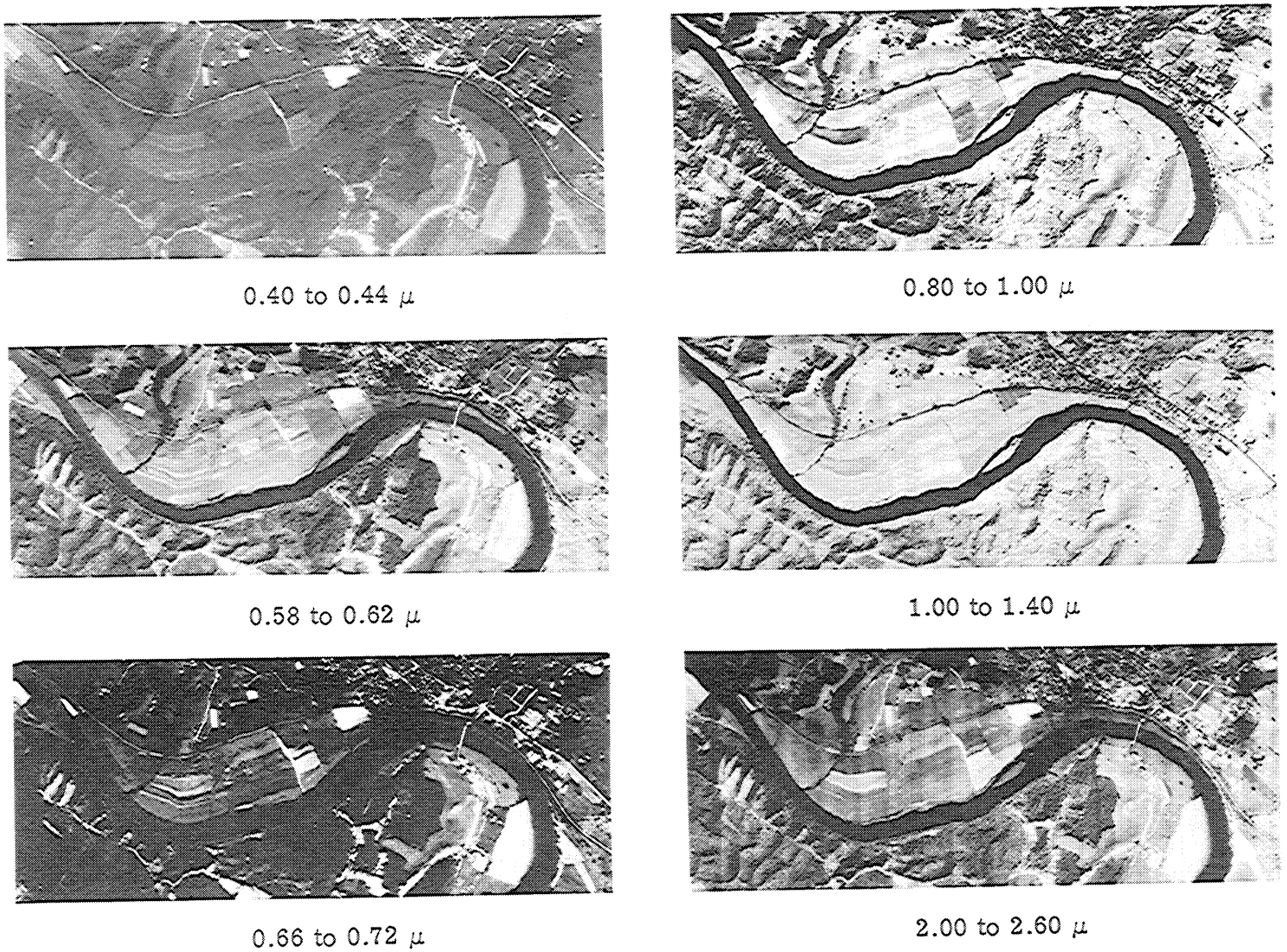


FIGURE 3-8. MULTISPECTRAL IMAGERY OF AGRICULTURAL TERRAIN. A section of the Holston River in six spectral regions between 0.4 and 2.6 microns. Note the variety of tonal patterns that exist in agricultural fields as a function of wavelength. For example, in the violet region, 0.4 to 0.44 microns, only cultural features and bare soils appear lighter in tone, whereas, in the near infrared, 1.00 to 1.40 microns, the most striking feature is the dark tone of the river. Close inspection of the intermediate regions shows a variety of tones for different crop types. This variety of tone suggests the existence of a spectral signature.

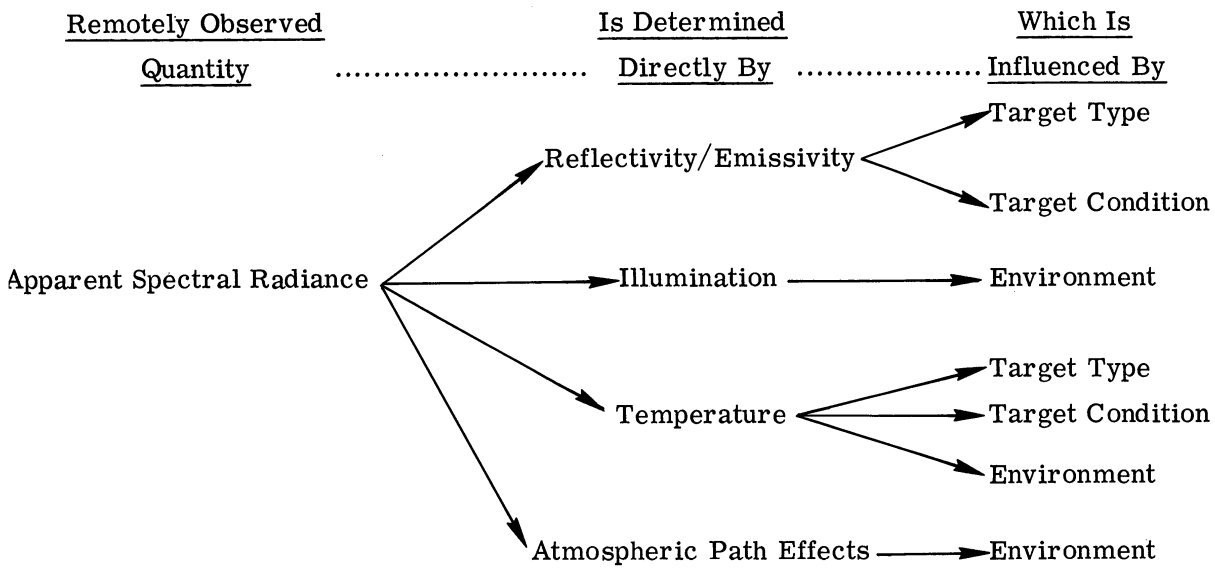


FIGURE 3-9. PARAMETERS AFFECTING APPARENT SPECTRAL RADIANCE

These Parameters	Are Described by Observing or Measuring These Quantities
Target Type	qualitative characteristics chemical composition mechanical composition physiologic structure etc.
Target Condition	physical dimensions density/mass moisture content texture surface contamination thermal conductivity thermal heat capacity geometry/orientation health/vigor maturation etc.
Environment	solar radiation intensity atmospheric thermal radiation intensity cloud cover air temperature wind speed and direction water vapor content of air rainfall atmospheric contaminants (dust, haze) distribution of other objects near target etc.

FIGURE 3-10. MEASURABLES OF TARGET PARAMETERS

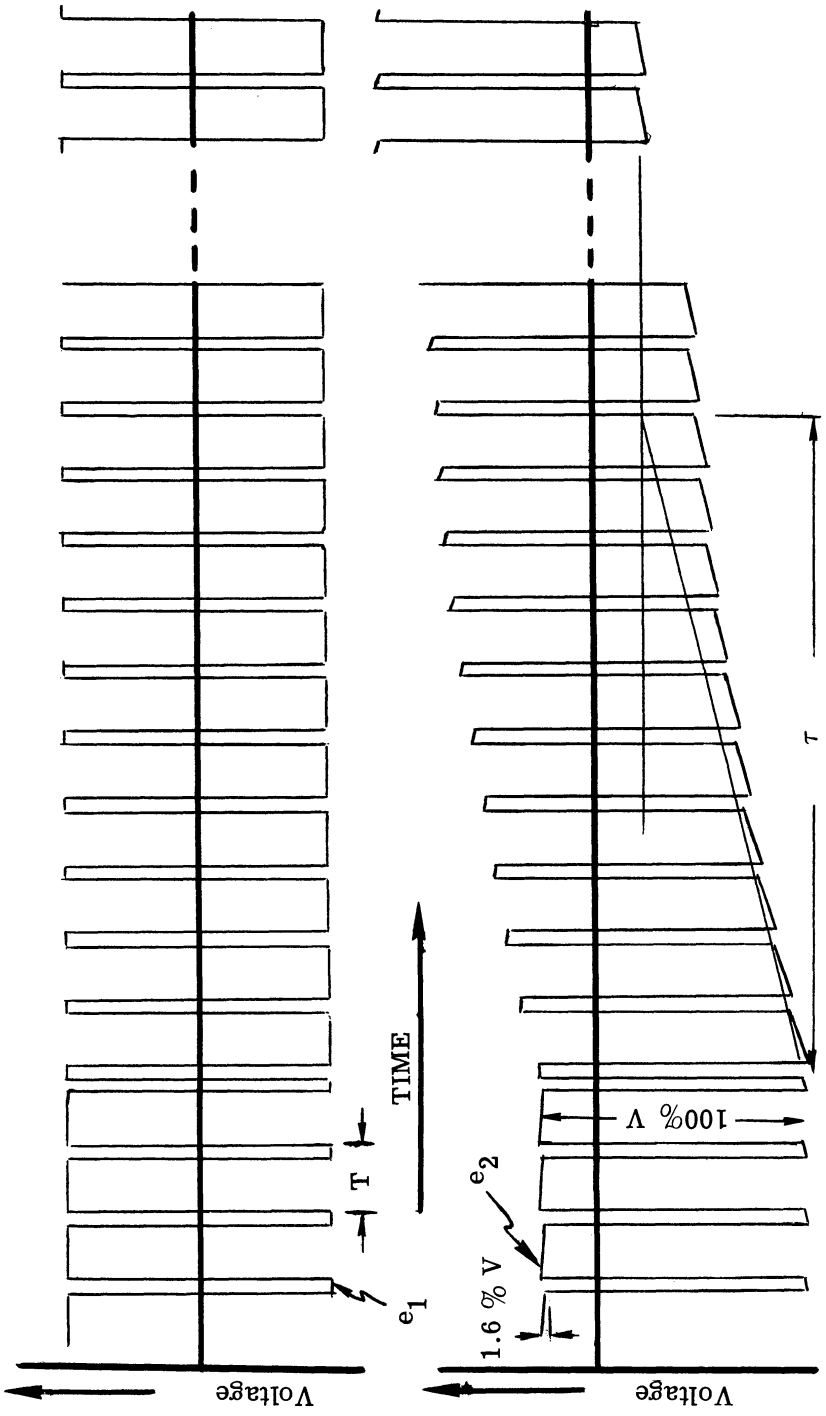


FIGURE A-1. SIGNAL DROOP CAUSED BY AC COUPLING

FIGURE A-2. TYPICAL IMAGERY WITH REFERENCES AND GREY SCALE PRINTOUT. Holston River at Surgoinsville, Tenn., 9/3/67, 1643, altitude 4000 ft., 0.8 to 1.0 micron imagery by University of Michigan, IST, IROS.

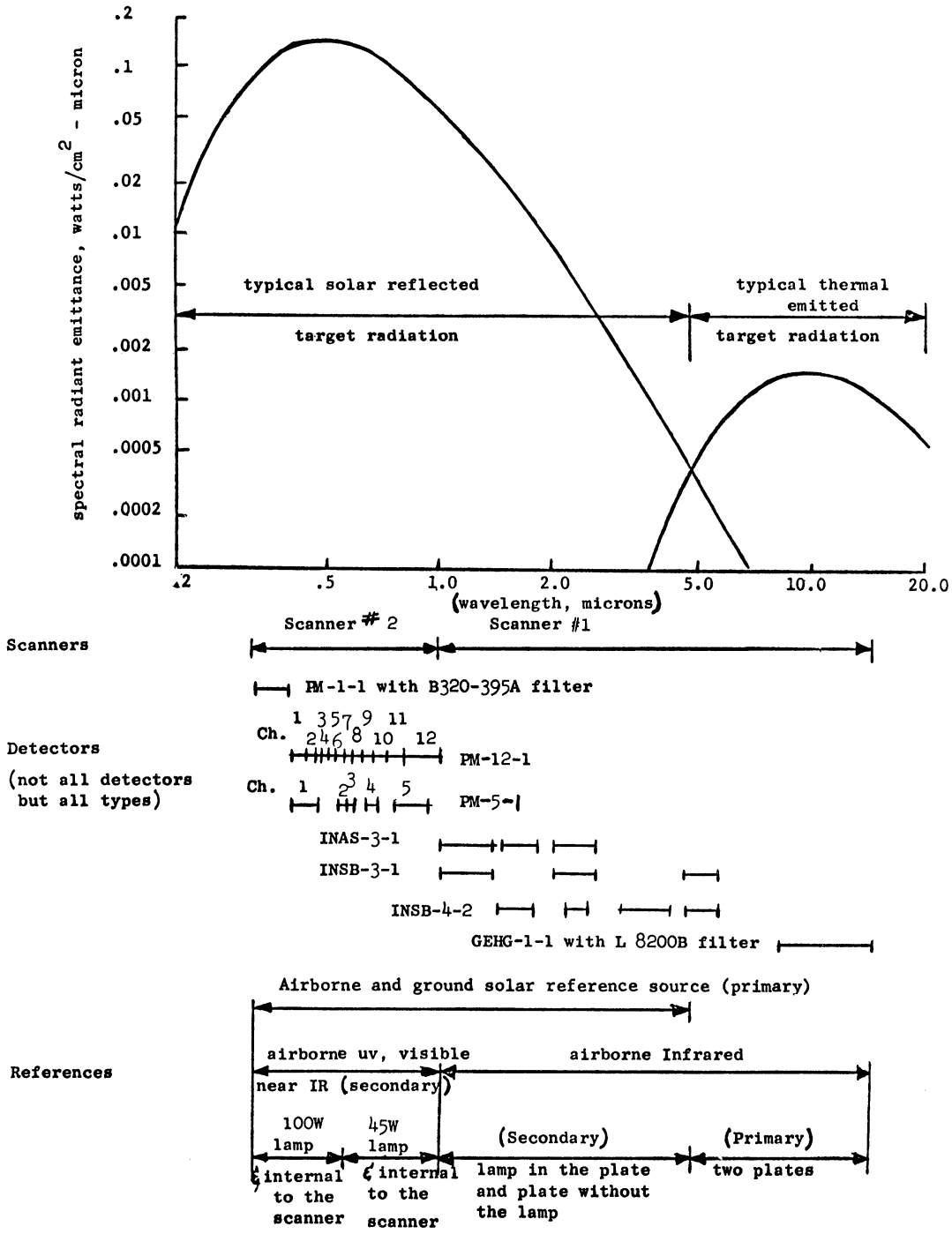


FIGURE A-3. EQUIPMENT SPECTRAL REGIONS

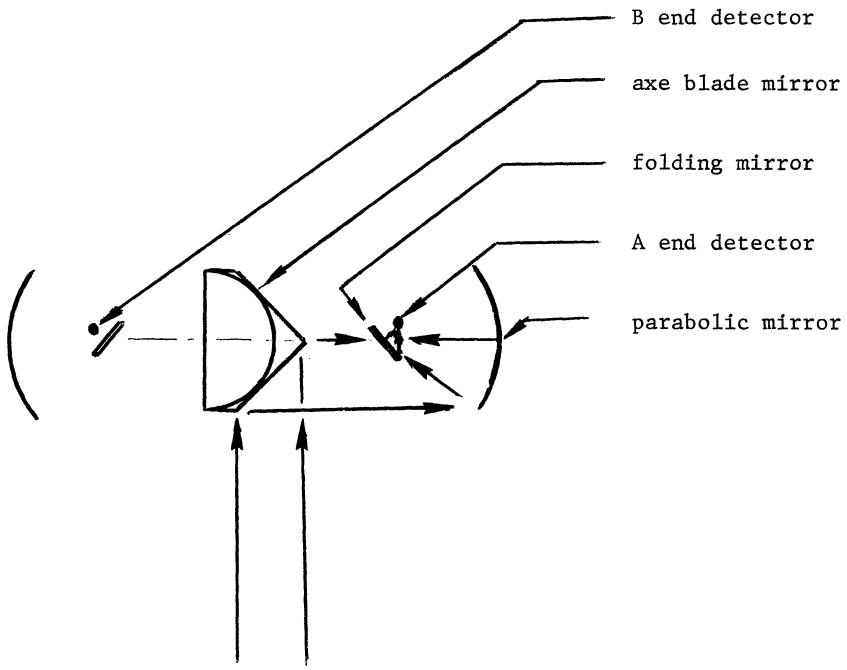


FIGURE A-4. SCANNER OPTICAL CONFIGURATION

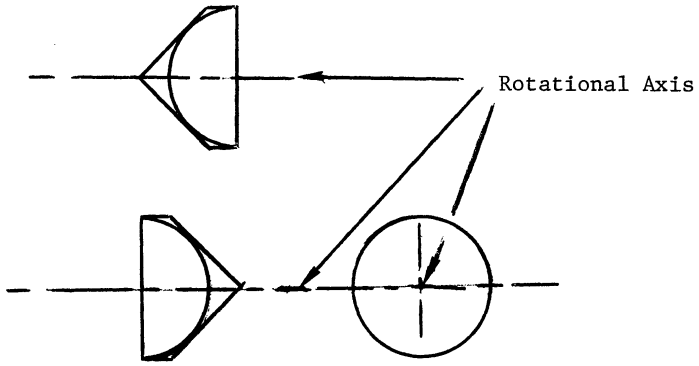


FIGURE A-5. AXE BLADE MIRROR CONFIGURATION

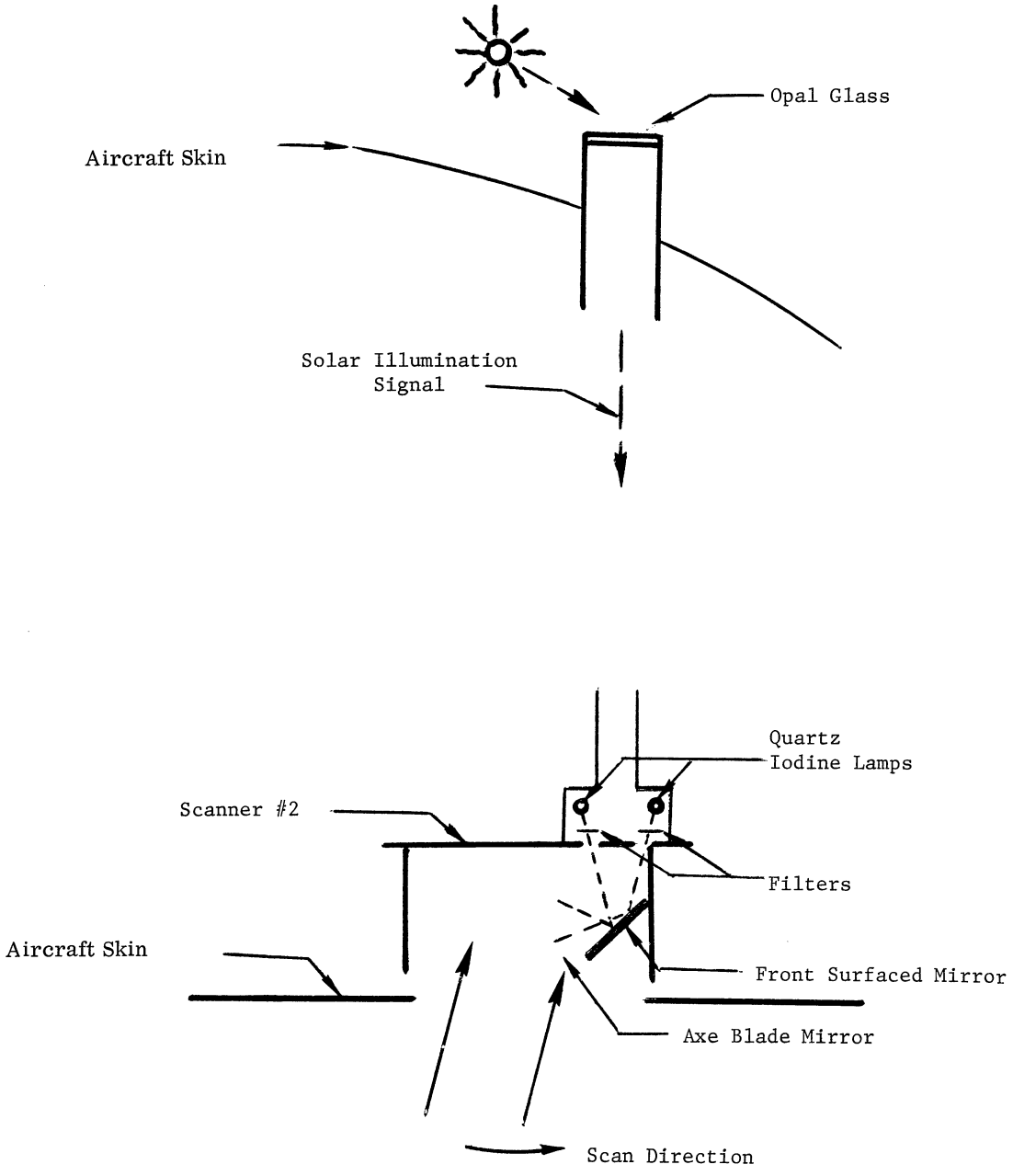


FIGURE A-6a. SCANNER #2 CONFIGURATION OF REFERENCE SOURCES

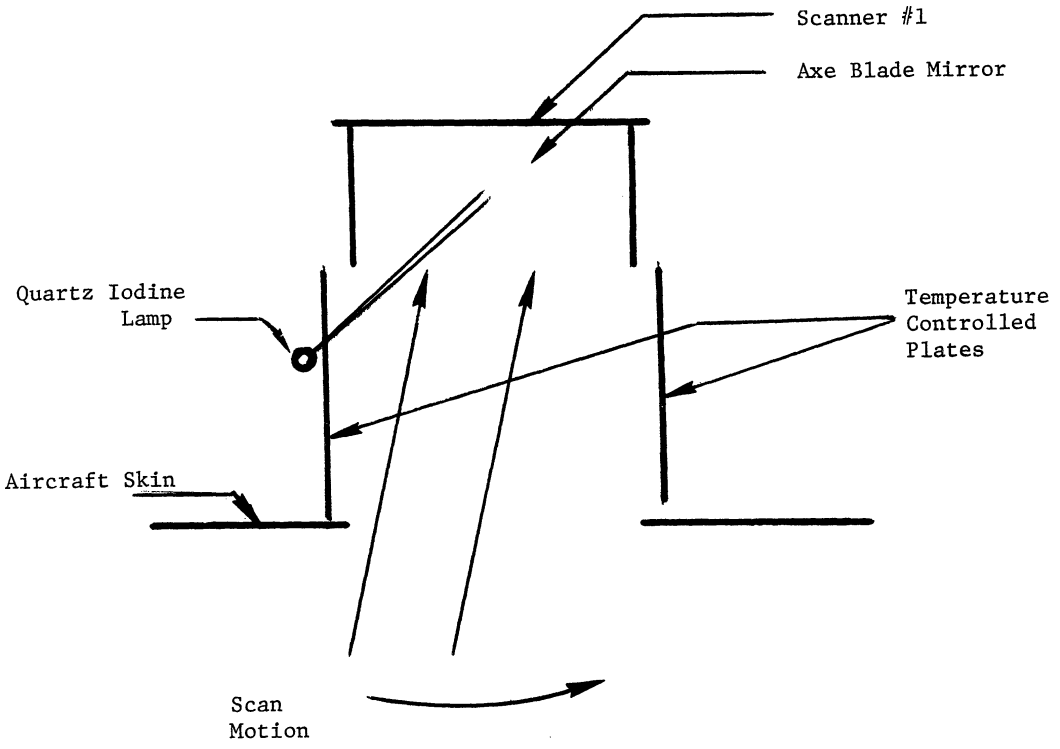


FIGURE A-6b. SCANNER #1 CONFIGURATION OF REFERENCE SOURCES

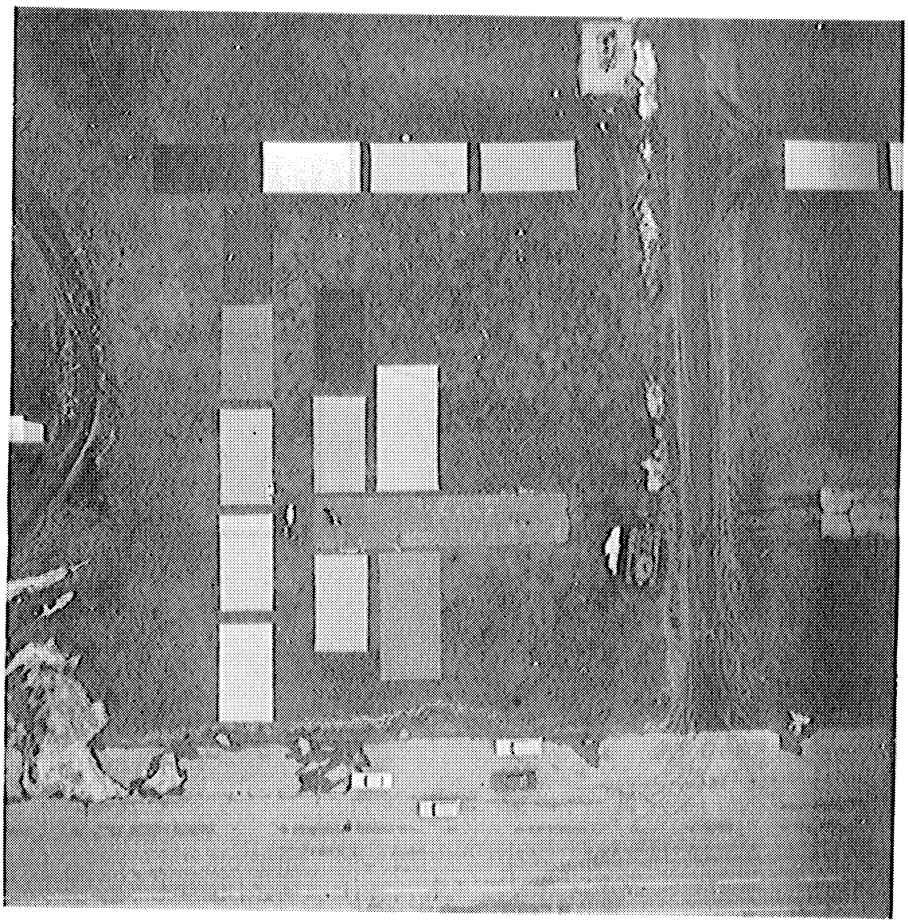


FIGURE A-7. IMAGERY OF REFERENCE PANELS. Panels on Willow Run Airport ramp, 3/15/67, 1156, altitude 500 ft., with Ansco D-200 film using a 1A filter.

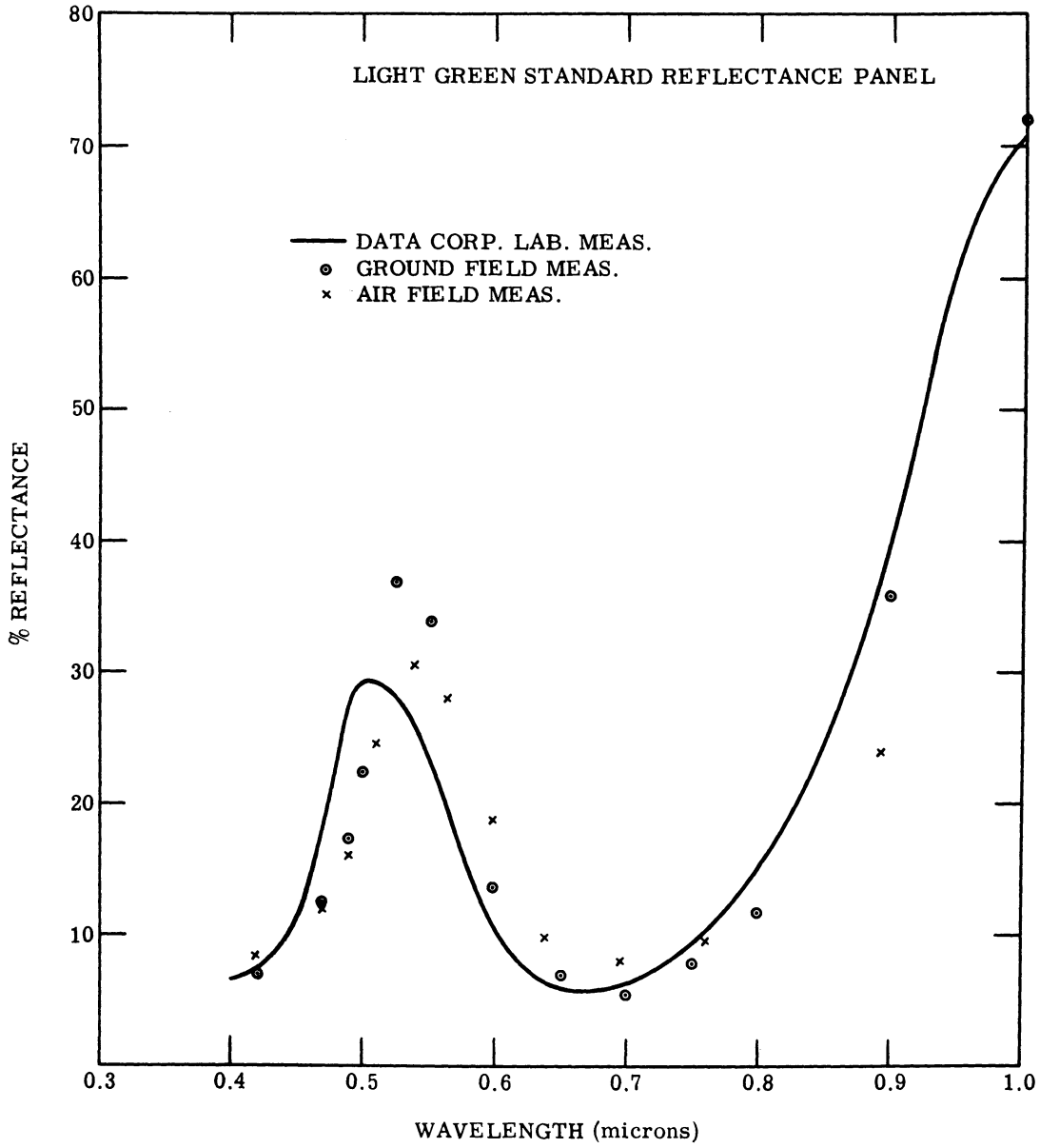


FIGURE A-8. SPECTRAL REFLECTANCE OF REFERENCE PANELS VS. FIELD AND AIRBORNE MEASUREMENT

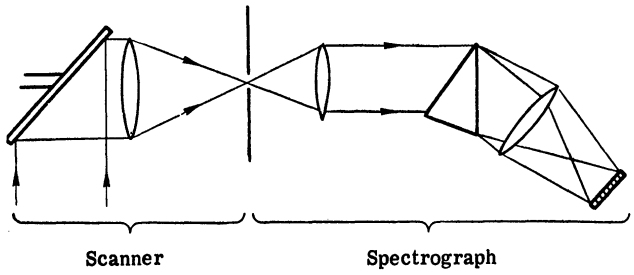


FIGURE C-1. GENERALIZED OPTICAL SYSTEM OF DISPERSIVE SCANNER

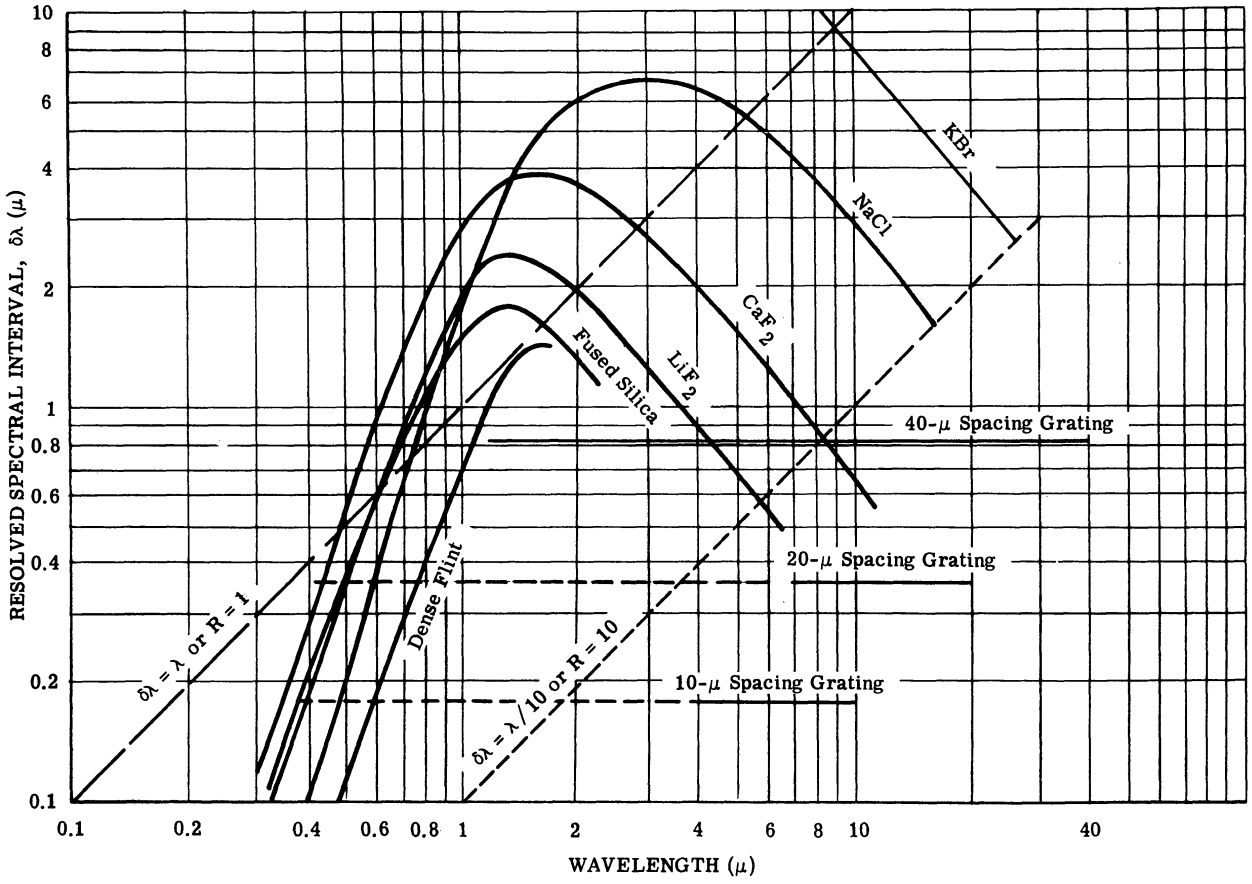


FIGURE C-2. SPECTRAL RESOLUTION FOR VARIOUS DISPERSING ELEMENTS

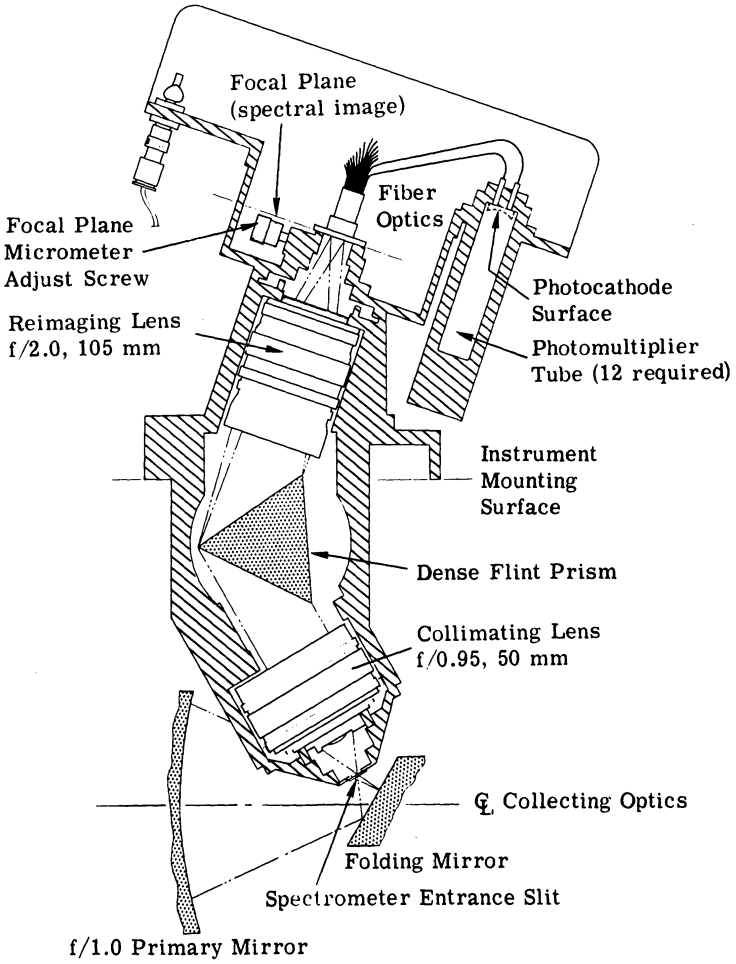


FIGURE C-3. CONCEPTUAL SECTIONED VIEW OF SPECTROMETER DETECTOR SYSTEM

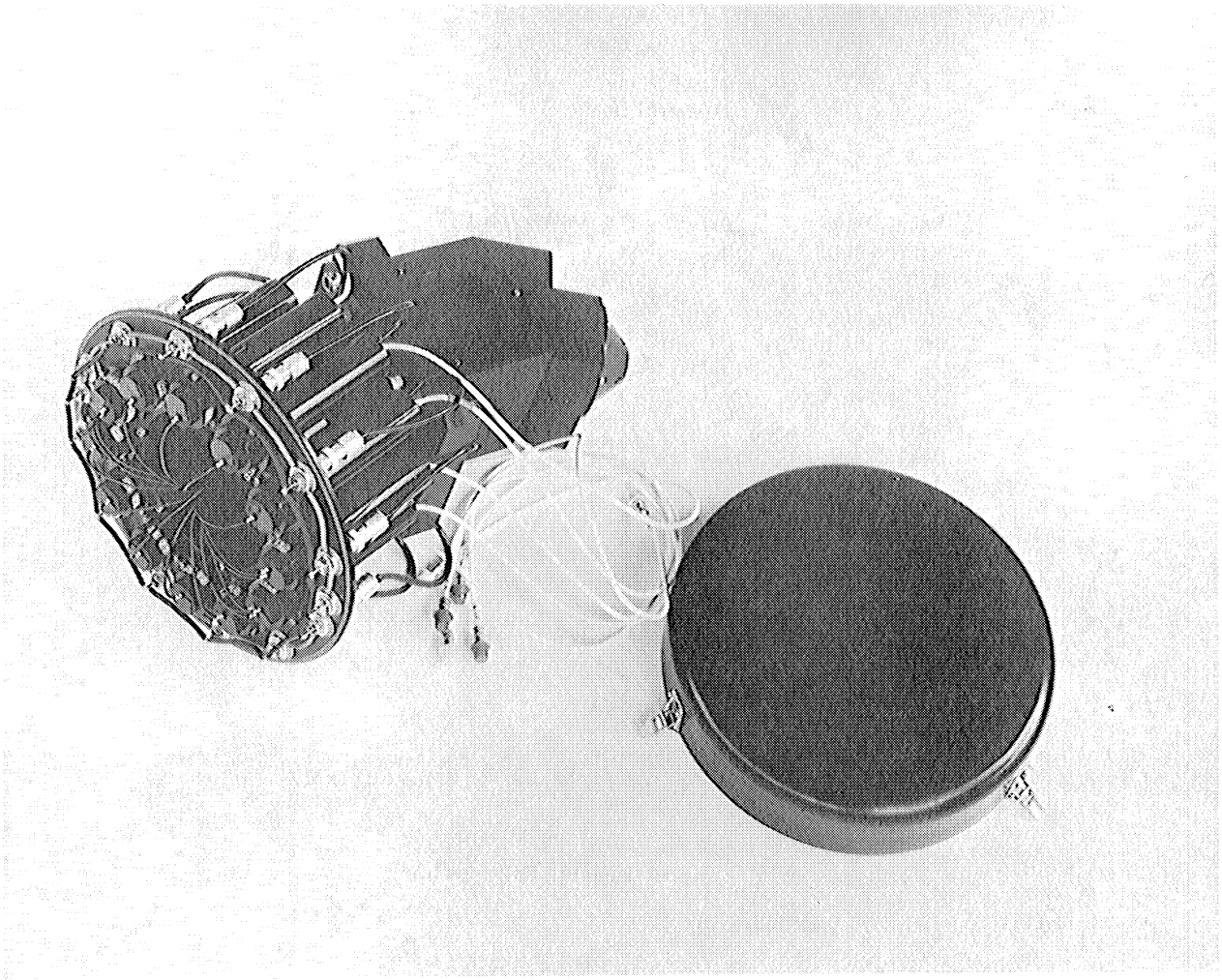


FIGURE C-4. SPECTROMETER DETECTOR SYSTEM WITH ACCESS COVER REMOVED

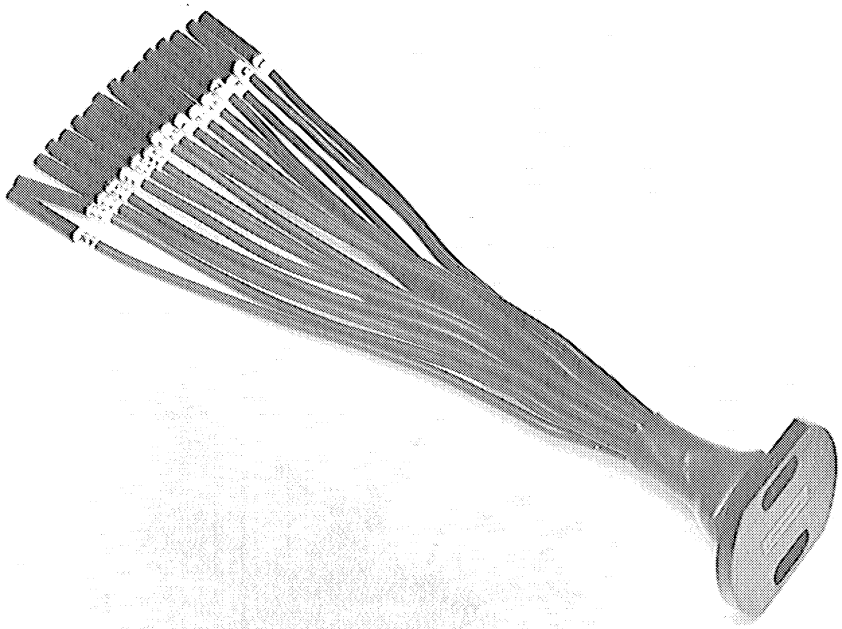


FIGURE C-5. FIBER OPTICS OF SPECTROMETER DETECTOR SYSTEM

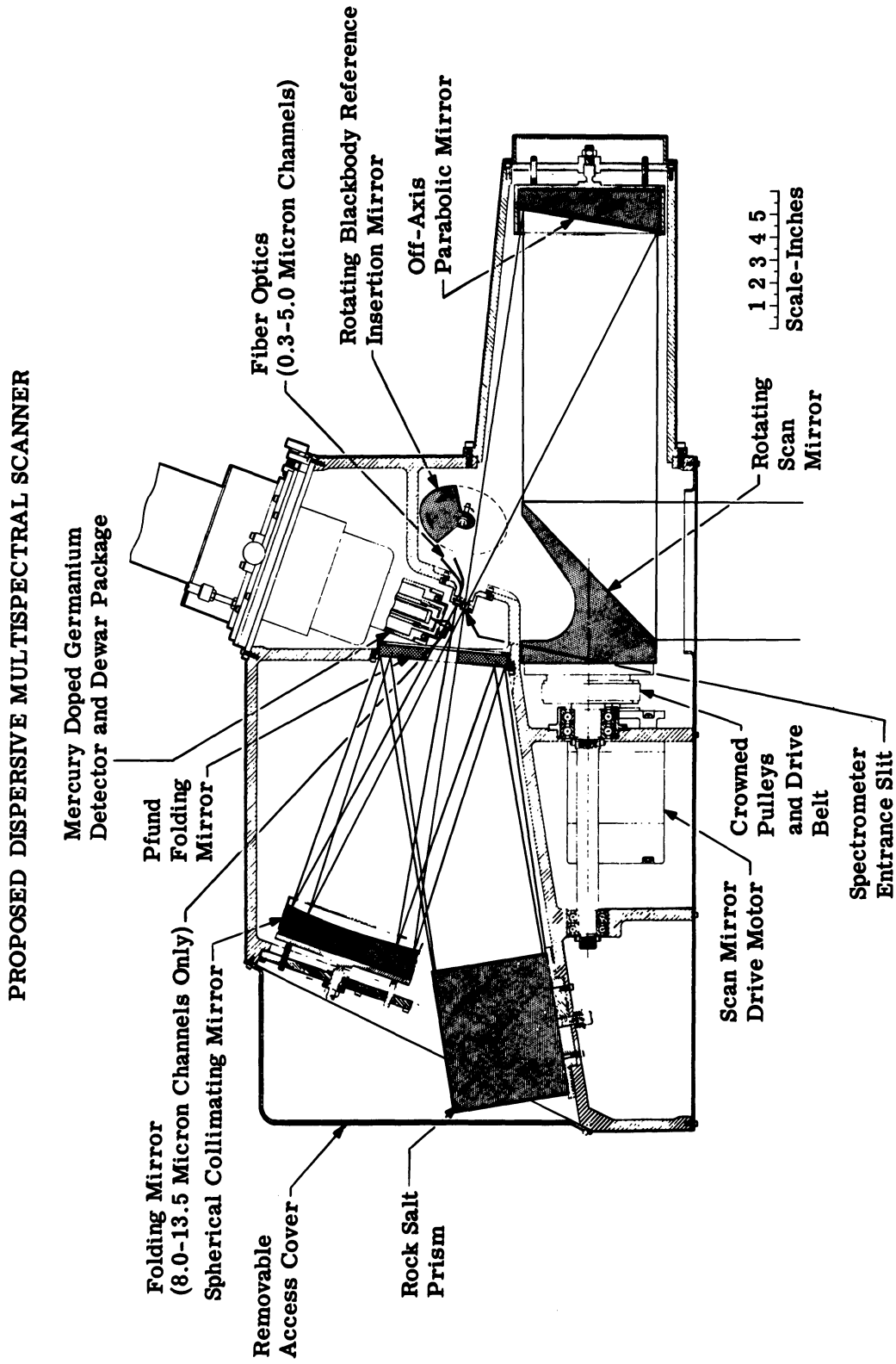


FIGURE C-6. THE SCANNER-SPECTROGRAPH

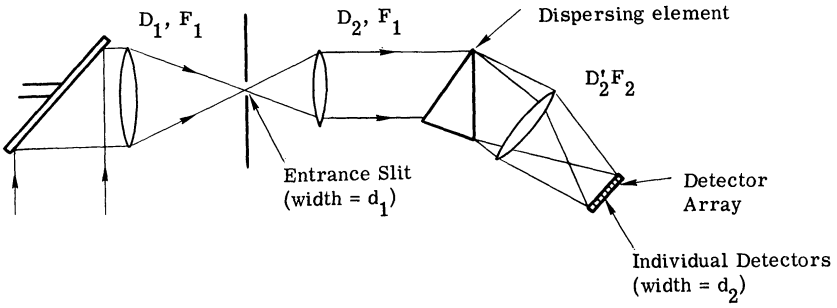


FIGURE C-7. SCHEMATIC OF MULTISPECTRAL DISPERSIVE SCANNER

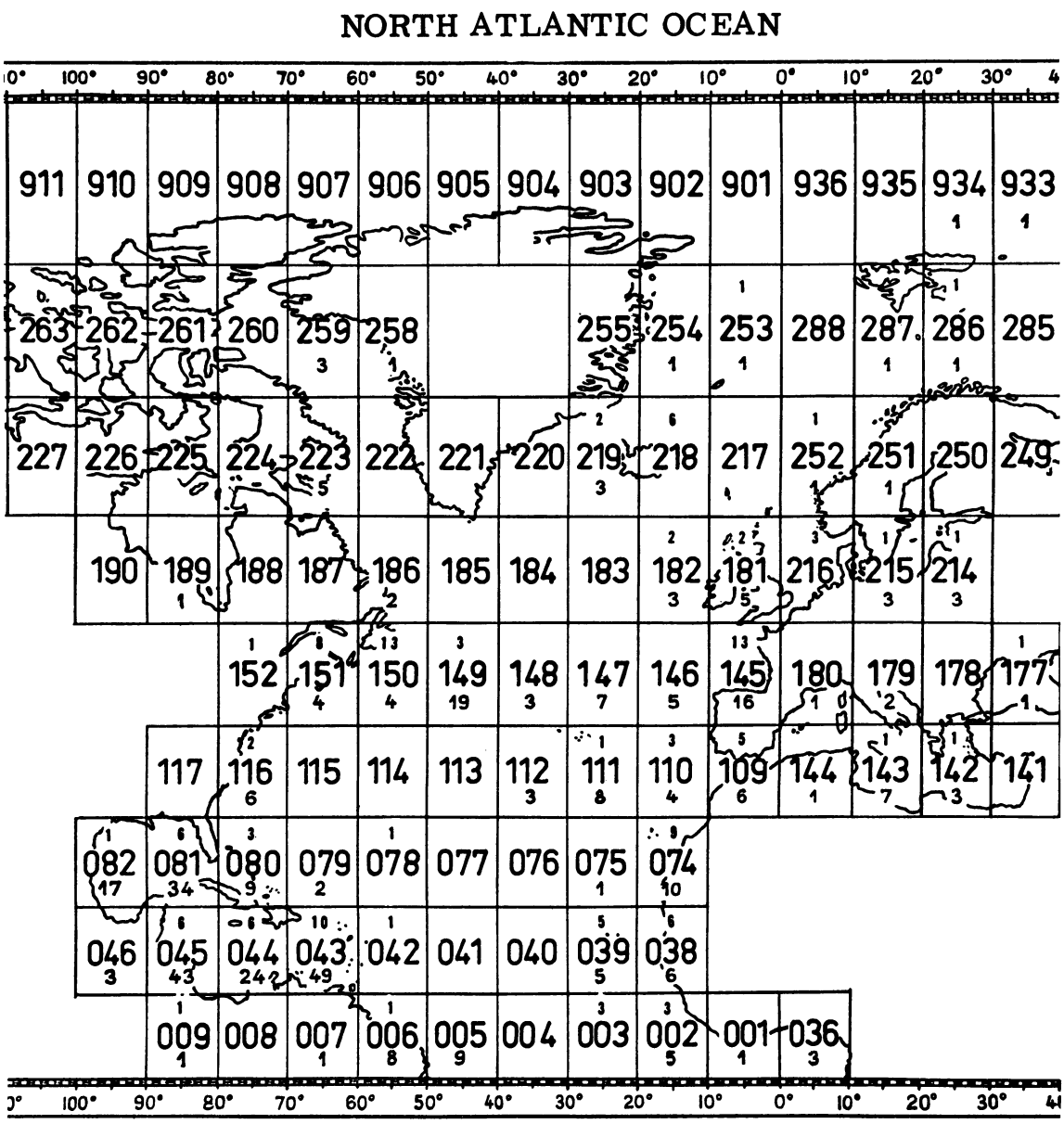
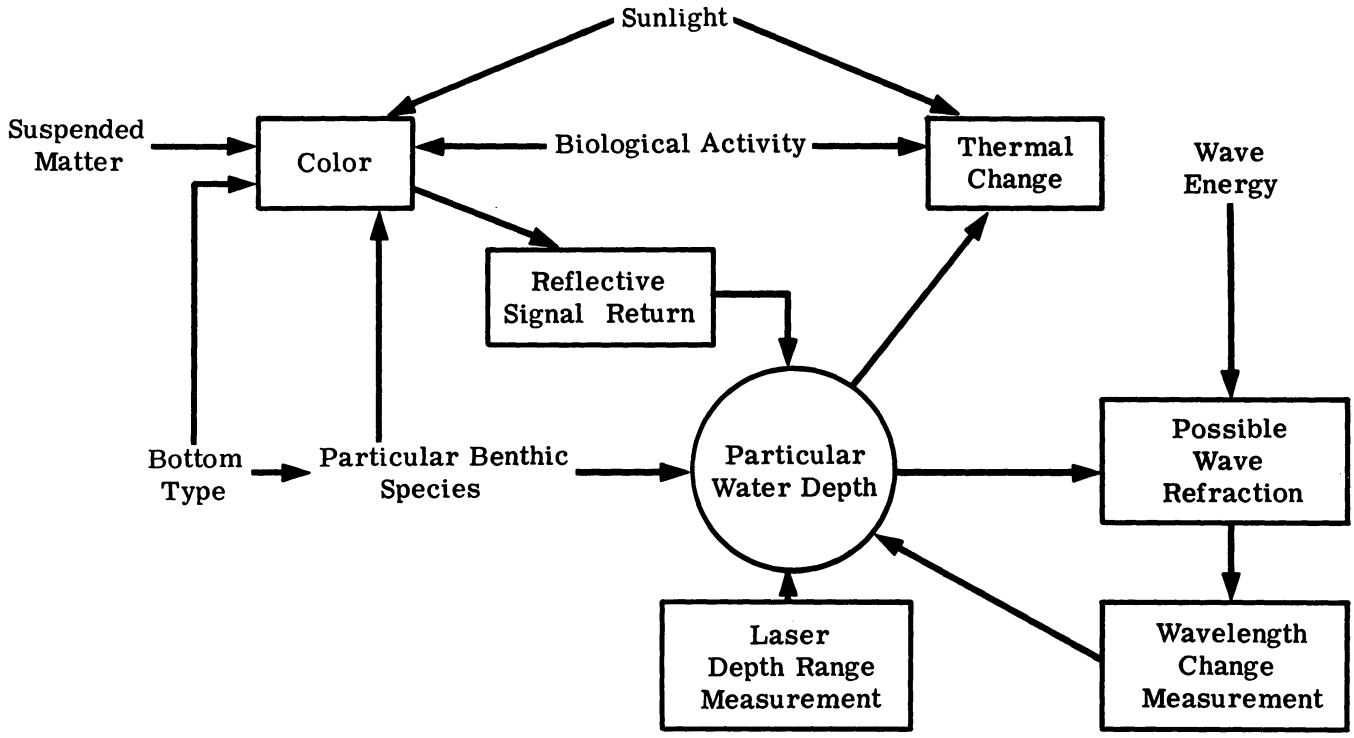


FIGURE D-1. DISTRIBUTION OF DOUBTFUL SOUNDING FOR NORTH ATLANTIC OCEAN



INFERENCE LOGIC CHART

FIGURE D-2. INFERENCE LOGIC CHART

OBSERVABLE	SENSOR
Color/Density	Camera
	Optical Mechanical Scanner (Multispectral)
	TV System
Thermal Anomaly	Optical Mechanical Scanner
	Radiometer
Wave Refraction	Photography: Camera, TV, O.M.S.
	Radar
	Laser Imager
Direct Measurement	Laser Depth Ranger

FIGURE D-3. SENSOR-OBSERVABLE RELATIONS FOR SHALLOW WATER DETECTION CASE

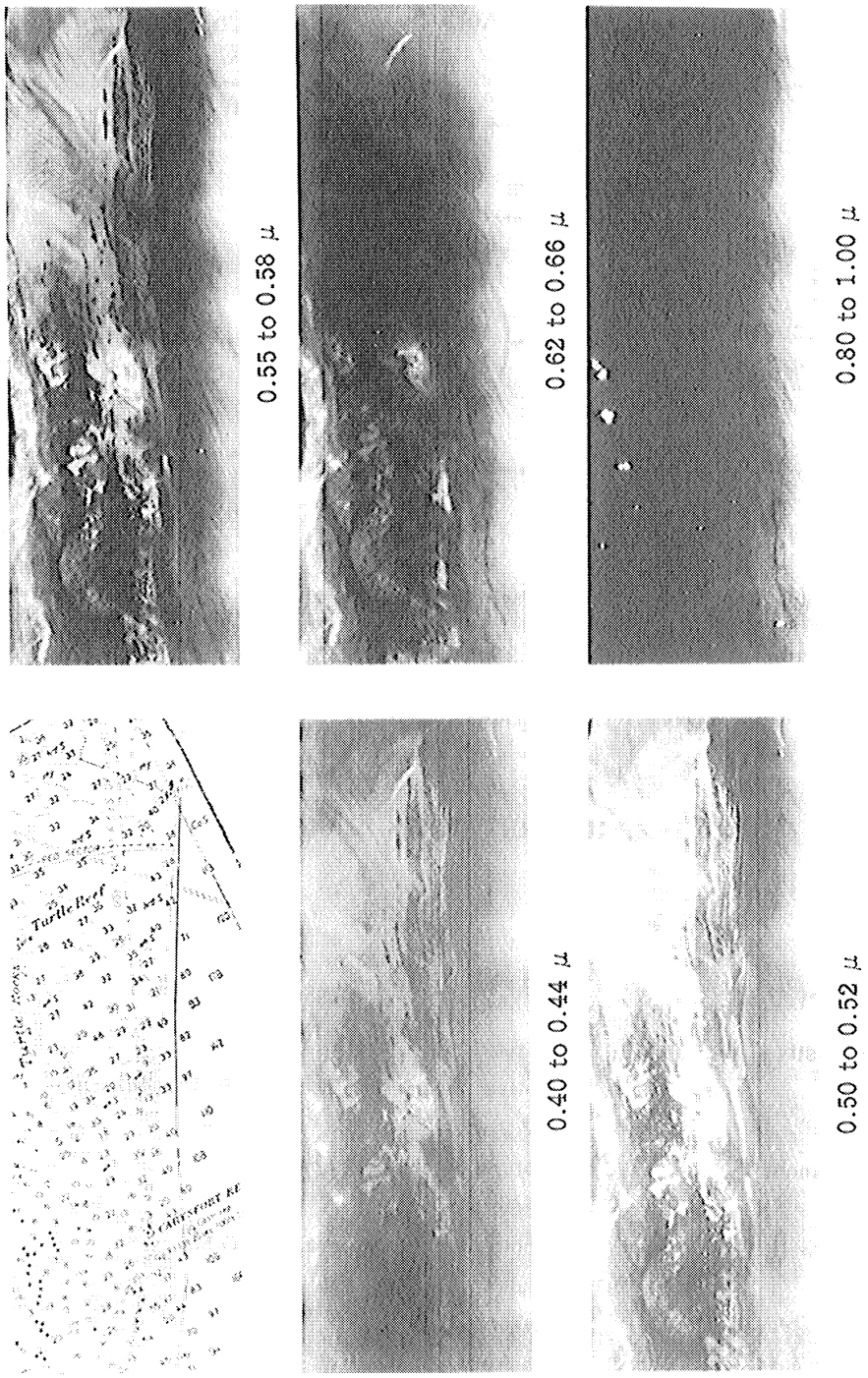
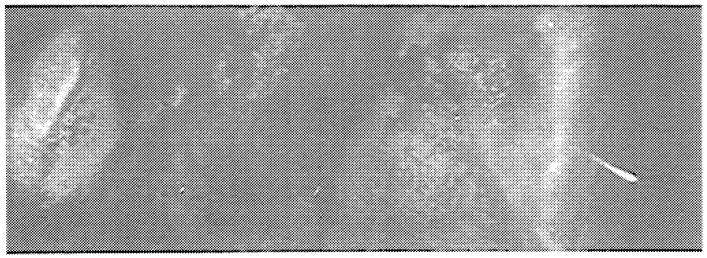
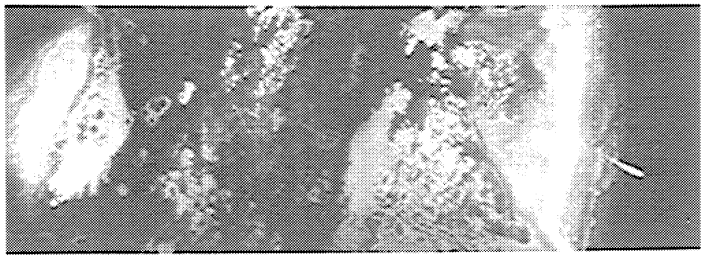


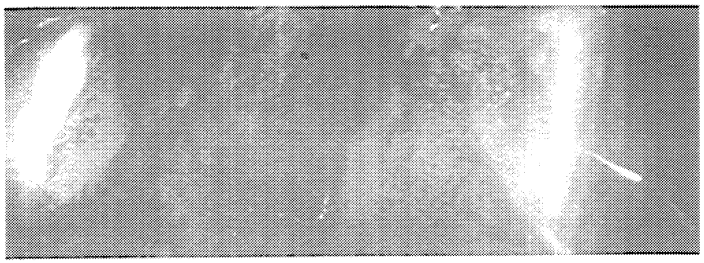
FIGURE D-4a. MULTISPECTRAL COMPARISON OF WATER TRANSMISSION



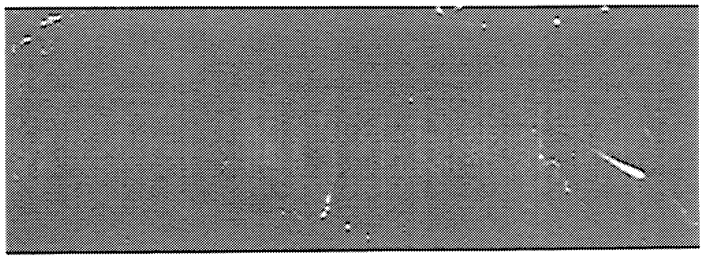
0.40 to 0.44 μ



0.55 to 0.58 μ

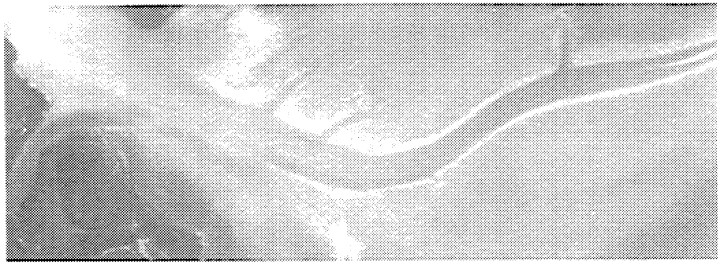


0.62 to 0.66 μ

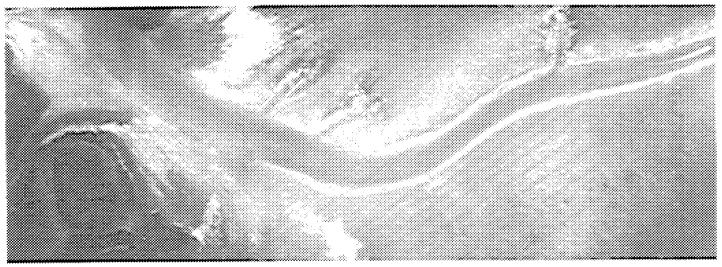


0.80 to 1.00 μ

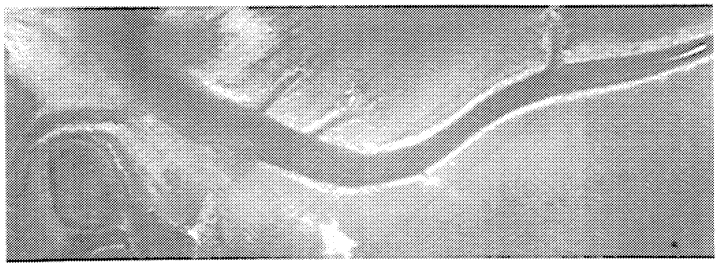
FIGURE D-4b. MULTISPECTRAL COMPARISON OF WATER TRANSMISSION



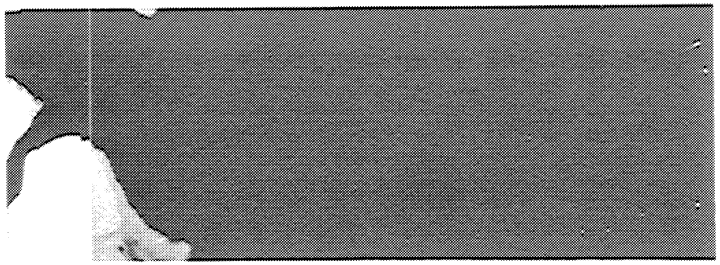
0.40 to 0.44 μ



0.55 to 0.58 μ



0.62 to 0.66 μ



0.80 to 1.00 μ

FIGURE D-4c. MULTISPECTRAL COMPARISON OF WATER TRANSMISSION

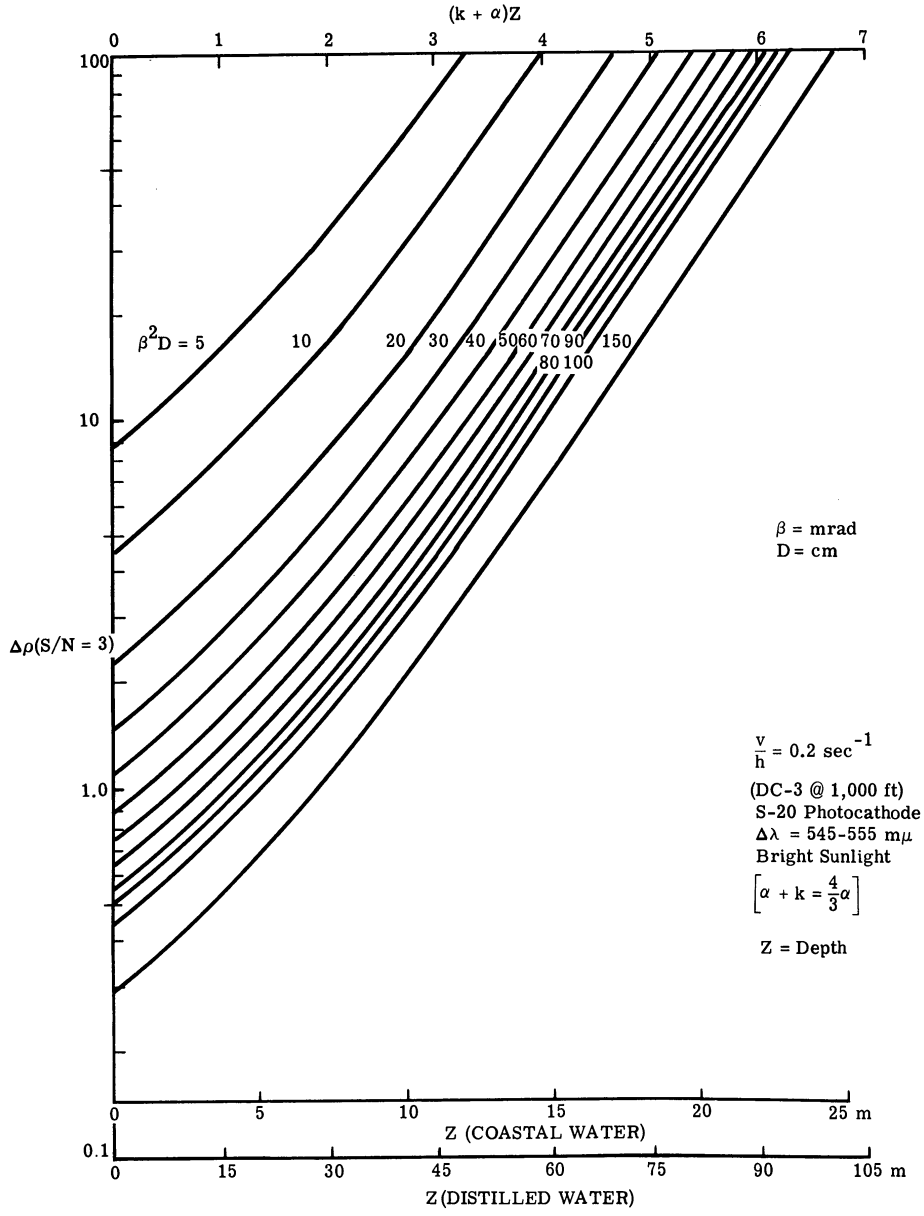


FIGURE D-5a. WATER DEPTH FOR USEFUL OPTICAL MEASUREMENT

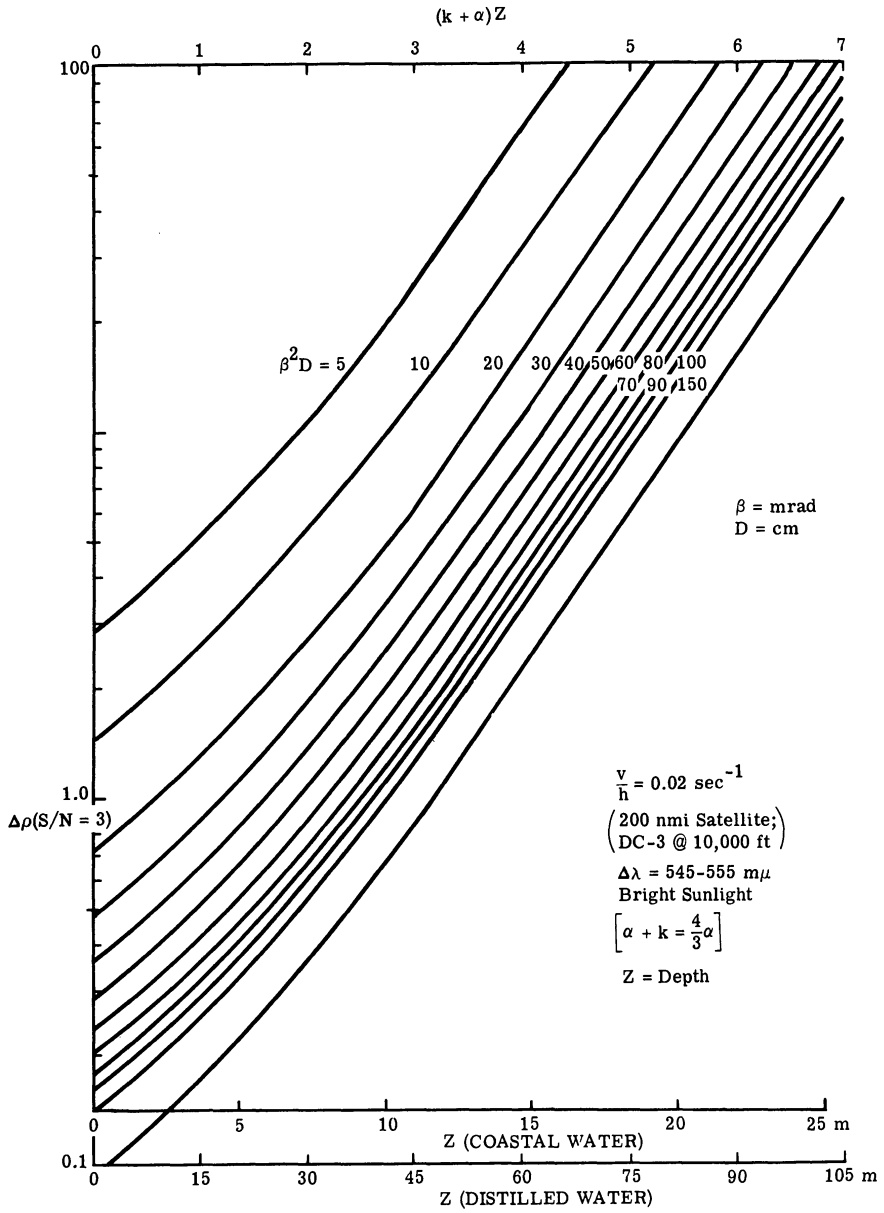


FIGURE D-5b. WATER DEPTH FOR USEFUL OPTICAL MEASUREMENT

DIFFRACTIVE OPTICAL SYSTEM

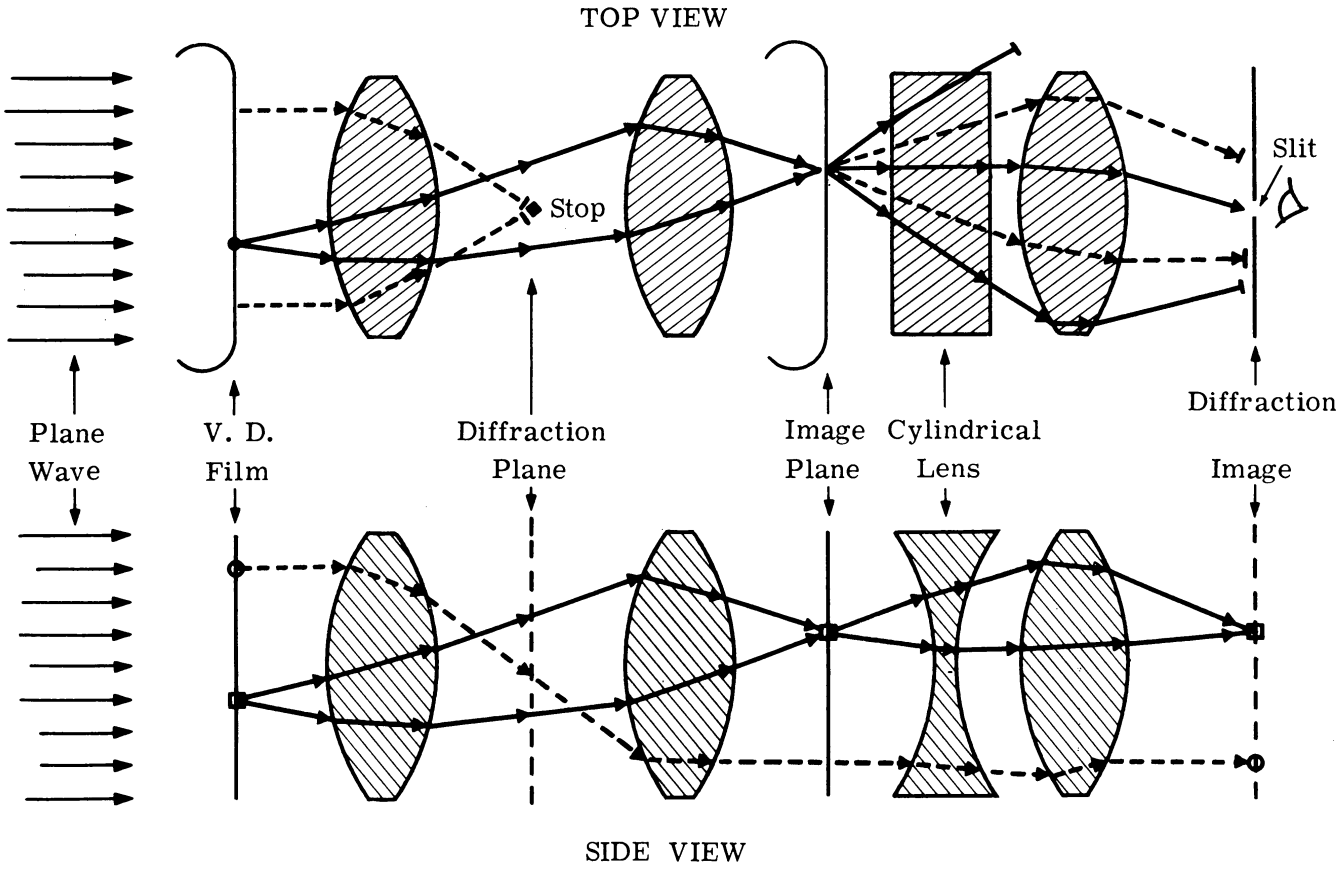


FIGURE D-6.

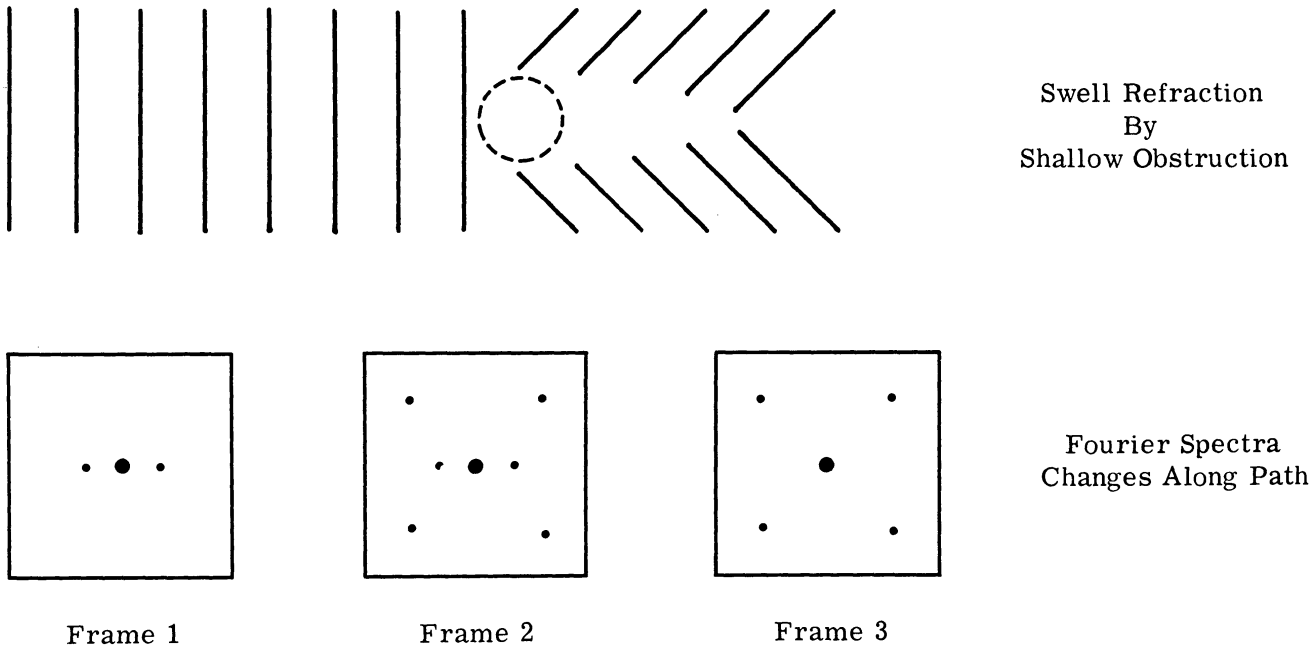


FIGURE D-7. SCHEMATIC FOURIER TRANSFORM CHANGE IN SUCCESSIVE FRAMES

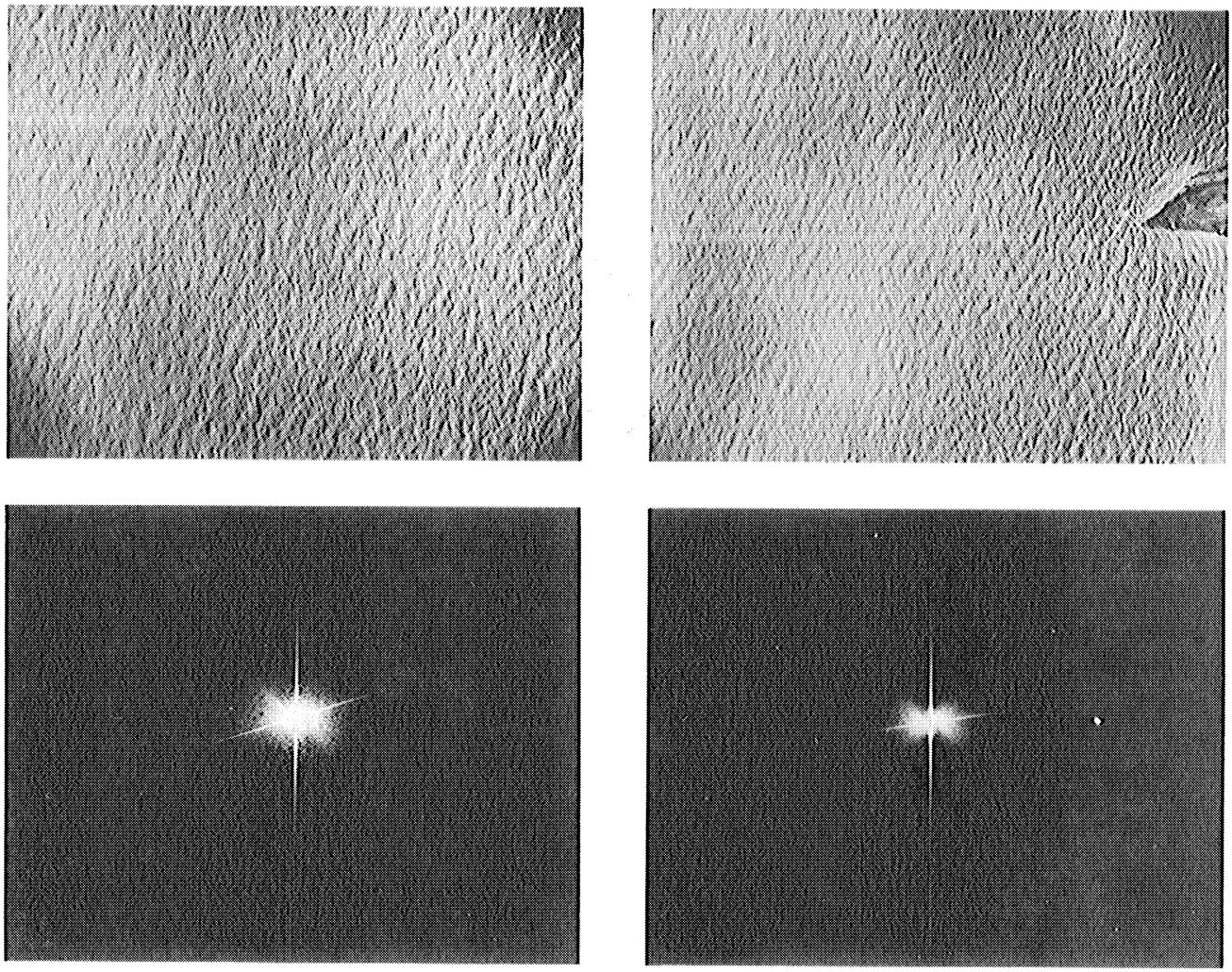


FIGURE D-8. FOURIER TRANSFORMS CHANGE DUE TO ISLAND

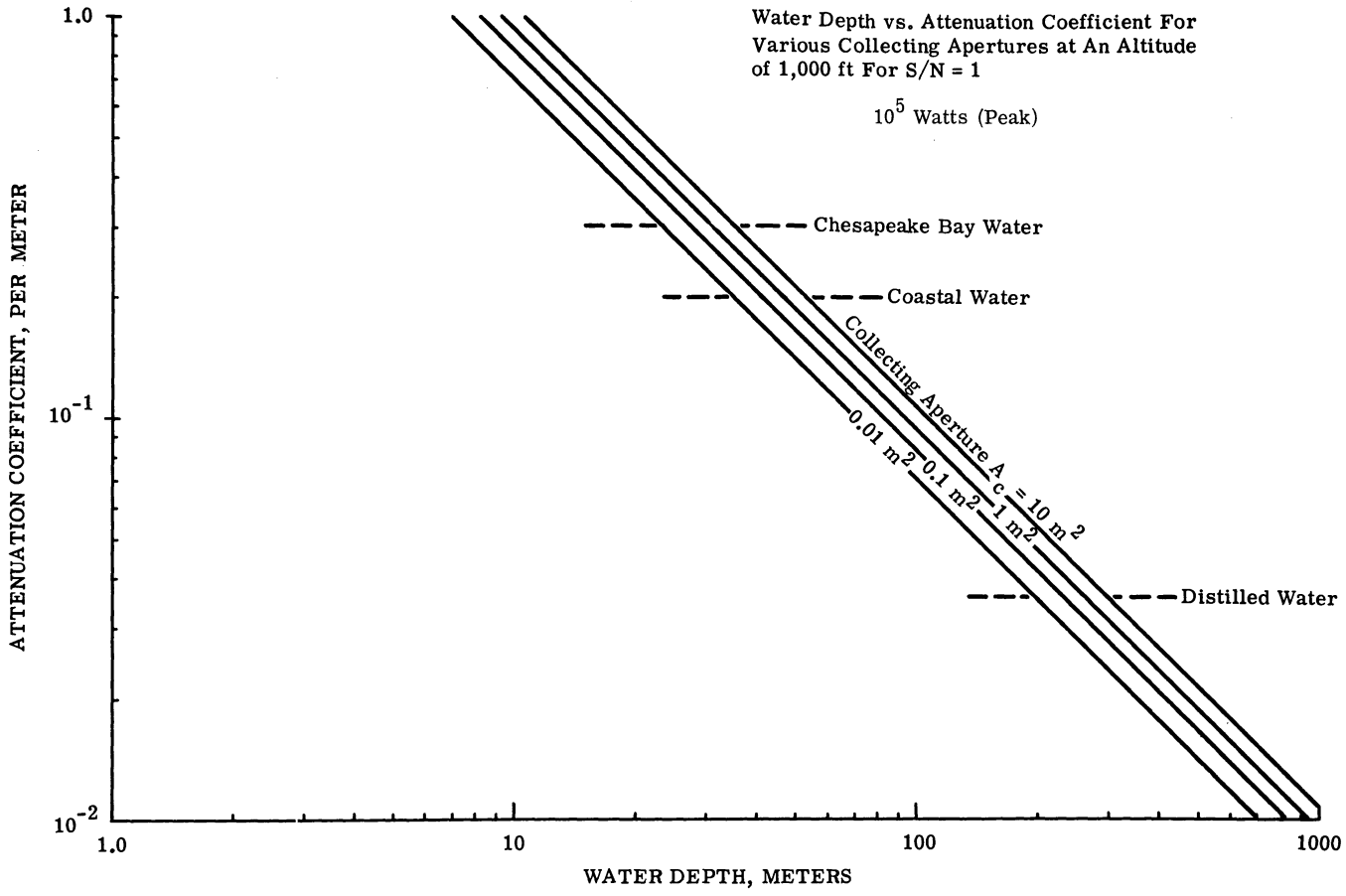


FIGURE D-9a. WATER DEPTH VS. ATTENUATION COEFFICIENT FOR VARIOUS COLLECTING APERTURES AT AN ALTITUDE OF 1,000 FT. FOR S/N = 1

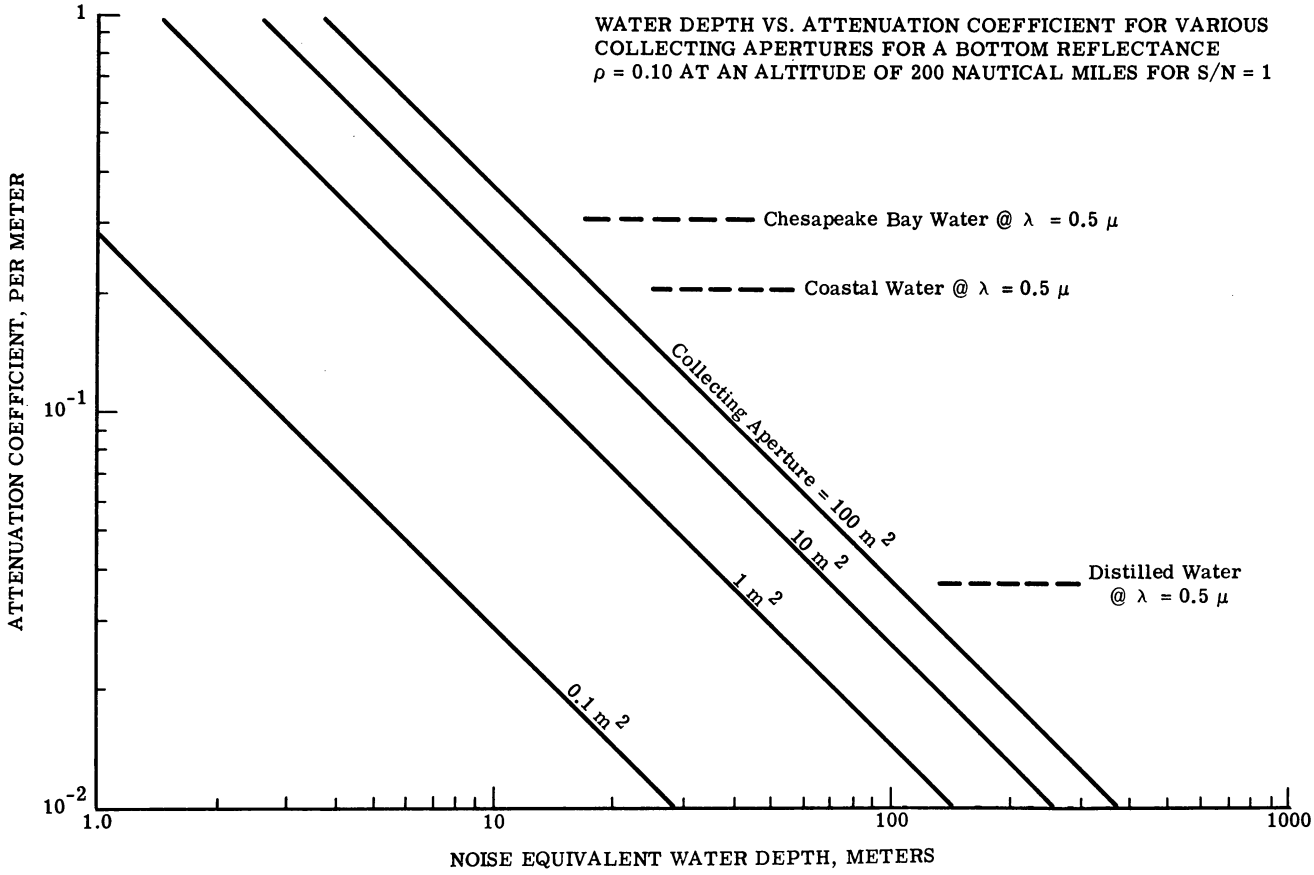

WATER DEPTH VS. ATTENUATION COEFFICIENT FOR VARIOUS COLLECTING APERTURES FOR A BOTTOM REFLECTANCE ρ = 0.10 AT AN ALTITUDE OF 200 NAUTICAL MILES FOR S/N = 1
Chesapeake Bay Water @ λ = 0.5 μ
Coastal Water @ λ = 0.5 μ
Distilled Water @ λ = 0.5 μ
Collecting Aperture = 100 m 2
10 m 2
1 m 2
0.1 m 2
ATTENUATION COEFFICIENT, PER METER
1
10-1
10-2
1.0
10
100
1000
NOISE EQUIVALENT WATER DEPTH, METERS

FIGURE D-9b.

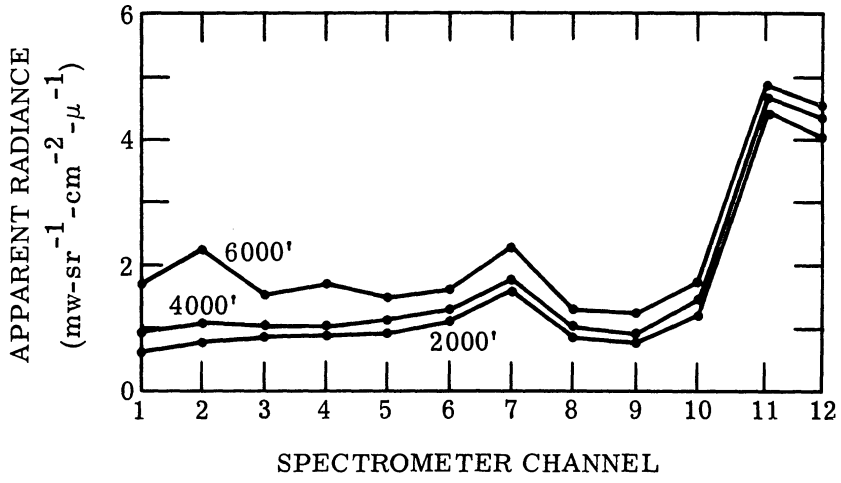
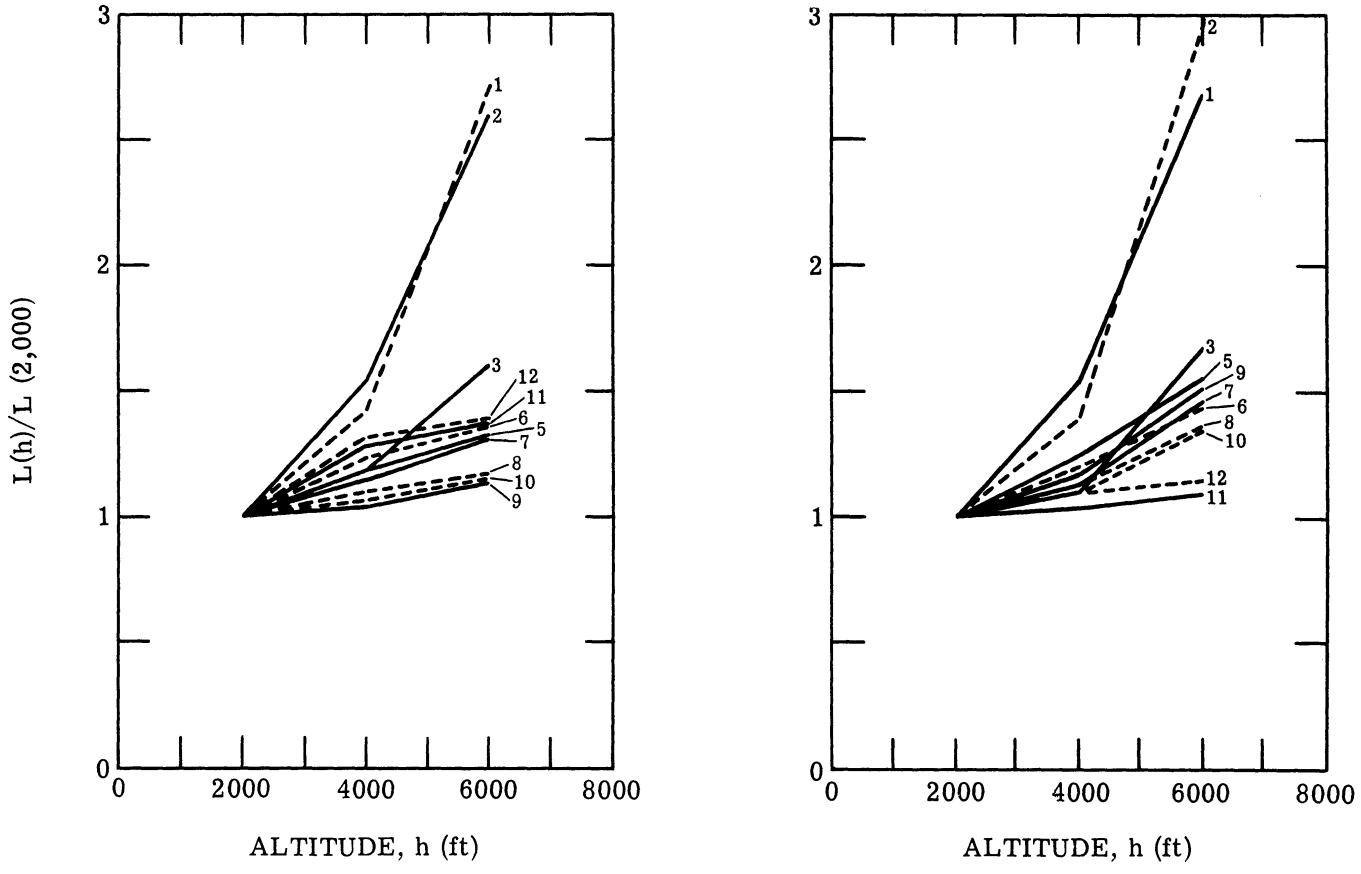


FIGURE F-1. APPARENT RADIANCE OF SOYBEANS VS. SPECTROMETER CHANNEL. July 27, 1966, 11:40 A.M.



(a) Wheat Stubble

(b) Soybeans

FIGURE F-2. RATIO OF APPARENT RADIANCE AT ALTITUDE h TO THAT AT 2000 ft. July 27, 1966, 11:33 A.M. to 12:17 P.M.

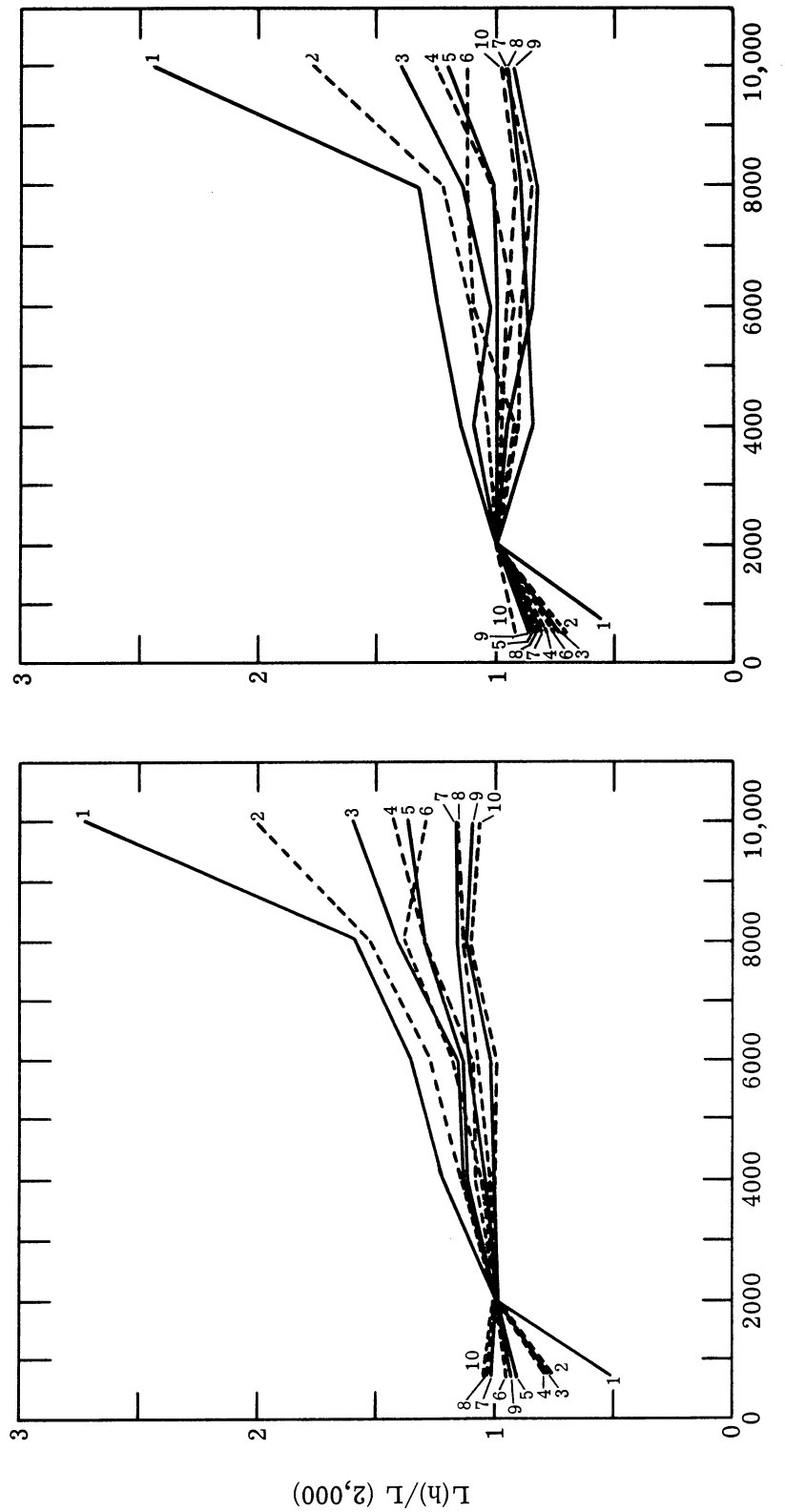


FIGURE F-3. RATIO OF APPARENT RADIANCE AT ALTITUDE h TO THAT AT 2000 ft. September 15, 1966, 8:50 A.M. to 12:24 A.M.

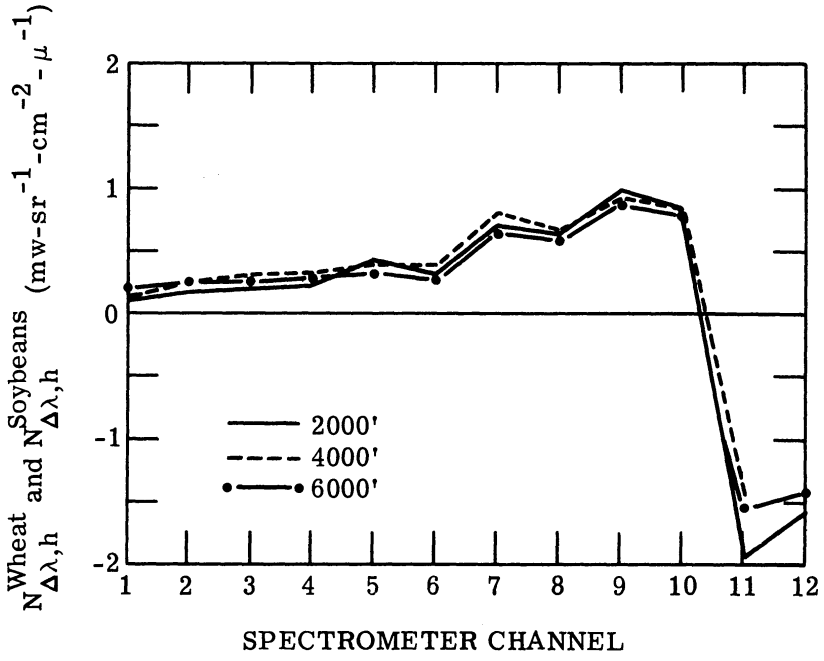


FIGURE F-4. APPARENT RADIANCE DIFFERENCE BETWEEN WHEAT STUBBLE AND SOYBEANS. July 27, 1966, 11:33 A.M. to 12:17 P.M. Hazy atmospheric conditions.

SECTION VI
RADAR AND PASSIVE MICROWAVE RADIOMETRY†

†This section was composed by N. Smith, B. Levy, and G. Suits following the original lecture notes of J. Wolf.

1.0. PASSIVE MICROWAVE RADIOMETRY

1.1. INTRODUCTION

Most of the important means of remote sensing utilize electromagnetic waves. The portion of the electromagnetic spectrum with which this section deals is broadly known as the "radio frequency" spectrum, which extends from about 10^4 to 10^{12} Hz;* the corresponding wavelengths are 30 km to 0.3 mm. It is characterized by the use of more or less conventional electronic devices and circuitry to generate and/or detect the waves. In this section we shall first discuss passive microwave radiometry, or the sensing of microwave radiation naturally emitted by any body at a temperature higher than absolute zero (0^oK). This is the part of the spectrum used by radio astronomers in measuring the temperatures of planets. The passive microwave discussion will furnish some background for the discussion of radar.

In contrast to the passive microwave technique, radar systems are "active," i.e., they illuminate the area being sensed with radiation of a specific radio frequency, and use the power that is reflected or back-scattered from the area and from objects in the area.

It is of interest to note that radar techniques are applicable over a broad frequency range, but because of technical and economic considerations, the radar development has tended to cluster in and around a few frequencies. These are designated by letters, as in table 1-I.

1.2. AVAILABLE POWER

The black-body radiation curve (fig. 1-1) shows that the power emitted at radar wavelengths by blackbodies at temperatures of 300-1000 deg K is much less than that emitted at infrared wavelengths of 1-10 microns. The spectral radiance in W m^{-2}-sr^{-1} Hz^{-1} is given to sufficient accuracy by the Rayleigh-Jeans approximation: $L_\nu = 2kT/\lambda^2$. Even though the emission may be six or more orders of magnitude smaller, it is quite detectable because of the effectiveness of radio amplifying circuits.

A microwave radiometer consists basically of a large paraboloidal mirror (called, in radar parlance, a "dish") at the focus of which is placed a small directive microwave antenna to collect the energy and feed it, via a waveguide, into a special microwave radio receiver which amplifies it until it can operate a recorder or other display devices. It is in principle

*The "hertz," abbreviated Hz, is the unit of frequency, formerly called the "cycle per second."

TABLE 1-I. DESIGNATION OF MICROWAVE BANDS

Designation	Frequency, MHz	Wavelength (cm)
P-Band	600	50
L-Band	1250	24
S-Band	3000	10
X-Band	9000	3.33
K~-Band	15000	2
K_a-Band	37000	0.8

like an IR radiometer, except that the mirror is larger (1 to 100 meters in diameter) and a conventional radio receiver is substituted for the detector and amplifier. Unlike the IR radiometer, however, the resolution is governed by the size of the mirror rather than that of the detector, the time of dwell on each resolution patch is long because the signal is very weak, and it can see through clouds, provided they are optically thin at the operating wavelength. The mirror is scanned or moved in angle so as to scan the area of interest. The output of the receiver is recorded as a function of the pointing direction of the mirror.

We shall now discuss in turn the factors that determine the resolution capability and sensitivity of a microwave radiometer, and introduce some terminology and concepts needed for describing radar systems.

1.3. RESOLUTION

The resolution of the parabolic antenna is limited by diffraction in the same way that the resolution of a telescope or camera lens is ultimately limited. If the diameter D of the antenna is very great compared with the wavelength being used, the beam width θ is given approximately by

$$\theta = \frac{\lambda}{D} \text{ radians.}$$

The directional pattern of the antenna (fig. 1-2) represents the effectiveness of the antenna in focusing power upon the receiver input from irradiance arriving at various angles to the direction of the antenna axis. The large maximum in the direction of the axis is called the "main beam" and the other, smaller, maxima are called the "side lobes." The "beam width" is defined as the angle between the "half-power" points, i.e., the directions at which the power-gathering effectiveness of the antenna is reduced to half of its value on the axis. The side lobes are generally undesirable but inevitable. They can, however, be reduced by appropriate "tapering" of the illumination, i.e., adjusting the directional pattern of the microwave antenna at the focus so as to reduce the illumination near the periphery of the paraboloid mirror. Such a reduction of the side lobes is generally accompanied by some broadening of the main beam.

For most practical purposes, it suffices to consider that the beam width is given by

$$\theta \doteq \frac{\lambda}{D}$$

For example, a dish one meter in diameter operating at an X-band wavelength of 3 cm would have a half-power beam width θ of:

$$\theta = \frac{3}{100} = 0.03 \text{ radians or } 1.7^{\circ}$$

Practical beam-widths are not very small. This is a serious limitation on the passive microwave technique as a remote sensor, for it permits only low resolution sensing.

1.4. SENSITIVITY

The microwave radiation incident on the dish and brought to a focus is there picked up by a microwave antenna, such as, e.g., a microwave horn, which feeds it to the input of a microwave radio receiver. If the receiving system did not add any microwave energy of its own, the power output of the receiver would be proportional to the power picked up by the antenna. Such is not the case, however, for a significant amount of radio noise is generated in the receiver input circuits and electronic devices by random thermal motions of electrons; this noise adds to and is virtually indistinguishable in character from the incoming microwave radiation. The "sensitivity" of the receiver is a measure of how strong the incoming signal must be in order to be detectable through the unwanted noise produced by the radiometer system itself. A noise equivalent power and noise equivalent irradiance at the aperture can be defined in the same way as for any radiometer.

An important parameter of the receiver is the "band width" or range of frequencies that will be accepted and amplified by its receiver. The band width $\Delta\nu$ is generally measured between the half-power frequencies. The noise power that lies in this band is effective in competing with the signal.

Figure 1-3a is an elementary diagram of a radiometer. The voltage source V, in combination with the resistance R, is a model of the microwave radiation input to the receiver; the input resistance of the receiver is symbolized by the resistor R'. The voltage V represents the effective voltage corresponding to the microwave power arriving at the focus of the parabolic antenna, and R represents the internal resistance of the source - largely the radiation resistance of the antenna. For the maximum signal power to be transferred to the receiver, R' must equal R. The maximum power that can be obtained from the source is then:

$$P_{max} = \left.\frac{V^2 R}{(R + R')^2}\right]_{R=R'} = \frac{V^2}{4R}$$

This is called the available power from the source.

Any resistance such as R', at a temperature T, is a source of noise power, owing to the random thermal motions of electrons in it. The noise power available from such a resistance, within the band width $\Delta\nu$, is P_n, given by:

$$P_n = kT\Delta\nu$$

Note that this does not involve the value of the resistance. The corresponding available rms noise voltage input to the receiver is:

$$V_n = \sqrt{4kTR\Delta\nu}$$

where k = Boltzmann's constant 1.38×10^{-23} joules per deg K,
T = absolute (Kelvin) temperature,
R = resistance in ohms,
ν = frequency band width of receiver in Hz,
V_n = root-mean-square (rms) noise voltage.

Even if we have no external input to the receiver, i.e., V = 0 and the temperature of R were 0 deg K, there is at least a noise power input corresponding to the temperature of the receiver. If the temperature of the receiver were 290°K, for example, and $\Delta\nu = 10^6$ Hz, the minimum input P_n is 4×10^{-15} watts.

This is not, however, the only source of noise in a practical receiver. Additional noise power is generated in the electronic devices and circuits of the receiver. The totality of the noise generated in a receiver may be represented by an "apparent noise input," which is the noise power output P_0 divided by the power gain G of the receiver with no external input. If the resistance R' is at a known temperature T we may define a "noise figure" as the ratio of the apparent noise input to the noise input due to the resistance R' - i.e., $kT\Delta\nu$.

$$\text{noise figure} = \frac{P_0}{kTG\Delta\nu}$$

The noise figure is a measure of how many times worse an actual receiver is than a theoretically perfect receiver; it is the way generally used to describe the threshold signal properties of a radio receiver.

It has become the practice, in microwave radiometry, to use an equivalent "noise temperature" for the receiver, obtained by dividing the apparent noise power input by $k\Delta\nu$. This equivalent temperature accounts for the noise produced by the receiver. The noise equivalent power of a microwave radiometer is proportional to the noise equivalent temperature, T_e

$$T_e = \frac{P_0}{Gk\Delta\nu} = \text{T of resistance R' } \times \text{ noise figure,}$$

where $\frac{P_0}{G}$ = noise equivalent power.

The radio frequency noise voltage input to a receiver may be regarded as a continuous spectrum of simple harmonic component voltages with amplitudes and phase randomly chosen. The receiver band width limits the spectrum at the output available for measuring the average power, so that we have the equivalent of only $n = t\Delta\nu$ independent samples of the average in time t. This determines the uncertainty in measuring the input microwave radiation power.

For analysis of the relationship between the average receiver output $\overline{V}$ during the sampling interval and the input power, let V_s be the rms (root-mean-square) value of the output voltage V determined over a time very long compared with the sampling interval, and define the parameter ρ:

$$\rho = \frac{\overline{V}}{V_s}.$$

Since the instantaneous phases and amplitudes of the input voltage components are random variables, the distribution of the magnitudes of ρ is described by the normalized distribution function $p(\rho)$:

$$p(\rho) = 2\rho \exp(-\rho^2)$$

That is, the probability that ρ lies between ρ and $\rho + dp$ is $p(\rho)d\rho$.

This distribution function was first derived by Lord Rayleigh in his "Theory of Sound," and bears his name. It is, in fact, the polar form of a bivariate normal distribution. The mean value of ρ is 0.8862 and the standard deviation from the mean is 0.4632; the rms value is, of course, 1.

Thus we may estimate the rms noise voltage output V_s from the average value of the output voltage, $\overline{V}$:

$$\overline{V} \doteq \left(0.886 \pm \frac{0.463}{\sqrt{n}}\right) V_s$$

where n is $(t\Delta\nu)$. V_s then = 1.128 $\overline{V}$, with an uncertainty of $\pm 52.3/\sqrt{t\Delta\nu}$ percent. This is related to the input noise power and the effective receiver temperature T_e as shown above, so that

$$kT_e G\Delta\nu = \frac{V_s^2}{R_0}$$

where R_0 is the output termination resistance.

1.5. MODE OF OPERATION

Now consider the radiometer antenna looking at a part of the outside world that might be represented as a black body at temperature T_1. The antenna then acts like a resistor at the temperature T_1. In the receiving system the effective receiver temperature T_e including the noise figure adds to the antenna temperature; thermal radio noise powers, and therefore the temperatures, are additive, since the noise signals are random.

The output power P_0 of the receiver with the antenna connected is then $Gk(T_e + T_1)\Delta\nu$. If the output load resistance is R, the rms output voltage V_0 is:

$$V_0 = \sqrt{P_0 R} = C\sqrt{T_e + T_1}$$

where $C = \sqrt{RGk\Delta\nu}$

One mode of operation of the radiometer involves connecting the receiver alternately to the antenna and to a real resistor equal in value to the antenna radiation resistance while simultaneously connecting the output to one or the other of two identical averaging circuits. See figure 1-3b. The switching, or chopping, is done at an audio frequency rate. The capacitor C_1 stores the time average of the output obtained while the signal comes from the antenna, while capacitor C_2 stores the output from the resistor at a standard temperature T_2. The voltage difference between the two is measured by a voltmeter.

Calibration of the voltmeter V in terms of antenna temperature T_1 is done as follows:

a) Point the antenna toward outer space, avoiding any solar or terrestrial input via side lobes. T_1 is arbitrarily taken as 3^oK.

b) Balance the meter circuit to give zero reading.

c) Point the antenna at warm ground or a black body. T_1 is now about 300^oK.

d) Adjust gain for full scale reading. The radiometer now reads from 3^o to 300^oK.

This calibration is accurate enough for most purposes.

It is important to note that if a uniform black body at temperature T completely fills the antenna beam, the noise power available from the antenna in the frequency range $\Delta\nu$ is exactly $kT\Delta\nu$, regardless of the antenna gain. The effect of the antenna size is to limit the area at which it looks. It follows that for a black body with a non uniform temperature across the area covered by the beam, the available power is $k\overline{T}\Delta\nu$, where $\overline{T}$ is the average temperature of the area. Thus if the beam covers an area of $(R\theta)^2$, at a temperature T, in which there is a smaller area A at temperature T_2, the average temperature is given by:

$$\overline{T} = \frac{\left[\frac{\pi}{4}(R\theta)^2 - A\right]T_1 + AT_2}{\frac{\pi}{4}(R\theta)^2}$$

If A is much smaller than $(R\theta)^2$, $\overline{T}$ may be only slightly different from T_1, even though T_2 may differ greatly from T_1.

If, for example, the area A is a small hot area at 1000^oK in a forest, like the beginning of a forest fire, the antenna beam width is 0.01 radians, the range is 1000 m, and the background temperature is 300^oK, the average temperature over the antenna beam is

$$\overline{T} = \frac{\left[\frac{\pi}{4}(1000 \cdot 0.01)^2 - 1\right]300 + 1.1000}{\frac{\pi}{4}(1000 \cdot 0.01)^2} = 308.9^{\circ}K$$

The hot spot is greatly diluted by the background, and might be missed on a routine search.

The radiometer receiver is connected for the same length of time to each source. Therefore, the reading of the voltmeter is proportional to the difference between the output values,

$$V = C\left(\sqrt{T_1 + T_e} - \sqrt{T_2 + T_e}\right),$$

where the C contains the constant of the integrating circuit. Since T_e and T_2 are both known, we can thus obtain a value for T_1, and thus for the temperature difference $\Delta T = T_1 - T_2$. The probable error of this value is the square root of the sum of the squares of the errors in measurement of the output voltages for the two sources.

Since the observation time t is divided between the two sources, the number of samples for each is $\sqrt{t\Delta\nu/2}$, so that the error in measuring the antenna temperature, for example, is σV_0 given by the error analysis above

$$\sigma V_0 = \frac{0.52\, V_0}{\sqrt{t\Delta\nu/2}} = \frac{0.37\, V_0}{\sqrt{t\Delta\nu}}$$

Now

$$\sigma_1 = \sigma(T_1 + T_e) = \frac{\sigma V_0}{dV_0/d(T_1 + T_e)} = \frac{0.74\, C\sqrt{T_1 + T_e}}{\sqrt{t\Delta\nu}} \cdot \frac{C}{2\sqrt{T_1 + T_e}}$$

since $V_0 = C\sqrt{T_1 + T_e}$

$$\sigma_1 = \frac{1.48}{\sqrt{t\Delta\nu}}(T_1 + T_e)$$

and similarly for the resistor at temperature T_2

$$\sigma_2 = \frac{1.48}{\sqrt{t\Delta\nu}}(T_2 + T_e)$$

So the temperature difference between the antenna and the resistor is

$$\Delta T = T_1 - T_2 \pm \sqrt{\sigma_1^2 + \sigma_2^2}$$

If $T_1 \doteq T_2$ as would usually be the case in looking at terrain, $\sigma_1 \doteq \sigma_2$ and

$$\Delta T \doteq T_1 - T_2 \pm \sqrt{2\sigma_2} = T_1 - T_2 \pm \frac{2.09}{\sqrt{t\Delta\nu}}(T_2 + T_e)$$

which is the working formula.

As an example, suppose $T_2 = T_1 = 300^{\circ}K$, t = 1 second and $\Delta\nu = 10^8$ Hz. Assume a low noise-figure, cooled receiver, with $T_e = 100^{\circ}K$. The probable error is then:

$$\pm 2\frac{300 + 100}{\sqrt{10^8}} = 0.8^{\circ}K$$

We have thus far assumed that (a) there are no power losses between the reference resistance switch and the antenna dish, (b) the target is a black body, and (c) there is no absorption of power in the path from the target to the antenna. None of these are strictly true.

1.5.1. POWER LOSSES AT THE SENSOR. There is always a power loss in a practical transmission line or wave guide. This is considerably increased if there is a standing wave in the line due to reflections of the traveling waves by irregularities or mismatches of impedance at the microwave antenna and the receiver. There is also a coupling loss between the microwave antenna and the electromagnetic field at the focus. It is very difficult to maintain a good impedance match over the broad frequency band used in radar and microwave radiometry. Such losses can be represented by a resistor at the temperature of the line - about $300^{\circ}K$ - which would add a corresponding noise power to the receiver input. Not only do we have less available power from the antenna, but the effective temperature of the receiving system is increased. The losses in the system ahead of the receiver input terminals must be kept small, if an acceptable noise equivalent temperature is to be obtained.

1.5.2. GREY BODY EMISSANCE. The efficiencies of emission and reflection of radiation from an object are closely related; the sum of the two is unity for opaque bodies. A good reflector is therefore a poor emitter, and vice versa. The black body temperature corresponding to the actual power emitted by a target at, say, $300^{\circ}K$, may be much lower. This is discussed later.

1.5.3. ATMOSPHERIC ABSORPTION. Clear air is relatively transparent for microwaves longer than 3 cm or so. For shorter wavelengths, however, there is significant absorption of power by the oxygen and water vapor in the air; these gases also emit radiation at the same frequencies as they absorb. The absorption and emission occur at a large number of discrete frequencies or spectral lines. These are too close together in frequency to be resolved by the microwave receiver we use for radiometry, so the effect is one of a general absorption region with some peaks and valleys due to the arrangement of the lines in spectral bands. Figure 4 indicates the absorption of microwave power over a representative path through humid air (66% relative humidity at $32^{\circ}F$ or $291^{\circ}K$), measured in decibels per kilometer.

The effect of atmospheric absorption and radiation is to dilute the target signal with that emitted by the air, since the emissance of a gas in a transmission path is equal to its absorptance. For example, for the atmosphere of figure 1-4, using 1.5 cm waves over a 10 km path, the path absorption is $10 \times 0.2 = 5$ decibels, or a factor of 0.316. The emissance is 0.316 and the transmittance is 0.684. If one looks at a $400^{\circ}K$ target over this path, the antenna temperature will be

$$T_1 = 0.684 \times 400 + 0.316 \times 291 = 366^{\circ}K$$

indicating a loss of contrast.

On the other hand, if we were to use 1.5 mm waves over a similar path, the transmittance would be only 0.001, and we would see only the emission of the air itself. Similarly, if we were to use the 1.5 cm waves and point the antenna straight up, so that it looked at the cold sky (at nearly $0^{\circ}K$) we would likewise measure only the emission of the water vapor, and thus be able to remotely measure the water content of the air.

1.6. TARGET CHARACTERISTICS

1.6.1. THE SKY. Aside from the sun, the sky is very cold at microwave frequency. The galactic radiation background brightness temperature (largely due to synchrotron radiation from electrons traveling with a velocity nearly that of light in a magnetic field of 10^{-5} gauss or less) decreases as the 2.5 power of wavelength for wavelengths shorter than 10 meters, reaching the intergalactic black body temperature of $3^{\circ}K$ at and below 30 cm. As mentioned above, if the antenna is pointed straight up the sky temperature for wavelengths below 1.5 cm is largely due to water vapor or water droplets in clouds, or to sunlight scattered by the atmosphere into the beam of the antenna.

If the angle of elevation of the antenna is ϕ (fig. 1-5) the temperature it sees depends on the wavelength λ and on how much air, clouds, haze, etc. is in the beam. For ϕ small there is much air and the temperature is that of the lower atmosphere. As ϕ increases the air path becomes shorter and the temperature decreases. It should be noted that, for ϕ small, radiation from the earth's surface or from objects on the surface can enter the antenna side lobes, and produce a spurious effect.

1.6.2. THE SUN. This is an extremely strong emitter at $6000^{\circ}K$; it is detectable in spite of clouds. Even when the antenna is pointed toward the ground, enough solar radiation can be seen in the side lobes to spoil the data.

1.6.3. WATER. A radiometer looking downward at open water in lakes or ponds will frequently detect only the sky temperature, since smooth water is a good reflector for microwave

power over a wide range of angles, and hence a poor emitter. The apparent temperature depends on the reflection coefficient, which is a function of both polarization and elevation angle. At small angles the reflection coefficient is nearly unity, so T is the sky temperature - about 300^{o} at $\phi = 0$. For vertical polarization, the reflection coefficient decreases rapidly with increasing ϕ to a very small value at the Brewster angle;* here the radiometer sees only the emission from the water, again about 300^{o}. For larger angles T is then a mixture of the water and the sky temperature, the latter being quite low; the water emission is diluted to a large extent by the cold sky. For horizontal polarization, there is no Brewster angle and the temperature is just a mixture of sky and water temperatures, corresponding to the variation of the reflection coefficient with angle.

If the surface of the water is rough, it is not as good a reflector, and the sky reflection plays a less important part. By comparing the effective temperatures for vertical and horizontal polarization, the roughness can be detected. If observations are made at two or more wavelengths, the scale of roughness of the water could be measured.

1.6.4. SOIL. Soil appears much warmer than water because soil emissance is greater and varies with the water content. For very dry sand, the radiometer can detect radiation from below the surface. Accordingly, radiometric measurements of the moon were made in an attempt to determine the thickness of an expected "dust layer" by subsurface radiation.

1.6.5. MISCELLANEOUS. Concrete and asphalt pavement are smooth with low emissance and reflect the sky, so they appear relatively cool, even though their actual temperature might be much greater than 300^{o}K.

Metal strongly reflects the sky or surroundings, and emits very little microwave radiation even when it is hot.

Vegetation is warm and is generally a good emitter.

1.7. SYSTEM SPECIFICATIONS

The selection or design of a passive microwave radiometric system depends on its purpose and proposed modes of operation and necessary boundary conditions of size, cost, weight, power supply, and other requirements. The principal technical parameters include wavelength, diameter of antenna, resolution, effective receiver temperature T_e, range, angular scan or search rate, aircraft speed if airborne, dwell time on target, expected target and background temperatures and variances, area of target and the interface between the radiometer and data

*The Brewster angle is discussed in the section on propagation.

processing. These parameters are not independent; optimization of some of them results in the degradation of others.

The search rate is governed by the antenna beam angle θ and the desired dwell-time t on a resolution element. The solid angle corresponding to a resolution element = θ^2 steradians, provided $\theta < 5$ degrees of arc. If the antenna is in a fixed location the angular scan rate must be θ^2/t steradians per second. If the antenna is airborne and scans from side to side, the aircraft speed and altitude must be such that successive scans will be approximately contiguous.

2.0. RADAR

2.1. INTRODUCTION

The term "Radar" is an acronym of the phrase Radio Detection and Ranging. It is an active system, depending upon the reflection of a radio wave from a distant object. A radio signal is transmitted and the echo, or signal reflected from the object, is received and processed to obtain information about the object, such as its range, direction and other characteristics. There are two general classes of radar. The first and most common is pulse radar; the signal is a short burst of radio frequency radiation, and the range is determined by the time delay between the emission of the pulse and the reception of the echo from the distant object. The second type is known as C-W (continuous-wave) radar; the emitted signal is an unmodulated wave, the phase of which is compared with the phase of the returning echo in order to determine the range. Sophisticated radar systems often use a combination of the two techniques.

The earliest use of radar (long before the term itself was invented) was for measuring the height of the ionosphere. Both C-W and pulse systems were used in 1925 and 1926 by Appleton in England and by Breit and Tuve in the United States. Radio wave reflections from aircraft were observed during the operation of a high frequency experimental radio link in the 1930's, and were regarded as a nuisance which interfered with the experiments!

Accidental radar has been observed by nearly everyone who has a television receiver. The reflection from an errant aircraft causes the TV signals to arrive via two paths, whose relative length changes with time. The signals over these paths are alternately in and out of phase, which produces flickering and horizontal bars on the TV screen. Again, reflection from a fixed object like a building can produce a "ghost" owing to the additional time delay of the reflected wave.

The basic elements of a radar set are a pulsed high-power transmitter, a receiver, an antenna, a switch for connecting the antenna alternately to the receiver and transmitter, and a device to display the received radar echoes, generally a cathode-ray oscilloscope. The transmitter operates at a fixed frequency, generally in one of the frequency bands listed in the Passive Microwave chapter. For the best range resolution the pulse is made as short as practicable; pulse lengths of less than a microsecond are common at the higher frequencies. The switch, called a TR (transmit-receive) switch, connects the transmitter to the antenna for the duration of the pulse and/or disables the receiver for the same period. The antenna is a large dish, often but not always paraboloidal, to obtain a narrow beam for concentrating the power. It is designed either to point or to scan in azimuth and elevation. The display equipment depends on the type of information desired.

In operation the direction of a detected object is the direction in which the antenna points. The range of the object is measured by the time delay until the echo arrives. The velocity of propagation in air is about 3×10^8 meters per second, so that the range is about 150 meters per microsecond of delay time, since the wave has to travel out and back. It is easy, in electronics, to measure one microsecond.

As a remote sensor, radar has many advantages. It can operate in darkness or under conditions of low visibility. It can see through clouds owing to the use of wavelengths longer than, for example, infrared or visible light. It can sense velocities of, for example, birds, hailstones and waves on the water. It can provide both a limited penetration of the surface, depending on the wavelength used, and an indication of surface roughness. It gives a view of the world in a different range of wavelengths. The use of techniques involving wave polarization can yield new information, since the polarization of the transmitted waves is known.

One of the outstanding advantages of radar is its unique capability for ranging. It is capable of operating over great distances limited only by the radio line-of-sight.* From an airplane the radar line-of-sight is 30 to 50 miles in all directions. An airborne radar can cover very great areas quickly; it is not necessary to fly directly over a region to get a radar picture, as would be the case for, say, aerial photography. In fact, radar is not effective for mapping directly below the aircraft. In order to get a map-like view an aerial camera must be directly over the area, while a radar can do it from the surface many miles away.

A radar on a ship, for example, can record a map-like view by using a radial scan on the cathode ray tube, with the spot starting at the center of the screen and moving outward in the direction in which the antenna points. The spot is made brighter (intensity modulated) when the echo is received, and the range is indicated by the distance the spot has moved from the center. It is almost as though the spot represented the pulse moving outward until it encountered the target (except that the pulse actually goes out and back). As the antenna scans in azimuth, the direction of the line changes accordingly. The result is a faithful map of the area, showing actual locations of neighborning ships, etc. The radar is at the center of the map. Such a display is known as a PPI (plan position indicator) display.

Among the disadvantages of radar are the following: It is very difficult to obtain fine resolution in angle, owing to the width of the antenna beam. It is also difficult to obtain fine resolution in range; if two objects differ by less than 150 meters in range, for example, they can not be resolved by a one-microsecond pulse. In mapping of terrain the pictures cannot be interpreted in great detail, for the image is grainy. The recording means may not yet be available for accurate mapping; this is particularly true of airborne radar. Also the equipment is somewhat heavy and bulky, and consumes much power.

*Since atmospheric refraction is much greater for radio waves than for optical waves, the radio line of sight is somewhat greater than the visual line of sight.

2.2. THE RADAR EQUATION

A transmitter radiating a power P_t which is spread uniformly in all directions is called an isotropic radiator. Its intensity I (isotropic) is $P_t/4\pi$ watts sr^{-1}. The power P'_r intercepted by an area A' normal to the direction of propagation and at range R is

$$P'_r = P_t \frac{A'}{4\pi R^2}$$

A directional antenna concentrates the same power P_t at or near the area A'. The intensity of the source in that direction is greatly increased. We define the power gain of the antenna as the ratio I(antenna)/I(isotropic). Thus, G = I(antenna)/I(isotropic) so that

$$P'_r = P_t \frac{GA'}{4\pi R^2}$$

The receiving area A' is, with exceptions to be noted later, just the projected area of the radar target normal to the radar beam direction. It is assumed that the power received by the target is reradiated or "scattered" in all directions.

The target can now be considered as a new transmitter, radiating the power P'_r as though from antenna with gain G' directed toward the radar receiver. As in the case of the original radar beam, the portion of P'_r received by the area A of the receiving antenna is P_r given by:

$$P_r = P'_r \frac{G'A}{4\pi R^2}$$

So, in terms of the transmitted power P_t

$$P_r = P_t \frac{GAG'A'}{(4\pi R^2)^2}$$

Now we do not necessarily know, nor do we need to know, the target properties A' and G' separately; it is sufficient to know their product, which is called the radar cross section, σ. The units of σ are square meters.

We then have the basic radar equation:

$$P_r = P_t \frac{GA\sigma}{(4\pi R^2)^2}$$

This is valid for a lossless, isotropic medium, like free space.

P_r is the power available to the receiver, P_t is the power actually radiated by the antenna, R is the range, σ is the radar cross section, G is the gain of the transmitting antenna and A is the effective area of the receiving antenna. In most radar systems the receiving antenna is also the transmitting antenna; the two functions are isolated by the TR switch, described

previously. The ratio P_t/P_r is called the "transmission loss." Expressed in decibels, it is a convenient way of relating the received and transmitted powers.

There is a very basic relationship between the gain G of an antenna used for transmitting and the effective area A of the same antenna used for receiving. It is:

$$G = \frac{4\pi A}{\lambda^2} \text{ or } A = \frac{\lambda^2 G}{4\pi}$$

where λ is the wavelength in meters and A is the effective area in square meters. G, of course, is dimensionless, being a ratio; it is the power gain referred to as isotropic radiator. This relation can be readily proven using a single short dipole antenna and the reciprocity of transmission loss over a path in an isotropic medium like the air.

If the same antenna is used for receiving and transmitting, the radar equation becomes:

$$P_r = P_t \frac{G^2 \lambda^2 \sigma}{(4\pi)^3 R^4},$$

or

$$P_r = P_t \frac{A^2 \sigma}{4\pi \lambda^2 R^4}.$$

Example: Let $A = 10\ m^2$, $\lambda = 0.1$ m, $P_t = 2.5 \times 10^5$ watts (peak pulse),

P_r minimum $= 10^{-13}$ watts, and

$\sigma = 10\ m^2$ (an airplane).

Then

$$P_r = \frac{2.5 \times 10^2 \times 10^2 \times 10}{4\pi \times 10^{-2} R^4} = 10^{-13} \text{ watts}$$

so:

$$R^4 = 0.2 \times 10^{23}$$

and

$$R = 3.76 \times 10^5 \text{ m or } 374 \text{ km.}$$

The maximum detection range for this radar is 374 km. It is this large because of the high power used, the great sensitivity of the receiver, and the high gain of the antenna (G = 12500). The high power can be afforded because it is only used during a small fraction of the time, i.e., when the short pulse is transmitted. If the pulse repetition frequency (p.r.f.) is

400 sec^{-1} and each pulse is one microsecond long, the full 2.5×10^5 watts is used only 1000 microseconds per second of operation; the so called "duty cycle" is 0.001, so that the average power radiated is only 250 watts.

2.3. RADAR EQUATION FACTORS

We shall now discuss in turn the various terms in the radar equation: P_t, P_r, G, A, and the cross section σ.

2.3.1. TRANSMITTER FACTORS. The power available for radiation is symbolized by P_t. This represents the actual electromagnetic power launched by the antenna; the average power input to the transmitter is considerably greater than the average power radiated, due to losses in the transmitter itself as well as losses in the antenna-transmission line system. The "peak power" is the maximum power in the pulse. The "average power" is the peak power times the ratio of the pulse duration to the time between pulses. This ratio is called the "duty cycle"; it is customarily less than 0.001.

A microwave radar transmitter generally uses either a high-power "magnetron" or "Klystron" or traveling-wave-tube to generate the radio frequency pulse. The limit on the peak pulse power that can be obtained is generally set by the available electron emission from the cathode, but sometimes by the area of the anode. The internal dimensions of these tubes must be proportional to the operating wavelength λ. The cathode and/or anode areas, and therefore the available powers, are thus proportional to λ^2; empirically the maximum power, P_{max} is:

$$P_{max} = 10^4 \lambda^2$$

If λ = 10 cm, P_{max} would then be 10^6 watts. Usually less power is used.

2.3.2. RECEIVER FACTORS. Some basic receiver characteristics were discussed in the Passive Microwave chapter. The sensitivity of a practical radar receiver is described by a "noise figure" (F) which represents the equivalent noise input power P_n. When the input is terminated in a resistance equal to the receiver input resistance at temperature T:

$$P_n = kT\Delta\nu F,$$

where T is the temperature of the receiver input circuit, usually the ambient temperature of about 300°K. This noise power is evident in the loudspeaker or headphones of any radio as a continuous hiss as one turns up the volume. The hiss sets a limit to how weak a signal one can detect. Beyond a certain point it makes the radio receiver no more sensitive to turn up the volume control; it merely makes the hiss louder.

The quantity P_r is called the power available or, more precisely, the signal power available from the receiving antenna, i.e., the maximum signal power that can be obtained if there is a perfect match between the antenna and the transmission line. In practice, there is attenuation in the transmission line and the transmit-receive (TR) switch, so that the signal power available at the receiver input is somewhat less than P_r. In other words, for a given signal noise ratio, P_r must be somewhat greater than would be needed if the above attenuation were not present. This is usually included in the design by increasing the noise figure of the receiver by an appropriate factor (1.5 or so).

The equivalent noise power input also includes any microwave noise power that is picked up by the antenna, e.g., $kT_a\Delta\nu$, from a background at temperature T_a. The equivalent receiving system noise input power P_n is:

$$P_n = k\Delta\nu[T_a + (F - 1)T_e].$$

2.3.2.1. Required Bandwidth. A transmitted radio-frequency pulse occupies a frequency bandwidth which depends on the duration of the pulse. The spectrum of the pulse varies with the pulse shape; for a short rectangular pulse, of duration τ, the power spectrum increases in width as τ decreases. Most of the power lies within $\pm 1/\tau$ of the carrier frequency. In order to accommodate this spectrum the receiver needs a bandwidth of $2/\tau$ Hz. For best signal detectability the bandwidth should be about $1.5/\tau$ Hz; this is not much different from the half-power bandwidth of $1.2/\tau$ Hz.

2.3.2.2. Minimum Detectable Signal. The problem of detecting a radar echo amid receiver and background noise differs from the passive microwave measurement problem in that the signal is a portion of a coherent radio-frequency wave, controllable at the transmitter in phase, pulse rate, amplitude, polarization and frequency. With this knowledge as to the kind of echo signal we are likely to receive comes the possibility of using sophisticated processing techniques to optimize a wanted signal with respect to unwanted and random echoes and noise.

Consider first a simple square coherent pulse received amid random noise. Let V_s be the rms value of signal voltage during the pulse, and V_n the rms value of the noise voltage. As in the passive microwave case, normalize the instantaneous noise voltage V to V_n, and also normalize V_s in the same way:

$$\rho_n = \frac{V}{V_n} \qquad \rho_s = \frac{V_s}{V_n}$$

We may now also use the phase of V_s as a reference, and express ρ_n as a complex quantity:

$$\bar{\rho}_n = \rho_n \exp(i\phi)$$

where ϕ is the phase of $\bar{\rho}_n$ relative to that of ρ_s. As before, $\bar{\rho}$ by itself has a bivariate normal distribution.

Let us designate this vector sum of the normalized signal and noise voltage, $\rho_s + \bar{\rho}_n$, as $\bar{\rho}$, and its magnitude as ρ. Then $\bar{\rho}$ has a similar distribution except that the center is displaced along the real axis as shown in figure 2-1.

Expressed in polar coordinates, this is no longer a Rayleigh distribution. The average ρ is larger but ρ still varies from zero to some large number. One cannot really distinguish in a single trial between the case with a signal and that without. If the signal is made very strong, so that there is a high probability of finding a large value of ρ, the probability of detection with a single pulse becomes good.

If many echoes, obtained on a large number of successive trials (pulses), are averaged, estimates of the mean value of the noise and the signal-pulse-noise can be obtained, and the detection probability thus improved. The standard error of the estimated average, σ_m, is inversely proportional to the square root of the number of trials. If the standard error is small enough, compared to the difference between the noise and the signal plus noise, detection is then possible. When one observes the indicator (oscilloscope) of the radar, one's eye integrates the signals in this way. A single radar trace may show no trace of an echo (fig. 2-2a), but many traces, one on top of another, will show a distinct radar return (fig. 2-2b).

For a single pulse, the signal power must be 10 times the noise power in order to have a good probability of detection; for many pulses averaged or integrated, the signal can be detected if its power is equal to, or even less than, the noise power.

We may thus take the minimum received power P_r to be of the order of the noise level ($kT\Delta\nu F$) provided that we integrate over many pulses. For very high probability of detection, however, the signal may need to be several times the noise power. The ratio of signal power to noise power is called the signal noise ratio S/N.

2.4. ANTENNA GAIN

The gain of a directive transmitting antenna, as noted above, is a measure of the concentration of power, in the main beam of the antenna, that would otherwise be spread over 4π steradians (fig. 1-2). Let $f(\psi, \phi)$ be the ratio of the intensity of the source, in watts per steradian, in a direction described by the angles ψ and ϕ, to the intensity, I, at the center of the main beam ($\psi, \phi = 0$), i.e., $f(\psi, \phi)$ is the "directive power pattern" of the antenna. The total radiated power is then I times the integral of $f(\psi, \phi)$ over 4π steradians:

$$P_t = I\int_{4\pi} f(\psi, \phi)d\Omega$$

If this radiated power were spread evenly over 4π steradians due to an isotropic antenna radiating the same total power, the isotropic intensity would be

$$I(\text{isotropic}) = \frac{P_t}{4\pi}.$$

Thus,

$$4\pi I(\text{isotropic}) = I\int f(\psi, \phi)d\Omega.$$

The power gain of the antenna in the center of the beam is

$$G = 4\pi\Big/\int f(\psi, \phi)d\Omega.$$

To obtain the relation between the gain and the effective area of an antenna, let us recall the relation, demonstrated in the Passive Microwave Section, that the half-power beam width θ of a linear antenna array of length D is:

$$\theta = \frac{\lambda}{D} \text{ radians.}$$

If we have a square array of antennas, D on a side, the half-power beam area has approximately a square shape, subtending a solid angle of θ^2 steradians. The gain G is then approximately:

$$G = \frac{4\pi}{\theta^2} = \frac{4\pi D^2}{\lambda^2}.$$

Since the physical area of the array A is equal to D^2, we can write

$$G = \frac{4\pi A}{\lambda^2}.$$

The A defined by this relationship is really the "effective area," which is also the area used in calculating the power received by the same antenna. The actual area of a paraboloidal antenna is somewhat greater, in part owing to the tapered illumination, i.e., the edge of the dish is not as brightly illuminated by the feed horn at the focus as in the central section.

It is interesting to note that the effective area of a dipole antenna which is a half wavelength long is approximately equal to that of a square whose diagonal is the length of the antenna. The effective area of a dipole antenna much shorter than a half wavelength is:

$$A = \frac{3}{8\pi}\lambda^2$$

independent of the length of the dipole!

For the frequently used circular aperture paraboloid antenna of half-power beam diameter θ, including the effect of taper, the gain is given approximately by:

$$G = \frac{30000}{\theta^2}$$

where θ is measured in degrees. This is a satisfactory approximation so long as θ is smaller than 20°.

2.5. IMPROVEMENT OF RADAR DATA

It is evident from the previous discussion in this section that the resolution of a conventional radar picture usually depends on the beam width and the pulse length. A vertical mapping mode similar to the one used by the airborne infrared scanner would have a resolution limited by the two dimensional beam width only. Such a mode is only of marginal interest because of the large antenna size needed to obtain a narrow beam at radar wavelength. The angular resolution is λ/D radians, or λ R/D meters at a range R. The range resolution, one of the unique features of radar, is $c\tau/2$, where τ is the pulse duration. The following constraints on resolution exist:

1. The antenna dimensions D is limited by the size of the aircraft involving economic considerations.
2. The wavelength λ is limited by atmospheric transmission (absorption increases for $\lambda < 3$ cm).
3. Larger wavelengths must be used if we are to see through foliage, or into the soil.
4. R must be large, for we are interested in remote sensing.
5. If the pulse duration is shortened to improve the range resolutions, the necessary band width $\Delta\nu$ increases and so does the noise level, $kT\Delta\nu$.

Two techniques are used to improve the situation, the use of "chirp" or a frequency modulated pulse, to make a long pulse look like a short one, and "synthetic resolution" to make a short antenna act like a long one.

2.5.1. CHIRP.

A long radar pulse produces a narrow spectrum and a short pulse produces a wide spectrum. The spacing between half-power frequencies for a pulse of a given shape is inversely proportional to the pulse duration. An increase in the band width increases the information that is available, and this information can be used to obtain finer range resolution.

For example, if two targets of approximately equal cross section are separated in range by ΔR, the relative phase of the echoes differ by:

$$\phi = \frac{2\pi}{\lambda}\Delta R = \frac{2\pi}{c}\nu\Delta R$$

When a pulse is transmitted with increasing frequency ν, the reflected irradiance at the antenna due to the two targets together varies periodically. The complete spectrum of the reflected power contains information from which the precise spacing of the targets can be found.

We can make a long pulse cover a large frequency band by varying the radio frequency with time during the pulse. For example we might use a one microsecond pulse which starts at 9050 MHz and ends at 9000 MHz. The frequency varies at the rate of 50 MHz per microsecond, or 5×10^{13} Hz per sec; it occupies a 50 MHz band width, like a 0.04 microsecond pulse, but has a pulse energy as great as a one microsecond pulse. This is called a "chirped" pulse, since, if it were done at audio frequencies it would sound like a chirp.

On reception, the echo is "dechirped" by passing it through a dispersive network, in which the time of transit varies with frequency. All the energy spread over the frequency band occupied by the chirped pulse thus comes out of the network at the same time, i.e., in a 0.04 microsecond period for the pulse described above. The range resolution has thus been improved by a factor of 5. Using this technique we can obtain a very high average power from the radar, without exceeding the peak power limitations of the transmitter.

2.5.2. SYNTHETIC RESOLUTION TECHNIQUES. It will be recalled that the expression for the beamwidth θ of a linear antenna array of length D is (approximately)

$$\theta = \frac{\lambda}{D} \text{ radians}, \qquad (1)$$

at a wavelength λ. If a radar antenna of length D were to be mounted on an aircraft and directed to the side (normal to the aircraft motion), an isolated target on the ground at a distance R from the antenna would produce an echo on the radar output display with an along-track dimension approximately equivalent to the distance the aircraft moves while the target is "in the beam," i.e., between the half-power points. This distance varies directly with the target range. If two targets are at the same range and are separated by less than this distance, the radar cannot "resolve" them as being two separate targets but sees them as one. Thus, the ground resolution along the flight line of an antenna of length D at a wavelength λ is, under these conditions,

$$\text{Resolution} = \frac{\lambda R}{D} \qquad (2)$$

The conventional technique for achieving better resolution along the flight line is to provide a longer antenna, thus creating a narrower beam, or by operating at a shorter wavelength, or both. In the case of airborne radars, however, physical considerations limit the resolution which can be obtained in this way. An additional constraint on resolution achieved by this means if that equation (1) is correct only in the "far field" of the antenna. The far field is usually regarded as beginning at a range

$$R_{ff} \cong \frac{D^2}{\lambda}. \tag{3}$$

If this expression is substituted into equation (1), it can be seen that the minimum resolution obtainable is D, the physical length of the antenna.

The above considerations make the possibility of achieving really good azimuth (along the flight line) resolution with a radar appear remote. By "really good" is meant of the order of $\lambda/2$, which is regarded as the theoretical resolution limit.

Within the last fifteen years or so, however, a technique has emerged which circumvents the above constraints, and permits finer azimuth resolutions to be achieved. This technique has been given the name "synthetic resolution."

Suppose that instead of having a physical antenna of length D, a short antenna is placed at successive positions along the line that the large antenna would have occupied if it existed. (See figure 2-3.) At each position (numbered 1 through 12 on the figure), the small antenna is used to transmit and receive radar echoes from all the targets within its broad beamwidth. The target information is recorded or stored, in both phase and amplitude for later processing.

It seems reasonable to suppose that by appropriately processing the stored information obtained in this way a synthetic beamwidth could be created equivalent to that which would have been obtained by using the much longer antenna of length D. In fact, as will be seen, the synthetic beamwidth thus created by appropriate processing can be made equivalent to an antenna whose length varies with range, such that, in theory, the resolution obtainable is independent of the range to the target.

The broad beamwidth of the real antenna is

$$\beta_r = \frac{\lambda}{D} \text{ radians} \tag{4}$$

The distance across this antenna beam at a radar range of R is $\beta_r R$. If a target, or group of targets, is located at range R, data from them can be collected and stored by the small antenna as it is moved over a distance of βR, from one "edge" of the real antenna beamwidth to the other. It can be seen that this distance increases directly with range: the greater the range, the more target echoes are collected and stored for later processing. Thus the synthetic antenna length $\mathcal{L}_s$ is:

$$\mathcal{L}_s = \beta_r R. \tag{5}$$

Before calculating the resolution obtainable with this synthetic antenna, it is necessary to point out an essential difference between a synthetic and a real antenna. Consider a physical array of length D made up of an array of elementary dipole antennas uniformly energized. All the dipoles radiate simultaneously and receive simultaneously, each one receiving the energy reflected from the target which was originally radiated from all the dipoles. In this case, the antenna beamwidth is given by λ/D. Now consider a synthetic antenna of the same length D, but made up not of a number of dipoles radiating simultaneously but by a single dipole which is moved successively from point to point along a line of length D, each point corresponding to the position of a dipole in the physical array. At each point the dipole is caused to radiate and receive radiation from the targets. The received signal is stored (in amplitude and phase) at the location of each position the dipole occupied during its movement from one end of D to the other. This information differs from that obtained by the real array of dipoles, in that in each position the dipole receives and stores only that information contributed by its own radiation. In this case, the antenna beamwidth (synthetic) is given by $\lambda/2D$. Thus the beamwidth of a synthetic array is half that of a real antenna of the same length, or equal to that of an antenna twice as long. Using equations (4) and (5) with this result, the beamwidth (or resolution) of a synthetic antenna can be expressed by

$$\beta_s = \frac{D}{2R}. \tag{6}$$

This indicates that the synthetic beamwidth β_s is independent of frequency, and decreases with range.

To provide a measure of resolution, it will be assumed that two objects can be resolved if they are one beamwidth apart. For this case, at a range R, the distance across the beam is:

$$\mathcal{L} = \beta_s R = \frac{D}{2} \tag{7}$$

Thus (in theory), the azimuth resolution of a synthetic resolution radar is independent of range, radar wavelength, and smaller than the physical size of the real antenna used to create the synthetic one.

Now one might expect that one can't get all these goodies for nothing. In the case of synthetic resolution the price can be quite high. The antenna must be moved relative to the targets in a very precise straight line. As an example, suppose the real antenna beamwidth is 1/10 radian, λ = 10 cm, and R = 10 km. Then $\mathcal{L}_s = \beta_r R = 1/10 \times 10^4 = 1000$ meters long. In order for the synthetic resolution antenna to be sufficiently straight to achieve the theoretical resolution, it must not deviate from a straight line anywhere more than $\lambda/4$, or 2.5 cm in 1 km of travel. This degree of precision cannot be achieved in an aircraft flying through the normal

turbulence of the atmosphere. The alternative is to measure deviations from the desired flight path and appropriately correct the phase of the radar information prior to storage. This can be done, but it adds considerably to the complexity and cost of the equipment. The high degree of phase coherence required of the radar equipment and the requirements for accurate information storage and readout also add to the difficulties of the technique. At the time of the present writing, synthetic aperture radar is regarded as a highly sophisticated radar technique.

References: L. J. Cutrona, W. E. Vivian, E. N. Leith, G. O. Hall, A High Resolution Radar Combat Surveillance System, IRE Transaction on Military Electronics, V. MIL-5, No. 2, April 1961, pp. 127-131.

C. W. Sherwin, J. P. Ruina, R. D. Rawcliffe, Some Early Developments in Synthetic Aperture Radar Systems, IRE Transactions on Military Electronics, V. MIL-6, No. 2, April 1962, pp. 111-115.

L. J. Cutrona, G. O. Hall, A Comparison of Techniques for Achieving Fine Azimuth Resolution, ibid. pp. 119-121.

2.6. RADAR CROSS SECTIONS

The radar cross section σ may be regarded as the product of the effective area A' of the target, projected on a plane normal to the direction of the arriving radar signal, and the gain G' of the target as a radiator in the direction of the receiving antenna. These factors depend on the reflectivity and roughness of the target, as well as its geometry and orientation, and also the wavelength of the signal. Normally the receiving antenna of a radar is also the transmitting antenna. This is called the "monostatic" case, as against the bistatic case, where they are separated. The cross sections we shall discuss here are the monostatic cross sections.

In remote sensing, the radar cross section and its properties yield information regarding the nature of the target. If the target is large, like a patch of terrain or trees, for example, the radar cross section observed will be an average over the illuminated patch, and the received echo may be made up of many small echoes differing in amplitude and phase. Smaller, discrete targets, on the other hand, will result in more coherent echoes, which may be interpreted in terms of geometry, size, material, etc. The radar cross section can often be estimated or interpreted in terms of combinations of one or more simple geometrical objects. We shall next discuss some of these objects.

2.6.1. A FLAT METAL PLATE. If a flat metal plate of area A is oriented as in figure 2-4a normal to the direction of arrival of the signal, so that the currents in it are all in phase, it radiates like an array of dipoles and thus has a gain of $4\pi A/\lambda^2$. The cross section is

$$\sigma = GA = 4\pi A^2/\lambda^2$$

Note that σ is not equal to A. Suppose the plate is 50 cm square and λ is 10 cm. Then A = 0.25 m^2 so that:

$$\sigma = 4\pi \times \frac{0.0625}{0.01} = 78 \text{ square meters}$$

Just as for the antenna array, the reflected power has a directional pattern; for any direction other than straight back to the radar, this would represent the bistatic cross section. It is referred to as the "diffraction" pattern.

If the plate is not perpendicular to the direction of the radar, as in figure 2-4b, the path length is different to different parts of the plate, and the induced currents are no longer in phase. Their phases now lag by $\left(\frac{2\pi x}{\lambda}\right) \sin \gamma$ where x is the distance along the plate from the nearest edge, and γ is the angle of incidence. The plate now behaves like an array with continuously varying phase, with the result that the maximum of the diffraction pattern is shifted away from the radar by the angle 2γ. The beam width of the pattern is also slightly increased (by the factor sec γ). As seen from figure 2-4b, the echo received at the radar now comes from a side lobe of the diffraction pattern. There is thus some echo even if the angle is not just right, but not very much.

Note that the direction of the center of the maximum of the diffraction obeys the well known law of reflection of light. In point of fact, the reflection of a ray of light from a mirror does have a diffraction pattern, but, with an extremely narrow main beam, since γ is usually no greater than 10^{-5} D in this case. (10^{-5} radians is only 2 seconds of arc.)

Flat plates are sometimes not very flat; their deviations from flatness may be comparable with a wavelength, or greater. This is shown in figure 2-4c, where one path (2) is longer than the other path (1) by 2Δ, where Δ is the depth of the irregularity. The currents induced in the plate by waves arriving over the two paths differ in phase by $\frac{2\pi}{\lambda}\Delta$. By the time the reradiated (reflected) waves have made the round trip to the radar, their phases differ by twice this, or $\frac{4\pi}{\lambda}\Delta$. The resultant is less than if they had been in phase, by the factor $\cos \frac{2\pi\Delta}{\lambda}$.

In order to be satisfactory as a reflector (e.g., the reflecting surface of paraboloid antenna) the surface must be smooth to 1/8 of a wavelength or better. This is called the Rayleigh criterion of roughness.

2.6.2. CORNER REFLECTORS. If two flat metal plates are joined at right angles, they form a dihedron or a two sided "corner reflector." A ray in the plane perpendicular to the line of intersection of the plates that hits the two plates successively will be reflected back in the direction whence it came, somewhat like a corner shot in billiards. This is illustrated in figure 2-5. The condition that a ray is doubly reflected may be represented by the construction in figure 2-6. The dashed lines represent the image of each side reflected in the other, while the dotted lines represent the extension of rays beyond their reflection points.

Rays 1 and 2 undergo two reflections each and are reflected back. Ray 3 is reflected only once, going off in a different direction, while ray 4 misses the reflector altogether. From the geometry, it is obvious that a ray can only be reflected back if it or its extension pass through a face of the corner reflector (e.g., AO) and the image of its image (OB'). The width of the aperture is thus limited to the projection of length BB' perpendicular to the direction of the arriving rays. More precisely, it is the smaller of the projections of BB' and AA', depending on the orientation of the reflector.

Once again, note that this "retroflection" applies only to rays perpendicular to the intersection of the metal plates. Also, the flatness of the plates must be adequate ($<0.125\lambda$) and the right angle must be precise, i.e., the deviation from 90° must be less than $1/2\ \lambda \div$ projection of BB' perpendicular to the ray. If the reflector is made of two plates, each ℓ long and w wide, joined along the ℓ dimension, the maximum effective area A is:

$$A = \sqrt{2}\,\ell w$$

and the cross section σ is as usual, $4\pi A^2/\lambda^2$.

This type of two sided corner reflector, called a "dihedron," is useful in describing the radar appearance of some objects. Because the retroflection is limited to one plane (normal to the intersection) it is not particularly useful in calibration of radars.

If three perpendicular metal plates are used to form the corner reflector, the ray need no longer be in a plane perpendicular to an intersection to be reflected back to the radar. A ray now is reflected once off each plate (fig. 2-7). This is the principle of the "cats-eye" reflector on the highways, for example; these and similar devices are called "retroflectors." A ray behaves something like a ball in the corner of a handball court.

It is difficult to depict, in two dimensions, the geometry of a triply reflected ray. For ray tracing, to find the effective area, one needs to consider seven image corners (fig. 2-8). Only one of these is of the right kind for a ray to be reflected back, the corner diagonally opposite the one being considered.

Consider a cubical box, with each edge twice as long as an edge of the corner reflector (fig. 2-9a). This box will contain all the images. Now cut off one corner as in figure 2-9b. This corner is the size and shape of the reflector. Then cut off the opposite corner, likewise, as in figure 2-9c. It is an image of its original corner. If now one looks through the missing corners of the box, the area that one can see through is the effective area of the corner reflector (fig. 2-9d). The maximum area appears when the line of sight is right through the center of the original box, so that the two missing corners are lined up. The area is a hexagon, each side of which is $\ell\sqrt{2/3}$ long, where ℓ is the length of an edge of the reflector.

The maximum effective area A of the corner reflector is in a direction making an angle of 35.3° with each side. It is:

$$A = \sqrt{3}\ell^2$$

and the radar cross section is

$$\sigma = 4\pi\frac{A^2}{\lambda^2} = \frac{12\pi\ell^4}{\lambda^2}$$

If, for example, ℓ = 1 meter and λ = 0.1 meter, σ is 3770 square meters (the actual area is only 1.73 square meters). The area diminishes if they are turned (fig. 2-10).

Corner reflectors must be made carefully, just as in the case of dihedrons. The total error in such a reflector must not exceed $\lambda/8$. The allowable error on any part must be less than this - perhaps $\lambda/16$.

Corner reflectors are often made with triangular sides, as shown in figure 2-11, but this causes a loss in cross section by a factor of 3.

For general use, when the direction of the radar is not specified, for example, a practical corner reflector is made from three planes as shown in figure 2-12. In this case we have actually made all the images. This works well in nearly all directions. For certain specific orientations, it may behave like a dihedron, or like a flat plate if the direction to the radar is perpendicular to one of the intersections.

Corner reflectors are used for standards of radar cross section, for calibration of radar sets, and also as markers; e.g., for balloon tracking, or to mark places on the ground so that they can be found in radar pictures.

2.6.3. SPHERES. Another kind of standard radar reflector is a metallic sphere. Because of its simple geometry it is easy to calculate its exact radar cross section, although the cross section is relatively small. A large sphere, several wavelengths in diameter, may be regarded as an isotropic scatterer, i.e., the power it intercepts from a wave is reradiated equally in all directions. The gain G is therefore unity, like an isotropic radiator, while the projected area A is πr^2, where r is the radius. Thus the cross section $\sigma = \pi r^2$, or just the projected area of the sphere on a plane perpendicular to the arriving ray.

The monostatic radar cross section of a sphere of radius r is shown in figure 2-13 as a function of r/λ. For $r \ll \lambda$

$$\sigma \doteq 9\left(\frac{2\pi r}{\lambda}\right)^4 A$$

This is called the "Rayleigh scattering" region. For $r \ll \lambda$, $\sigma = A$ as above. In the intermediate region for r the graph shows resonance effects, with an amplitude which decreases with increasing r or decreasing λ.

Spheres of water, like raindrops, reflect radar signals in the same manner. The radii of such drops are generally in the Rayleigh scattering region, for normal radar frequencies. For such reflectors:

$$\sigma \doteq 4\left(\frac{2\pi r}{\lambda}\right)^4 A$$

Scattering by clouds is negligible at the usual radar wavelengths. For wavelengths less than 10 cm, however, there is significant scattering by rain. Rain areas can be seen and mapped by a radar using such wavelengths. Since rain is generally associated with cloud and wind patterns, radar measurements are quite useful in meteorology. High frequency airborne radars are also used to determine weather conditions lying ahead on the flight path.

The radar cross sections of many small or rounded objects may be estimated by replacing them with a small sphere. Birds, insects, spacemen, etc., can be replaced by a sphere of water of equal weight. A metallic object like a bullet can be replaced by a metallic sphere of equal volume.

2.6.4. DIPOLES. A wire one-half wavelength long is resonant, i.e., if illuminated by an electromagnetic wave polarized in the direction of the wire, a relatively large current is induced in it. This current will produce radiation in accordance with the directional pattern of the wire. Figure 2-14 shows such a wire; the dashed lines indicate the reradiated pattern. The gain G for such a resonant dipole is 1.64, found by integrating the power pattern over 4π steradians. The effective area is accordingly:

$$A = 1.64\frac{\lambda^2}{4\pi} = 0.1305\lambda^2$$

Since ℓ, the length of the dipole is $\lambda/2$,

$$A = 0.522\ell^2$$

or about half the area of a square whose edge is ℓ long. The monostatic radar cross section of the resonant dipole is therefore

$$\sigma = 1.64 \times 0.522\ell^2 = 0.857\ell^2$$

The cross section of a dipole like this decreases sharply as the frequency (or length) departs from the resonant condition.

A large number of such dipoles, made of light aluminum foil, can be thrown out of an airplane to mark a spot in the air.

2.6.5. A LARGE THICK CYLINDER

A large cylinder of diameter D and length $\mathscr{L}$ has an area A = D$\mathscr{L}$ and a gain G = $2\mathscr{L}/\lambda$. Its radar cross section is thus:

$$\sigma = GA \doteq \frac{2D\mathscr{L}^2}{\lambda}$$

For example, the cross section of a tree might be estimated by considering the trunk as a cylinder and neglecting the top part. The tree may be considered as made largely of water. The cross section would be no more than $2D\mathscr{L}^2/\lambda$.

It is very difficult to calculate the exact radar cross section of an object. A reasonable estimate of the cross section, however, can often be made by analogy to the simple objects - a sphere, a cylinder, a dipole, a flat plate, and a two or three sheet corner reflector. An approximate cross section may be synthesized from a wisely chosen set of these simple objects, of appropriate sizes.

2.6.6. REFLECTIONS FROM MULTIPLE OBJECTS. When more than one object exists in the field of view illuminated by a radar, the resultant radar cross section is not in general equal to the sum of the separate cross sections. The reason for this is that the power received at a given instant depends on the relative phases of the reflections from the object. Given two identical spheres, for example, the received voltage can be anywhere between zero and twice the voltage due to a single sphere, but the power received would be zero to four times that for a single sphere. Voltages add linearly; cross sections, which are proportional to the squares of the voltages, do not.

To convert the cross section σ into a quantity proportional to the voltage, one must take the square root. Thus, if all the reflections have the same phase, the effective cross section σ_E is given by:

$$\sqrt{\sigma_E} = \sqrt{\sigma} + \sqrt{\sigma}$$

But the voltages add like vectors and may not be in phase, in which case:

$$\sqrt{\sigma_E} = \left| \sqrt{\sigma} e^{i\phi_1} + \sqrt{\sigma} e^{i\phi_2} \right|$$

where the ϕ's are the phases of the reflected voltages. If $\phi_1 = \phi_2$, i.e., the echoes are in phase:

$$\sqrt{\sigma_E} = \left| 2\sqrt{\sigma} e^{i\phi_1} \right| = 2\sqrt{\sigma} \text{ and } \sigma_E = 4\sigma$$

If $\phi_1 = 0$ and $\phi_2 = 180^o$

$$\sqrt{\sigma_E} = \sqrt{\sigma} - \sqrt{\sigma} = 0 \text{ and } \sigma_E = 0$$

If we have no information about the relative phase, or when the spheres have relative motion, a sort of an average phase would be 90^o and $\sigma_E = 2\sigma$. Thus, if the phases are random, the σ's can add linearly, i.e., $\sigma_E = 2\sigma$, giving an average cross section.

The echo from one raindrop, for example, is small, but one can see millions of raindrops at one instant. The echo from n raindrops is n times as great, on the average, as that from one, although at any instant the echo can vary from zero to a very high number. Here we again meet the phenomenon of adding many small signals with random phase, as we did in the discussion of noise, and the result is a Rayleigh distribution, provided n is a large number.

$$\sqrt{\sigma_E} = \sqrt{\sigma} \sum_{i=1}^{n} e^{j\phi_i} \text{ where } \phi_i \text{ is a random angle.}$$

The most probable value of the voltage is the root mean square value; the most probable value of σ_E is then the sum of the individual cross sections:

$$\sigma_E = \Sigma\sigma_i$$

This also applies to collections of objects with different cross sections, e.g., trees, bushes, a corn field, or rough terrain. The radar reflection in this case is said to come from a "patch" or resolution element, defined by the pulse duration τ and the beam width at the range of the patch, for the vertical angle the arriving radar signal. Figure 2-15 shows the geometry. If the vertical beam width is wide and the arrival angle ϕ is not too great, the patch area A_p is given by:

$$A_p = R\theta \cdot \frac{c\tau}{2 \cos \phi}$$

where R is the range, θ the beam width, c the velocity of light, and τ is the pulse duration.

The cross section $\bar{\sigma}$ is $\sigma_o A_p$, where σ_o is the average cross section per unit patch area, for the angle ϕ.

In the case described above, where the area giving the radar echo is limited by the beam width and pulse length, rather than by the total area of the ground, the terrain is called an "extended target." Since one dimension of the patch is proportional to the range R, the received echo power is proportional to R^{-3} rather than R^{-4} as in the basic radar equation. If

the length of the patch were limited by the vertical beam width, and thus were also proportional to R, the range-dependence of the received power would be only R^{-2}. This would apply, for example, to radar exploration of the surface of a planet from a spacecraft.

If there are large objects, like houses or dense trees, in a patch, they will cast radar "shadows" on objects or terrain back of them. See figure 2-16. In this case the patch will not return as large an echo as it would were the shadowing object not there. In the example shown, the corner reflector formed by the ground and the wall of house B is not illuminated; such a corner is the major source of an echo from a house. With shadowing, the σ of the two houses is obviously not the sum of the individual σ's.

Another condition under which the resultant cross section is not equal to the sum of the individual cross sections exists if there are multiple reflections. Figure 2-17 illustrates how, although neither mirror m_1 nor m_2 would separately reflect a beam back to the transmitter, together they form a corner reflector, which reflects the beam back very strongly. In cases where reflection from one object to another can be expected, care must be used in interpreting the radar echoes. The reflection can be much greater than the sum of its parts.

2.7. POLARIZATION EFFECTS

Polarization Types

Recall from the discussion on fundamental principles that the electromagnetic wave is a transverse wave when propagating in free space. The direction of polarization is defined as the direction of the electric field which is transverse to the direction of propagation. A wave is horizontally polarized when the electric field is in a direction so as to be parallel to a horizontal plane. A wave is vertically polarized when the electric field is at right angles to the horizontally polarized direction. These two polarizations are shown in figure 2-18. An elementary radio antenna as shown in these figures emits radiation with a polarization in accordance with the orientation of the antenna.

When two antennas with horizontal and vertical orientations are used, the resulting polarization is a superposition of these two polarizations. The electric fields of each wave adds vectorially to produce a polarization at 45 degrees to the horizontal provided that the frequency and phase of the radiation is the same. If the amplitude of one of the waves is less than the other, then the polarization is linear but at some other angle than 45 degrees from horizontal when the phases are the same. When the frequency is the same but the phase is changed between the two waves by separating the antennas by a quarter of a wavelength, then circular polarization results from the vector addition of the electric fields of the two waves. These polarizations are illustrated in figure 2-19. The polarization is elliptical if the phases are different. Thus, it can be seen that different kinds of polarizations can be generated by the

superposition of electromagnetic waves propagating in the same direction with the same frequency by adjusting the amplitudes and relative phases of two orthogonally polarized waves.

All possible polarization types can be neatly charted upon the surface of a sphere. Such a chart is called the Poincaire sphere. It is shown in figure 2-20. The point at the north pole represents right circular polarization. Right circular polarization is that polarization for which the electric vector direction rotates in the direction of the curved fingers of the right hand when the right thumb is pointed in the direction of propagation of the wave and as one advances along the wave. Left circular polarization is defined similarly using the left hand. Graphically, the vector diagram of a right circularly polarized wave would look like a right hand screw with the tips of the electric vectors forming the sharp edge of the screw thread. The circular polarized wave of figure 2-19, is of the right hand type. Each specified point on the Poincaire sphere has a corresponding point diametrically opposite called the antipode. The polarization type at the antipode is the exact "opposite" or orthogonal polarization type to the polarization type represented by the specified point. Thus, right circular and left circular are opposite or "orthogonal" sets of polarizations. Horizontal is the opposite of vertical, 45 degrees linear polarization tilted toward the right side is opposite to 45 degrees linear tilted toward the left. (Artistic license was used to emphasize orthogonality. The chart is technically incorrect here.) All right hand elliptical polarizations are represented by points in the northern hemisphere with the corresponding antipodal left elliptical polarizations in the southern hemisphere. Any other polarization type can be made by the superposition of two waves with orthogonal polarizations by choosing the appropriate phases and amplitudes. Thus, a wave having any kind of polarization can be considered as consisting of two component waves with orthogonal polarizations. It is often convenient to choose certain polarization components which are best suited to the analysis of a problem involving the interaction of radiation with matter. The resulting interaction using these components separately and recombining the resultants will be the same as the resulting interaction using the single wave having these components. It is often easiest to see the results using vertical and horizontal polarization components. Thus, every wave incident upon matter can be considered to be composed of vertical and horizontal polarization components and similarly for every reflected wave. Thus, the general interaction can be represented as

$$\mathcal{E}_{H,R} = A_{11}\mathcal{E}_{H,I} + A_{12}\mathcal{E}_{V,I}$$

$$\mathcal{E}_{V,R} = A_{21}\mathcal{E}_{H,I} + A_{22}\mathcal{E}_{V,I}$$

where $\mathcal{E}_{H,R}$ is the horizontal component reflected,

$\mathcal{E}_{V,R}$ is the vertical component reflected,

$\mathcal{E}_{H,I}$ is the horizontal component incident
$\mathcal{E}_{V,I}$ is the vertical component incident.

The values A_{ij} represent the action of the type of material and its geometry upon the amplitude of the reflected wave.

If we know what has been transmitted and hence what is incident and if by reception we measure the reflected amplitudes then we can establish the values of the A_{ij} which will be complex numbers to identify the magnitude and phase of the reflected wave relative to some zero phase angle to be set. Since the zero phase depends upon when one chooses to begin timing, one can set the zero phase arbitrarily. Therefore, one can choose one of the phase angles of the incident waves to be the zero phase as convenience indicates. Now, by the reciprocity theorem, an object which can change incident horizontally polarized radiation into reflected vertically polarized radiation can as readily change incident vertically polarized radiation into reflected horizontally polarized radiation. Thus the value of A_{21} must be the same as the value of A_{12}. There are two additional restrictions on the values of A_{ij}. For passive reflectors the reflected power must not exceed the incident power. For horizontally polarized incident waves:

$$\mathcal{E}_{H,R} = A_{11}\mathcal{E}_{H,I}$$

$$\mathcal{E}_{V,R} = A_{21}\mathcal{E}_{H,I}.$$

The incident power is proportional to $\mathcal{E}_{H,I}\mathcal{E}_{H,I}^*$ and the reflected power $\mathcal{E}_{H,R}\mathcal{E}_{H,R}^* + \mathcal{E}_{V,R}\mathcal{E}_{V,R}^*$ so that $A_{11}A_{12}^* + A_{22}A_{22}^* \leq 1$. Using only incident vertically polarized waves one finds similarly that

$$A_{12}A_{12}^* + A_{22}A_{22}^* \leq 1.$$

The magnitude of A_{ij} must therefore be less than or equal to one and not entirely independent of each other.

One can deduce what the A_{ij} values should be for some simple materials and structures by examining the character of the reflection in detail. First, let us limit our reflections to those which reflect backwards in the direction of the transmitter from perfectly conducting material. No power will be lost so that

$$A_{11}A_{11}^* + A_{21}A_{21}^* = 1, \text{ and}$$

$$A_{12}A_{12}^* + A_{22}A_{22}^* = 1.$$

2.7.1. FLAT MIRROR. When an electromagnetic wave is incident upon a flat conductor, that component of the electric field which is parallel to the surface is reflected in opposite

sense but that component of the incident field which is perpendicular to the surface is reflected in the same sense. See figure 2-21. If the wave is normally incident, the electric field, being transverse, is entirely parallel to the surface so that both horizontal polarization and vertical polarization reflect in opposite sense. See figure 2-22. Therefore,

$$A_{11} = -1,\ A_{21} = 0$$

$$A_{12} = 0,\ A_{22} = -1,$$

and the values of the A_{ij} are real numbers in this case. Thus,

$$\mathcal{E}_{H,R} = -1\mathcal{E}_{H,I} + 0\mathcal{E}_{V,I}$$

$$\mathcal{E}_{V,R} = 0\mathcal{E}_{H,I} - \mathcal{E}_{V,I}.$$

From the above, we can see that a flat conducting plate retroreflects horizontally polarized radiation when horizontally polarized radiation is incident normally and retroreflects vertically polarized radiation when vertically polarized radiation is incident normally. Polarization type is not changed but the phase is changed in both by π radians (or 180^{o}).

2.7.2. VERTICAL DIHEDRON. A vertical dihedron is formed by two conducting planes intersecting at right angles with the line of intersection perpendicular to the horizontal plane. It can be seen from the reflection rules for electric fields and from figure 2-23 that horizontal polarization changes phase by π radians but vertical polarization reflects with the same phase as the incident wave. Thus, for a vertical dihedron the A_{ij} values are

$$A_{11} = -1$$

$$A_{12} = 0$$

$$A_{21} = 0$$

$$A_{22} = +1.$$

It is a convenient short notation to list the values of the A_{ij} in an array where the position in the array identifies its position in the above equations. That is,

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

For the interactions of a vertical dihedron, one can list,

$$\begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}$$

2.7.3. DIAGONAL DIHEDRON. Suppose we wished to ascertain the effect in polarization of a vertical dihedron for 45 degrees tilted to the right polarized radiation we can consider that such radiation is equivalent to two polarization components of equal amplitude, one vertical and one horizontal and in phase. The incident wave has amplitude $\mathcal{E}_{I,45}$. Thus,

$$\vec{\mathcal{E}}_{I,45} = \vec{\mathcal{E}}_{H,I} + \vec{\mathcal{E}}_{V,I}$$

where the magnitudes are

$$\mathcal{E}_{H,I} = \frac{1}{\sqrt{2}} \mathcal{E}_{I,45}$$

$$\mathcal{E}_{V,I} = \frac{1}{\sqrt{2}} \mathcal{E}_{I,45}$$

By substitution

$$\mathcal{E}_{H,R} = -\frac{1}{\sqrt{2}} \mathcal{E}_{I,45}$$

$$\mathcal{E}_{V,R} = +\frac{1}{\sqrt{2}} \mathcal{E}_{I,45}.$$

Upon reflection the horizontal component of the reflected wave has changed phase by π radians but the vertical component of the reflected wave has not changed phase. Therefore, the reflected wave has changed to a wave polarized 45 degrees tilted to the left.

It is easy to see that a similar argument can be made for the effect on vertical polarization by a dihedron whose axis is slanted relative to the vertical by 45 degrees.

Vertically polarized waves incident upon a dihedron slanted at 45 degrees to the vertical would reflect horizontally polarized waves. Therefore the A_{ij} values for a dihedron slanted at 45^{o} to the horizontal will have the following A_{ij} values.

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Some reasonable approximation to the structures, the flat plate, the vertical dihedron (or horizontal dihedron) and the slanted dihedron can be found in the environment. The sides of a building frequently make a horizontal dihedron with the ground. Slanted dihedrons may be made accidentally by rock strata and fissures. The change of polarization can be used to discover the presence of these accidental structures.

2.8. PROPAGATION EFFECTS

2.8.1. RADIO WAVE PROPAGATION THROUGH THE GROUND. Consider a plane wave propagating through a partially conducting dielectric medium like the ground. In general, there

will be absorption of the power in the wave as it travels through the ground. Such a wave is represented by

$$\mathcal{E} = \mathcal{E}_0 \exp(-\alpha x) \cos(\omega t - \beta x)$$

where $\mathcal{E}_0$ is the maximum amplitude of the electric field and $\mathcal{E}$ is the amplitude of the electric field at any other position, x, α is the attenuation constant,* and β is a constant related to the electrical properties of the soil.

It is convenient to use the complex notation to express the value of the electric field. Hence,

$$\mathcal{E} = \text{Re}\,[\mathcal{E}_0 \exp(-\alpha x) \exp(i\omega t - \beta x)] \text{ or}$$

$$\mathcal{E} = \text{Re}\left\{[\mathcal{E}_0 \exp(-\alpha x + i\beta x)][\exp(i\omega t)]\right\},$$

where Re means the "real part of."

So long as only one frequency is to be considered at a time, it is not necessary to repeat the frequency dependent part. Also the rest of the discussion is concerned with the real part of the complex value. It will not be necessary to repeat the symbol, Re. With these understandings the electric field may be expressed in what is called "phasor" form as

$$\mathcal{E} = \mathcal{E}_0 \exp(-\gamma x)$$

where $\gamma = \alpha + i\beta$.

The important quantities to know are the values of α and β for the ground which is to be penetrated. The following results can be found in Kerr.

$$\alpha = \frac{\omega}{c}\left[\frac{\left(\epsilon_1^2 + \epsilon_2^2\right)^{1/2} - \epsilon_2}{2}\right]^{1/2}$$

$$\beta = \frac{\omega}{c}\left[\frac{\left(\epsilon_1^2 + \epsilon_2^2\right)^{1/2} + \epsilon_2}{2}\right]^{1/2}$$

where ϵ_1 = dielectric constant of the ground,

$$\epsilon_2 = \frac{K}{\omega \epsilon_0},$$

*The use of α here should not be confused with the symbol for absorptance used in earlier sections.

K = ground conductivity,

ϵ_0 = permittivity of free space, a universal constant.

The attenuation of the electric field is due to the factor exp $(-\alpha x)$. Since the power is proportional to the square of the electric field, the power is attenuated by the factor exp $(-2\alpha x)$. This additional attenuation can be introduced into the radar range equation using it as an additional factor twice, once for outgoing power and once for returning power. Thus,

$$P_r = P_t \frac{GA\sigma}{(4\pi R^2)^2} \exp(-4\alpha R).$$

Some likely values would be, for example,

$$\frac{\omega}{2\pi} = 300 \text{ MHz}, \ \epsilon_1 = 30, \ K = 10^{-3} \text{ mho/m},$$

The attenuation becomes

$$\alpha = 0.0344 \text{ meter}^{-1} = 0.298 \text{ db per meter}.$$

The distance for which the amplitude is reduced by a factor e^{-1} is $(1/\alpha)$. It is called the loss distance and is 29.2 meters in the present example. The two way attenuation in the path through the ground is 0.596 db per meter of path length. We have lost about 30 db for a 50 meter path. This is why high frequency radio is not used for long distance communication through the ground.

We can consider the use of radar for seeing into the ground. For this purpose we should use vertical polarization with an incident direction close to Brewster's angle. See figure 2-24. Brewster's angle is that angle of incidence upon a dielectric surface for which the angle between the reflected ray and the refracted ray is a right angle. At the Brewster's angle vertically polarized radiation reflects the least. In a perfectly non conducting dielectric, vertically polarized radiation incident at the Brewster's angle is not reflected at all. All radiation penetrates the dielectric.

The attenuation of the power in the ground is less the smaller the conductivity and the larger the dielectric constant; thus dry sandy soils are easier to see into than moist clay soils.

The velocity of propagation in the ground is obtained from β. For high frequencies, if $\epsilon_1 >> \epsilon_2$,

$$\beta \cong \frac{\omega}{c}\sqrt{\epsilon_1}$$

and

$$v = \frac{\omega}{\beta}\frac{c}{\sqrt{\epsilon_1}}$$

As far as soils are concerned $\epsilon_1 \geqq 1$ and can be 4 or 5 for dry, sandy soils. The velocity v is then substantially smaller than c (the velocity in vacuo). The time of travel over a path in the ground is greater than over the same path length in air by the factor $\sqrt{\epsilon_1}$. This makes objects in the ground appear farther in radar range. In the example above, $\sqrt{\epsilon_1} = 5.47$. See figure 2-25.

If the ground surface is rough, there will be back scatter from it to the radar receiver producing "clutter." In the figure 2-25, the target return from T appears to come from T' and is displayed at the same range as the reflection from the spot X on the ground. This will tend to reduce the radar visibility of the target, T.

On the other hand, if we bury a corner reflector in the ground, its cross section will be enhanced. If we let A be the corner reflector area, then the radar cross section is

$$\sigma = \frac{4\pi A^2}{\lambda^2}$$ in the ground as compared with

$$\sigma_0 = \frac{4\pi A^2}{\lambda_0^2}$$ in free space.

Here λ is the wavelength in the ground, and λ_0 that in free space. Now $\lambda/\lambda_0 = c/v = \sqrt{\epsilon_1}$. Thus

$\sigma = \epsilon_1 \sigma_0$ or

$\sigma = 30\ \sigma_0$ in the example above.

2.8.2. PROPAGATION THROUGH FOLIAGE. A leaf is very thin compared with a radar wave length, but it is also very moist (dielectric constant of water = 80). Specular reflection can be dominant, but there will also be at least partial transmission, with attenuation of $e^{-\alpha \ell}$ where ℓ is the thickness of the leaf. Also the reflection from the front of the leaf is partially cancelled by reflection from the back. Attenuation in the leaf tends to increase the net reflection. The reflection is greater, the shorter the wavelength, due to the greater phase shift.

The random spacing of the leaves, with the slight retardation in each leaf, breaks up the wavefront, so that the wave is no longer coherent. After passage through a tree, for example, the signal, besides being weakened, also has a significant scattered component, with the scattered radiation going off in all directions. If the tree is not too thick, there may be enough coherent radiation to provide retroflection from a corner reflector, for example, which is

hidden behind the tree. Most of the reflected wave, however, will be scattered again in passing back through the tree. Unless the tree is very thin therefore, or the wavelength is sufficiently long, little or no echo will be received from the target.

2.8.3. ATMOSPHERIC EFFECTS ON PROPAGATION. Because of the long-distance potential of radar sensing and the fact that water vapor has a very high dielectric constant at radar wavelengths, small temporal and spatial variations in the atmosphere are much more important than for visual or infrared sensing. It is well known that light rays are bent in the atmosphere, e.g., one sees the sun after sunset. Radio waves are also bent, but to a greater extent, owing to the presence of water vapor in the air. Also, scattering of the radio waves by turbulent inhomogeneities or by raindrops, for example, can disperse the waves and result in attenuation of the radar signal, thereby decreasing the range. Rain drop scattering on the other hand produces diffuse echoes that are useful for remote location of rainy areas and tornados, mapping of hurricanes, etc.

A vertical gradient of air and water vapor is always present in the earth's atmosphere. Since the atmospheric density and therefore the refractive index increases with increasing height, the velocity of propagation increases with height, so that the wave front bends. Figure 2-26 shows a wave front at two times differing by dt; v_1 and v_2 are the velocities at two heights differing by dh; s_1 and s_2 are the directions of propagation, i.e., the ray directions, at the two times; and R is the radius of curvature of the ray.

The rate of change of velocity with height is d v/d h so that $v_2 = v_1 + (\partial v/\partial h)dh$, or

$$\frac{v_2}{v_1} = 1 + \frac{1}{v_1}\left(\frac{dv}{dh}\right)dh$$

By similar triangles,

$$\frac{v_2}{v_1} = \frac{v_2\,dt}{v_1\,dt} = 1 + \frac{dh}{R}$$

Equating the two ratios gives the curvature of the ray, 1/R

$$\frac{1}{R} = \frac{1}{v_1}\frac{dv}{dh} \doteq \frac{1}{v}\frac{dv}{dh}$$

where v is the mean velocity in the heights interval dh. The curvature of the earth is $1/R_e$, where R_e is the earth's radius. If we subtract (1/v)(dv/dh) from both the earth's curvature and the ray curvature we have a fictitious earth's curvature R_f.

$$\frac{1}{R_f} = \frac{1}{R_e} - \frac{1}{v}\frac{dv}{dh}$$

relative to which the ray is a straight line. This applies only to rays which are nearly horizontal, and it also applies only to rays in the troposphere, or lower atmosphere, where dv/dh is approximately constant with height. It is readily shown that the curved radio ray plots as a straight line if the fictitious earth radius R_f is used. For a given value of x the height of the ray above the actual earth is

$$y = -\frac{x^2}{2R} + \frac{x^2}{2R_e} = \frac{x^2}{2}\left(\frac{1}{R_e} - \frac{1}{R}\right).$$

For the fictitious earth

$$y = \frac{x^2}{2R_f}$$

and they are the same if

$$\frac{1}{R_f} = \frac{1}{R_e} - \frac{1}{R} = \frac{1}{R_e} - \frac{1}{v}\frac{dv}{dh}$$

as before.

For the so called "standard atmosphere" (which is a sort of mean atmosphere) near the surface, $R_f \doteq \frac{4}{3}R_e$ and is equal to 8500 km.

For most applications, it is convenient to represent the earth's surface by a paraboloidal approximation. The equation of a great circle on the earth relative to a point on the circle as origin: x = 0, y = 0.

$$x^2 + (y + R_e)^2 = R_e^2$$

Since we are dealing with low heights and limited distances along the earth, we can neglect y^2 with respect to $2yR_e$, so the equation becomes parabolic.

$$x^2 = -2yR_e$$

This is a good approximation for horizontal distances less than 1000 km. The vertical scale can be exaggerated relative to the scale of distance along the earth by any desired factor.

By plotting the actual surface profile (terrain, mountains, cities, etc.) on this scale, using the fictitious earth radius R_f to set the scale factor, the geometry of low-angle ray paths can readily be depicted and used to determine radar coverage and line-of-sight distances. See figure 2-27. All radio rays are still straight lines, representing the actual heights of rays above the terrain. The height-distance relation for the parabolic earth approximation, using standard atmospheric refraction, is:

$$d = 4.123\sqrt{h}$$

where h is measured in meters and d is measured in km. In English units it takes a very simple form, easy to remember:

$$d = \sqrt{2h}$$

where h is measured in feet and d in miles. Note that the height h of a horizontal ray above the earth at the distance d is also the height at which a radar antenna must be to "see" an object on the ground at the distance d.

Examples:

(a) A horizontal ray passes 50 feet above the ground at a distance of $\sqrt{2 \times 50} = 10$ miles.

(b) An airborne radar at a height of 3000 meters can see objects on the ground out to $4.123\sqrt{3000} = 216$ km.

The distance d (km) = $4.123\sqrt{h(m)}$ is often called the horizon distance or the radio line-of-sight distance.

It is often convenient to use the refractive index n to describe atmospheric refraction, since this can be measured. The velocity v = c/n where c is 3×10^8 meters per second - the velocity of light in a vacuum.

$$\frac{1}{v}\frac{dv}{dh} = -\frac{1}{n}\frac{dn}{dh} \doteq -\frac{dn}{dh}$$

since n is always quite close to unity. The quantity $\frac{dn}{dh}$ is called, in meteorological terms, the "lapse rate of refractive index."

If the lapse rate of n becomes large R_f can become infinite, or even negative. In this case the ray has a greater curvature than the earth; the fictitious earth is concave! One then "sees" distant objects in the sky. When this happens at optical wavelengths we call it a "mirage." Radar mirages are more common than optical mirages (fig. 2-28).

Variations in the refractive index lapse rate, due to meteorological factors, are continually taking place. One such variation is characterized by an increase in the lapse rate in a relatively thin layer aloft. If the elevation angle of the radar beam is small enough, the beam can be refracted around and returned to the earth (fig. 2-29). It is then reflected by the earth and can go through the same process again, and perhaps several times, depending on the geographical extent of the refracting layer. This phenomenon is called "ducting," since it is analogous to the propagation of waves in a wave guide. It can make the radar range unusually large. A duct like this in effect confines the electromagnetic waves between two spherical shells; the power spreads out in only two dimensions so that the round trip power attenuation is proportional to the square of the distance rather than the fourth power:

$$P_r \sim P_t \frac{GA\sigma}{(2\pi R)^2}$$

Two such refracting layers, with opposite lapse rate changes, sometimes form a duct aloft (fig. 2-30) and the trapped radar energy can travel to great distances before coming down to the ground again.

These phenomena occur only at low elevation angles; when a remote-sensing radar looks down at large depression angles, the waves go right through the duct, and the operation is not affected.

2.8.4. SPURIOUS RADAR RETURNS. At shallow angles strange results will sometimes be found, due to weather conditions. In the situation shown in figure 2-31, some radar power reflected from ship 1 to ship 2 can in turn be reflected by ship 2 back to the radar in another reflection from ship 1. The radar screen would then show an echo or ghost ship at position 3. Echoes are sometimes received from weather conditions or invisible objects. These are called "angles." Some are known to be due to flocks of birds.

Another type of spurious radar return is called "second go-around echoes." These are caused by the repetitive nature of the radar pulses. Suppose a radar transmits 1000 pulses per second. The interpulse period, slightly less than 0.001 second, is allotted for reception of echoes. The radar range corresponding to 0.001 second is 150 km. Now suppose a large object exists at 165 km distance. The echo from this arrives after 0.0011 second, or 100 microseconds later than the emission of the next pulse, i.e., the start of the next range sweep on the oscilloscope. It would be interpreted as an echo from the next pulse, at a range of only 15 km (fig. 2-32).

To avoid such misleading echoes it is necessary to allow sufficient time for all echoes to be received before transmitting the next pulse. In particular, one must allow for the distance beyond the horizon at which tall objects would be in the line of sight. This may be evaluated by plotting a profile on the parabolic earth chart described above, including all the tall objects that might be seen from the radar site (fig. 2-33). The maximum range would be represented by the ray from the top of the tall object that just grazes the terrain. Alternatively, this range may be calculated as the sum of the horizon distances for the radar and the top of the tall object. For example, if the radar is 300 meters high and the tall object is a building 400 meters high the maximum range, using standard atmospheric refraction, would be:

$$d_{max} = 4.123(\sqrt{300} + \sqrt{400}) = 149 \text{ km}$$

The maximum permissible pulse repetition rate would then be

$$\frac{c}{2d_{max}} = \frac{300000}{298} \cong 1000 \text{ pulses per second.}$$

This will not guarantee the elimination of second go-around echoes; if ducting or strong refraction is present the radar range and hence the probability of second go-around echoes, is increased. Normally, however, such echoes are likely to be weak and should not be noticeable in a mapping radar system. They could, however, place echoes from distant mountains in a foreground plain. Should such echoes occur on a map, the chances are that they would not appear in the same place on a resurvey, since they depend on a special viewing angle or on special weather conditions.

2.9. MOVING TARGET INDICATION

We have considered polarization as a means of extracting additional information from the radar echoes. "MTI" is another way.

It is well known from everyday experience that if either a source of sound or the listener is in motion, an apparent shift in frequency (pitch) will be observed (change in pitch of an auto horn of a passing car for example). This is known as the Doppler effect; and it also occurs in the case of electromagnetic radiation (the so-called "red shift" of the radiation of distant galaxies is the basis for the expanding universe theory). It is the phenomenon on which MTI radar is based.

If R is the range from a radar to a target, the number of waves of length λ in the path from radar to target and back to the radar is $2R/\lambda$. Each wavelength corresponds to an angular excursion of the electromagnetic wave of 2π radians; hence the total angular excursion ϕ made by the electromagnetic wave when traveling from radar to target and back is $\phi = \frac{4\pi R}{\lambda}$ radians. If the target is in motion with a velocity V toward or away from the radar, R and ϕ are consequently changing. Since a change in ϕ with respect to time represents a change in frequency, a "Doppler shift" is the result. This shift is given by

$$\omega_D = 2\pi\nu_D = \frac{d\phi}{dt} = \frac{4\pi}{\lambda}\frac{dR}{dt} = \frac{4\pi V}{\lambda} \text{ radians per second} \qquad (8)$$

In terms of frequency,

$$\nu_D = \frac{\omega_D}{2\pi} = \frac{2V}{\lambda} = \frac{2V\nu_0}{c} \qquad (9)$$

the last expression resulting because λ, the wavelength, is equal to the velocity of propagation of the wave, about 3×10^8 m/sec in air, divided by the radiated frequency ν_0. In case V is measured in km per hour and λ in cm,

$$\nu_D = \frac{55.6V}{\lambda} \text{ Hz.} \tag{10}$$

As an example, consider an automobile traveling at a speed of 100 km per hour directly away from a radar operating at a wavelength of 10 cm. The Doppler shift in frequency received at the radar after reflection from the automobile will be:

$$\nu_D = \frac{55.6(100)}{10} = 556 \text{ Hz}$$

a note between $C^{\#}$ and D, in the octave above middle C. This is an extremely small shift in frequency relative to that of the radar, which is at a frequency of

$$\nu_0 = \frac{3 \times 10^8}{0.1} = 3{,}000{,}000{,}000 \text{ Hz.}$$

The frequency of the wave reflected by the target back to the radar is the radiated frequency minus the Doppler shift (because the target is moving away from the radar):

$$\nu_r = \nu_0 - \nu_D$$

where ν_r = frequency received.

$$\nu_r = 3{,}000{,}000{,}000 - 556 = 2{,}999{,}999{,}446 \text{ Hz.}$$

It is evident that no filter can be constructed which will distinguish between these two nearly identical frequencies in order to detect target motion. How then can an MTI radar be made to work? Consider a radar consisting of two separate but adjacent pieces of equipment, each with its own antenna. One set will be used for transmission, the other for reception. The transmitter section consists simply of a continuous-wave (CW) oscillator coupled to the antenna. The receiver section comprises a similar antenna connected to a "mixer" or "detector" (a non-linear device), followed by an audio-frequency amplifier and an acoustic transducer, such as headphones or a loudspeaker.

The transmitting antenna emits a continuous stream of electromagnetic energy at, say, 10 cm wavelength, some of which is reflected from moving and fixed targets and is received by the receiving antenna, and some of which enters the receiving antenna directly, due to its proximity. Signals reflected from moving targets, shifted in frequency by the Doppler effect, combine in the detector with the leakage signal from the transmitter to produce ν_D, the Doppler shift or difference frequency, as a result of the non-linear characteristic of the detector. This difference frequency is amplified and heard by the operator as a tone.

The described equipment resembles the radars used by police for detection of speeders. By appropriate design, a single antenna can be used for both transmission and reception, and a frequency meter calibrated in mph or km per hour replaces the headphones. No range

information is provided by such a radar. Echoes from fixed objects of course produce no Doppler difference frequency, and if the audio amplifier is designed to "cut off," or not amplify, at very low frequencies, fixed target returns will have no effect on the output.

Objects receding from or moving toward the radar at the same radial speed produce identical audio outputs; the direction of radial motion is lost in the detection process.

Radars such as those described above are more properly classified as "CW" radars. By appropriately frequency modulating the signal, it is possible for such radars to measure range; such a radar is called an "FM-CW" radar. Some types of radar altimeters operate on this principle.

A radar which uses a portion of the transmitted energy, "leaked" to the receiver, as a reference frequency for the detection of Doppler shifts, is by virtue of this also a member of a subclass of radar called "coherent radars." All radars of this type use the transmitted signal as a reference for comparison with the received echo.

All MTI radars must have a reference signal; however, an MTI radar need not be coherent. In "noncoherent" or "clutter-coherent" MTI, the reference is the transmitted signal reflected from a non-moving object such as a tree, the earth or various objects near the moving target which it is desired to detect. This type of radar does not require the preservation of the transmitted signal as a reference in the equipment; the preservation is performed externally. The phase of the reference signal as reflected from the fixed objects, is not predictable but is not required for moving target detection.

Consider a simple example in which there exists a single moving target located near a single fixed object such as a building. The radar will receive power reflected from both objects simultaneously if they are at the same range, and will produce an output on the display proportional to the square of the sum of the two reflection amplitudes. The sum is a vector sum, dependent on the relative phases of the two reflections (see fig. 2-34).

The radar is a pulsed radar, and in the figure two successive radar returns from two successive transmitted pulses are shown. It can be seen that as the phase of the reflection from the moving target changes relative to that of the fixed target from one pulse to the next, the sum or resultant will change in amplitude. The radar can be arranged to make visible or audible this change in resultant echo, thus providing means to detect the presence of a moving target at that range. One way of doing this is by displaying the receiver output on an amplitude versus range oscilloscope ("A" scope; see fig. 2-35).

As successive radar returns are received and displayed in rapid sequence on the oscilloscope, there will be a relatively small amount of fluctuation in echo amplitude except at the range of the moving target. At this range the fluctuation in echo amplitude from pulse to pulse

produces a characteristic fluttering pattern similar to that shown in the figure, indicating the presence of the moving target. This means of detection has been termed "butterfly" because of its appearance suggesting butterfly wings. This is not a particularly sensitive means of detection. A much more sensitive method is to arrange a means for allowing the ear to detect the presence of a target. A way of doing this is to provide the operator with a controllable gate which he can adjust in range; the gate allows only that portion of the received signals which are at the range of the gate to be amplified in an audio frequency amplifier and sent to headphones. The sensitivity of this means of detection is greatly increased if a "sample-and-hold" or "boxcar" circuit is used in conjunction with the gate. This circuit automatically locks to the amplitude of the received signal at the time (range) of its occurrence and preserves this amplitude until the next time the signal from that range is received, when it adjusts to the new value, preserving that until the next reception, and so on. By this means the audio-frequency power, representing the Doppler frequency, is tremendously increased, thus enhancing detectability. A filter must also be used in the audio circuit to remove the radar pulse repetitive frequency component, which would otherwise drown out the return from the target. This means of MTI detection is one of the most sensitive known. Its disadvantage is that it is relatively time consuming to search all ranges and many directions (of antenna pointing) for moving targets.

A pulsed MTI radar such as the one described, while being capable of supplying target range information, possesses the disadvantage that under certain circumstances it may fail to indicate the correct Doppler frequency of a moving target, or on occasion may even fail to detect the presence of the moving target at all. The reason for this is that the target is observed only when a reflected pulse is received and not during the time intervals between pulses. Thus the motion of the target is perceived at the radar in a manner similar to the way in which the wheels of a moving stagecoach are seen in a Western movie, i.e., sometimes rotating correctly, sometimes backward, sometimes not at all. If the radar pulses are transmitted once per rotation of the rotating vector representing the moving target (see fig. 2-34); no motion will be detected, since successive vector summations will be identical. This situation occurs when

$$\nu_D = \nu_r = \frac{2V}{\lambda}$$

A similar situation occurs at integral multiples:

$$\nu_D = n\nu_r = \frac{2nV}{\lambda}$$

where n = 1, 2, 3 . . . . . These are called MTI blind speeds.

Suppose the radar pulse rate is slightly less than ν_D (or $n\nu_D$). This means that the radar receives information on each successive pulse at times such that the rotating vector representing

the moving target (see fig. 2-34) has rotated slightly more than one complete revolution, and appears to the radar (and to the operator) as though it were rotating very slowly, whereas its actual speed might be very high. Thus, under some conditions, a pulsed MTI radar gives false indications of target Doppler frequencies, or target speeds. The maximum target speed which can be measured by a pulsed radar is that which produces a Doppler frequency equal to one-half the radar pulse repetition frequency. All other indicated target speeds will be false. This phenomenon has been called "Doppler fold-over." Figure 2-36 shows a plot of radar-measured target Doppler frequency vs. actual frequency.

An MTI radar such as the one described, comprising a transmitter, receiver, movable range gate, sample-and-hold circuit audio frequency amplifier and headphones, is frequently called an "aural butterfly" radar. It can be used to detect, or measure the velocity of (within the constraints mentioned above) moving objects such as animals or birds, rainfall, waves, moving automobiles.

2.9.1. ADVANTAGES OF NON-COHERENT MTI. Because a non-coherent radar measures the speed of a target relative to the surrounding fixed environment rather than relative to the radar, the radar may be carried on a moving vehicle such as an aircraft without fundamentally affecting its capabilities. A coherent radar requires modification in order to operate under these conditions. Because of this, a non-coherent MTI radar to be carried on a moving vehicle such as an aircraft is considerably simpler than a coherent one would be.

2.9.2. DISADVANTAGES OF NON-COHERENT MTI. Under some conditions, a non-coherent MTI radar may not indicate the presence of a moving target because of the absence of sufficient background reflections to serve as a reference. Moving aircraft viewed at high angles, and boats moving in relatively calm water are examples of situations in which the target may not be identified as having a Doppler frequency. A non-coherent radar may be shown to be in theory somewhat less sensitive in detecting targets than a coherent MTI radar.

2.9.3. OTHER LIMITATIONS OF MTI RADAR. Heretofore the background of reflecting objects from which the moving target is to be distinguished by virtue of its Doppler frequency has been assumed to be fixed in space. Such is, of course, not the case when the background consists of trees and vegetation which can move in the wind. Such movement produces random Doppler frequencies which tend to mask the presence of slow-moving targets. Winds of around 12 km/hr or above move the trees themselves; below this velocity the wind moves the leaves only. Both motions produce random Doppler shifts which tend to reduce the ability of the radar to detect slow-moving objects. The spectrum of Doppler frequencies due to wind may extend to about 2 km/hr and has a distribution of energy as shown in figure 2-37.

In case the MTI radar is carried on a moving platform such as an airplane, a further degradation can occur. Suppose the radar antenna is pointed directly sideways, at right angles to the aircraft motion (see fig. 2-38).

If the antenna has a beamwidth θ, it can be seen that those fixed targets located forward of the perpendicular to the flight path (region X) will be moving toward the aircraft, while those in region Y will be moving away from the aircraft. This effect occurs at all ranges. These two motions will interact in the radar mixer and produce an apparent Doppler motion of all fixed targets, proportional to the difference in the radial speeds of objects in regions X and Y. The greatest difference is due to objects at the "edge" of the beam:

$$\frac{d(R_2 - R_1)}{dt} \cong V\theta$$

where V is the velocity of the airplane. The effect decreases if the antenna is pointed other than broadside to the aircraft motion, and almost disappears if the antenna is pointed directly forward or aft. The residual effect under these conditions is caused by the altitude of the airplane over the ground.

As an example, assume V = 200 km/hr, θ = 1/20 radian.

Then, $V\theta$ = 10 km/hr maximum. This effect tends to make detection of moving targets at 10 km/hr or less increasingly difficult.

2.9.4. "AUTOMATIC" MTI. An MTI radar can be made which will selectively present the echoes from moving targets on a plan position indicator. Various techniques exist for doing this; one of the simplest and best-known is to preserve or store the echoes from one transmission or pulse until the echoes from the next pulse arrive, and subtract one set from the other. If all returns are from fixed objects, both sets will be equal and there will be no residual signal to present on the indicator; if a moving object is present, there will in general be a residual signal from this object because of the Doppler phase shift (see fig. 2-34) which will be displayed on the indicator. The traditional device for accomplishing the storage is a delay line, consisting of a quartz crystal or column of mercury (early implementation) into which the radar returns are channeled as acoustic vibrations. The processing portion of the radar will be as diagrammed in figure 2-39.

The delay line must have a delay precisely equal to the interval between the transmitted pulses, in order that the subtraction can be performed on each return echo without leaving any residue in the case of fixed targets.

The delay line canceller, as this device is called, of course has the same limitation insofar as blind speeds are concerned as any pulsed MTI radar. It can be regarded as a selective

filter of special characteristics, with a frequency response (relative visibility of moving targets) as sketched in figure 2-40.

It can be seen from the figure that the greatest sensitivity of such a canceller is at Doppler frequencies at or near multiples of 1/2 the radar PRF, the sensitivity falling off to zero at the blind speeds. More sophisticated canceller design, incorporating more complex processing and involving more than one delay line, makes it possible to create a filter passband possessing more nearly an equal response to all moving targets except those at the radar blind speeds.

As with any MTI system, "fixed" targets which actually have motion, such as moving trees, or fixed targets viewed from a moving platform, will not cancel completely.

2.10. SYSTEM CONSIDERATIONS

Radar systems are capable of more flexibility in design than infrared scanners or cameras. For example, one has a wider choice of radiated wavelength; the transmitter power can vary within wide limits; it is possible to choose among a variety of waveforms for transmitter modulation, e.g., CW, FM-CW, short or long pulses, or frequency-swept pulses. One may choose to employ a coherent or a non-coherent radar, among a wide variety of antenna beamshapes, polarizations, and signal processing (MTI is one form of signal processing).

2.10.1. CHOICE OF RADAR.

As a consequence of this wide availability of different radar configurations, most existing radars are designed for a specific purpose. Therefore, it is not likely that a prospective user will find an existing and available radar exactly suited to his purpose. Instead, one must design and build the equipment to suit the purpose, or specify its characteristics sufficiently to allow a manufacturer to design and built it, or adapt an existing one. The latter alternative is often the preferable one if time and money are important. New radar designs are very expensive and often take an extremely long time to construct and get into full operation.

Radar components can be bought from commercial sources and assembled; however, the antenna must usually be specially designed, and thus should be one of the first items considered.

The first question to answer when considering the application of a radar-type sensor is: "what is the purpose for which the radar will be used?" For example, one might wish to: 1) observe the thickness of ice in Antarctica; 2) measure the polarization matrix of a moon crater; 3) locate the depth of the water table in the desert; 4) measure the speed of an Olympic runner; 5) study the flight of a condor; 6) find a treasure hidden in the waters of a lake; 7) survey agricultural crops; 8) locate the site of an ancient city in Yucatan; 9) map the Amazon valley; 10) track a satellite; 11) find geologic fault lines.

Evidently, many of these tasks would require a different radar design. As an example, suppose we consider example no. 8, locating the site of an ancient city in Yucatan.

Step 1. Evidently, from the size of the area to be searched, the radar must be an airborne one. A DC-3 type aircraft can carry a large radar, is inexpensive and has sufficient range and reliability.

Step 2. Assume the target is a mound on level ground. If it has a frontal area of 1000 square meters, and is rough, we can assume it to have a radar cross-section σ_1 of about 1000 square meters, since G', the target gain, would be approximately unity for a rough surface approximately normal to the radar line of sight. The level ground can be assumed to have a radar cross-section of about 1% of its physical (patch) area. Thus σ_2, the cross-section of the level ground, will be

$$\sigma_2 \cong 0.01 \frac{ct}{2} R\theta,$$

where $\frac{ct}{2}$ is the pulse length in the range direction and $R\theta$ is its extent in the azimuth direction. If we select 1/5 microsecond as the pulse length.

$$\sigma_2 \cong 0.3R\theta$$

In order to be able to see the target easily, the return from the target plus the "clutter" (σ_2) must be much greater than that from the clutter alone. Since the radar return is proportional to the cross-section of the target, this means that

$$\frac{\sigma_1 + \sigma_2}{\sigma_2} >> 1$$

or,

$$\sigma_1/\sigma_2 >> 1 \tag{11}$$

$$\frac{1000}{0.3R} >> 1.$$

In order to facilitate the search of large areas, we would like to make R, the range, as great as possible. In order to keep $R\theta$ small, we would like to keep θ small, which implies a long antenna or a short wavelength. It would reduce cost to keep the antenna small; thus a short wavelength, λ, would appear desirable. On the other hand, a longer wavelength will permit the radar to more easily penetrate screening vegetation. Therefore, a compromise must be reached.

Let us assume D is fixed at 3/4 meter. There are X-band and S-band radars available as military surplus items with antennas about that length. If we assume λ = 0.033 m (X-band).

$$\theta \text{ (beamwidth)} = \frac{\lambda}{D} = \frac{0.033}{3/4} = 0.044 \text{ radians}$$

substituting into equation 11,

$$\frac{1000}{0.3R(0.044)} >> 1; \quad R << 76 \text{ km}$$

A reliable detection range might be R = 7.6 km. If $\lambda = 0.1$ (S-band) is assumed,

$$\theta = \frac{0.1}{(3/4)} = 0.133 \text{ radians.}$$

$$\frac{1000}{0.3R(0.133)} >> 1; \quad R << 25 \text{ km}$$

A reliable detection range in this case might be R = 2.5 km.

It would seem that X-band, giving the longer detection range, might be the appropriate choice.

Step 3: Obtain a surplus AN/APS-15 X-band radar (try to obtain three, if possible, for spares!) and modify it to produce a pulse length of 1/5 microsecond, increasing the receiver bandwidth to 10 mHz (5 mHz video bandwidth).

Step 4: Provide as required: 400 Hz inverter for power supply, a flux gate compass to operate a heading marker on the PPI scope, and a camera to photograph every scan of the PPI. Operate the radar on the ground until the equipment is working well and the radar operators are well trained.

Step 5: Install the radar in the aircraft with minimum aircraft modification, to save money. Aircraft modifications are very expensive, time-consuming and fraught with red tape. Mount the antenna externally in a radome. Mount the internal parts on a cargo platform adequately strapped down, but easily removable. Provide a navigator's drift sight and a tape recorder for recording pertinent data by voice.

Step 6: Now that the equipment is installed and operating and the operating crew is trained, it is appropriate to begin flight operations. In general, this is where the main problems begin. Experimenters who plan airborne equipment operation and data gathering should be cognizant of the considerable difficulty that almost always besets such attempts. Among the traditional problems almost always encountered may be listed the following:

1) Mandatory aircraft down-time for inspection and maintenance always seem to occur at crucial and inconvenient times.

2) It is difficult for even a well trained crew to operate equipment satisfactorily while airborne. Unaccountable mistakes and lapses of memory occur, equipment adjustments are not made correctly, and even minor problems seem insurmountable.

3) The weather has a tendency to be unflyable for extended periods, except when the aircraft is down for maintenance or the equipment is not performing well.

4) Equipment that works perfectly on the ground will repeatedly fail for mysterious reasons when airborne.

Planners of experiments should make over-generous allowances of time for an airborne experiment to begin to produce satisfactory data. Three months from the time the equipment is installed and operating on the ground is not excessive. To help avoid some of the problems one should:

1) Make check lists of everything to be done in the air.

2) Operate the equipment on the ground with aircraft engines running at various throttle and rpm settings and with the equipment connected to the aircraft power supply so as to reveal, insofar as possible, problems caused by vibration, voltage fluctuation, and electrical noise.

2.10.2. OPERATION. When the equipment and operators have demonstrated ability to function while airborne, flights should be made over a known area to learn the working range of the radar and establish the operating procedure to be followed. Then, the unknown area should be searched in a regular pattern. In the case of the example radar experiment discussed above, make strip maps of the PPI photographs and examine them very carefully. Examine all the data that is taken immediately, while the experiment is in progress, even if this means long hours. If data examination is postponed until after the experiment has been completed, it is too late to re-take data that has been ruined by unsatisfactory system performance.

The final difficulty that may be encountered is that the equipment and crew may begin to operate very efficiently and flood the system with data. This has happened many times. Do not lose control of the experiment.

2.11. LASERS

No exposition on Radar is complete without a brief discussion of lasers. The laser is a device which can produce highly monochromatic coherent radiation in the infrared, visible, and ultraviolet spectral ranges so that all radar techniques previously discussed can equally well be applied to these short wavelength spectral ranges. Until the invention of the laser, the exploitation of all of the radar techniques in these spectral ranges was not feasible. Reception of laser radiation is done with the usual detecting elements that respond to the power in the spectral range of operation.

2.11.1. PROPERTIES OF LASERS. A comprehensive series of articles on lasers may be found in "Applied Optics," Vol. 5, No. 10, October 1966.

A laser has three basic parts - the amplifying medium, the primary source of power or "pump," and a resonant cavity to assist in providing regeneration at the laser frequency and propagation-direction.

The amplifying medium is an ensemble of atoms distributed in space generally by dispersal within a host material. The host materials may be solids, liquids, or gases. The active atoms must be carefully chosen so that the combination of host material and active atom achieves a suitable fluorescent property. The function of the pumping source is to cause a larger population of active atoms to remain in the excited state than can be achieved by any amount of heating. Such a condition is termed a "population inversion."

The emission of radiation by the active atom is triggered by radiation of the laser frequency passing by so that each atom contributes to the radiation in phase with the radiation moving by. The resonant cavity is designed to maintain sufficient triggering radiation of only one frequency and propagation direction within the cavity so that nearly all contributions add both in phase and in one direction. It is desired to have all possible pumped power go into one direction and at one frequency of laser light.

If a laser operates so as to transform the pump power continuously into laser light, the laser is said to be operating in the cw (continuous wave) mode. If the laser operates so as to store the energy from the pump source and release that energy by intermittent pulses of laser light, the laser is said to be operating in the pulsed mode. Also, if the power can only be delivered to the laser in the form of short pulses to avoid overheating the laser host material, the laser is said to be operating in the pulsed mode.

The important requirement for laser operation is to establish the inversion population. The thermal motion of the active atoms in the host material may tend to work against the maintenance of that condition by allowing de-excitation through mechanical perturbation. Hence, some lasers must be operated at reduced temperature.

Because of the current popularity of lasers, many people tend to consider the application of lasers to problems which are better and more economically done by traditional and unromantic means. It is worth while to state the principle new features offered by lasers which make lasers appropriate for a given application.

1. A laser can be a relatively high power monochromatic source at optical wavelengths (watts per micron).
2. A laser can be a source of high intensity (watts per steradian).
3. A laser can be a phase coherent source at optical wave lengths.

The use of lasers for remote sensing is not well demonstrated. The most promising application is for laser altimeters and ground profile measurement. Very similar work can be

done with radar without the atmospheric scattering problem. It may be possible to detect suspended pollutants in water and air by laser wavelengths which may be not measurable with microwaves.

It is important to remember that the radiation from lasers obeys the same optical laws as ordinary light and, therefore, laser light differs in degree and not in kind from any other optical radiation.

2.11.2. WAVE FRONT RECONSTRUCTION. An important new application of laser radiation is lens-less photography that requires coherent laser light. A good review article is "Photography by Laser," E. Leith and J. Upatnieks, Scientific American, Vol. 212, June 1965.

An important fact to bear in mind is that the principle of lens-less photography can be applied to the recording of any wave motion provided that a suitable substitute for the photographic recording is used. Thus, it is quite possible to use the principles of wave front reconstruction for sound waves, seismic waves, infrared and radiowaves. The applications to remote sensing are not as yet mature so that such applications should be considered as future possibilities requiring sensor research in advance of applications research.

The mathematical formulation of lens-less photography is elegant and shows that the concept is parallel to many others already used in electrical engineering for years. An intuitive grasp of the concept can be achieved by describing the process in terms of Huygen's principle and diffraction - grating behavior.

Huygen's principle states that each section of a wave front may be considered to be a new source of spherical waves beginning with the same phase as the original wave front. An experiment where two slit apertures are separated by a distance, d, and illuminated from the left with plane monochromatic waves, produces, on the right side circular wave fronts expanding from the two apertures in phase with the original wave. At a long distance away these two waves overlap; their amplitudes add.

The addition is in phase (constructive) when the path difference from the two apertures to the field point is an integral number of wavelengths. Thus the condition:

$$n\lambda = d \sin \theta_n, \quad n = 0, \pm 1, \pm 2, \pm 3 \ldots\ldots\ldots$$

are the conditions for constructive interference in the directions θ_n. The number n is the order number.

A grating is a multiple slit aperture with slit spacings, d, all equal. The direction θ_n still gives the direction for which all slit contributions add.

We can make a grating from a photographic plate by producing a two slit interference pattern on an unexposed plate. When the plate is developed the regular spacing of the interference pattern produces a slit grating pattern of transparent and opaque lines. We can also make the same grating using unfocused laser light if we split the laser beam so that one beam is normal to the plate and the other beam is incident at an angle, θ_i. A sketch will show that the spacing of maxima on the plates will be

$$d = \frac{\lambda}{\sin \theta_i}$$

If we develop the photograph and again allow the normal beam to strike the developed grating, we can expect that a diffracted wave in first order will be produced at an angle such that for first order

$$\lambda = d \sin \theta$$

as expected for that grating spacing. But since λ and d are the same as used to make the grating then $\theta = \theta_i$. Thus, on the right side of the plate we will find the normal beam (as zeroth order) and the other beam (as first order) coming from the plate at angle θ_i although the left side of the plate has only the normal beam incident. In this way we have reconstructed the beam at angle θ_i from the normal beam and the photograph.

It is only an extension of this idea to reconstruct a curved wave front on the right from a photograph of the interference pattern of a normal reference beam and the curved wave front on the left. For a curved wave front the θ_i varies across the plate forming diffraction spacings d which vary across the plate. If θ_i is large, d is small, etc. Hence, when the normal beam is made incident again on the developed photograph the parts of the normal beam which fall at small spacing positions produce first order diffracted waves at the corresponding large θ_i angles and the parts which fall at large d spacings yield wave fronts on the right propagating at small θ_i angles. Hence the curvature of the curved wave will be reconstructed.

It is in this manner that a most complicated curving wave front reflected from objects to be photographed are reconstructed. The normal beam need not be normal to the plate but can be at any convenient angle; it need not be a plane wave itself, however, it must be reproducible with good accuracy or the "playback" process will fail. Thus the normal beam is simply any reference beam and is usually a plane wave at some convenient angle of incidence which is easily reproduced later.

In wave motion which does not effect photographic film, the interference pattern must be recorded by other means. If the recording is by electronic means then it is possible to introduce the effect of the reference beam artificially by appropriate oscillators without an actual reference beam being formed. The combined data can be printed on photographic film and

viewed by laser light to see the image by the same kind of CRT line scan printers used in airborne scanners. On the other hand, it would be possible to compute what the image would be if brought to focus by a lens and have a computer print out the image.

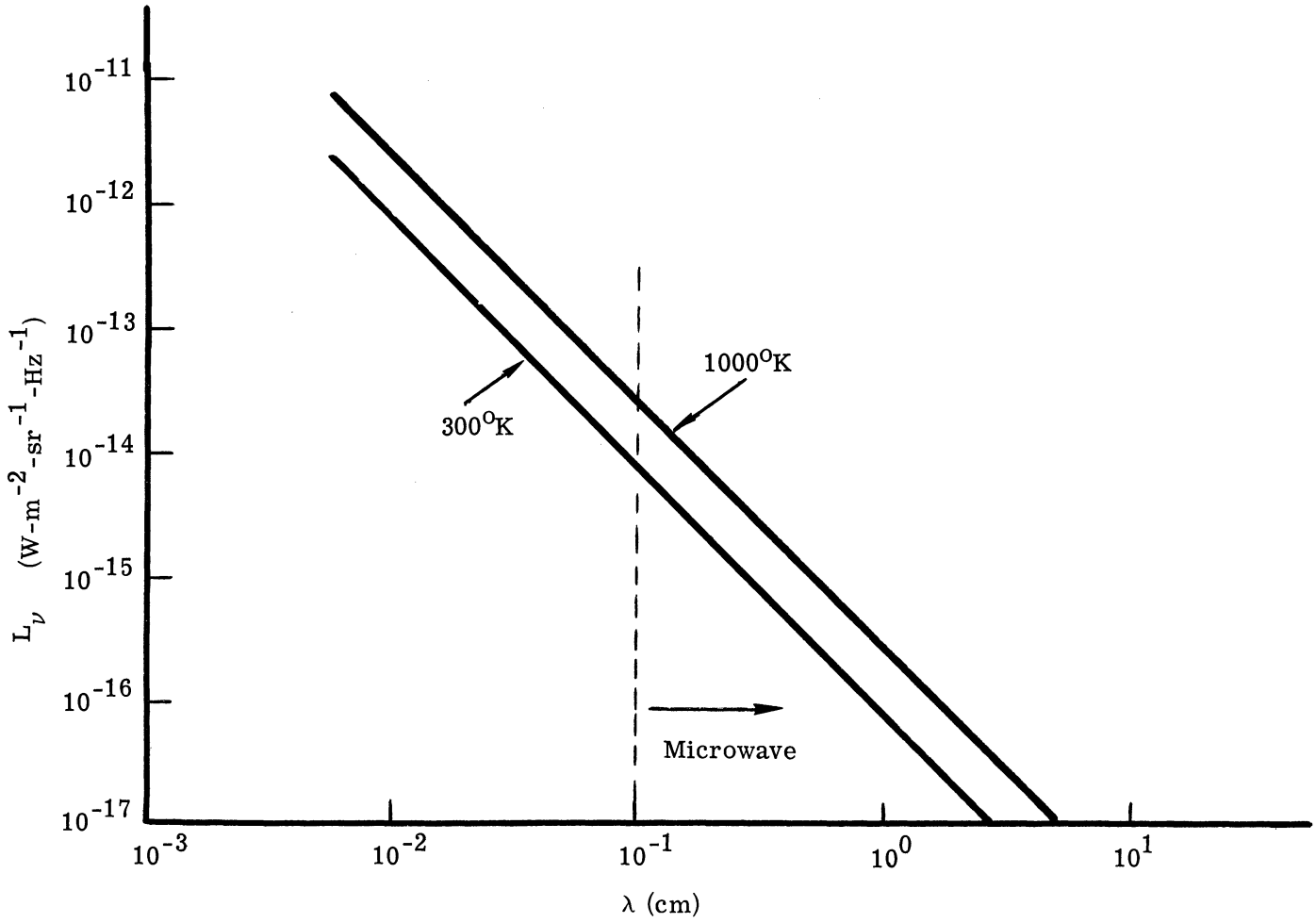


FIGURE 1-1. SPECTRAL EXITANCE, M_λ, FOR 300°K AND 1000°K BLACKBODY

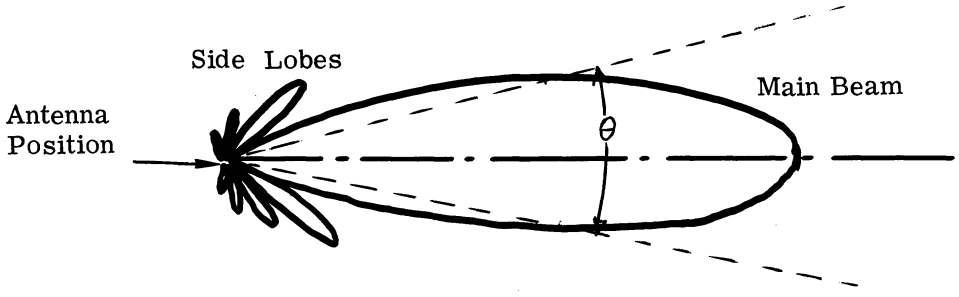


FIGURE 1-2. POLAR PLOT OF ANTENNA GAIN. Increasing gain is plotted as increase in radial distance from the antenna position.

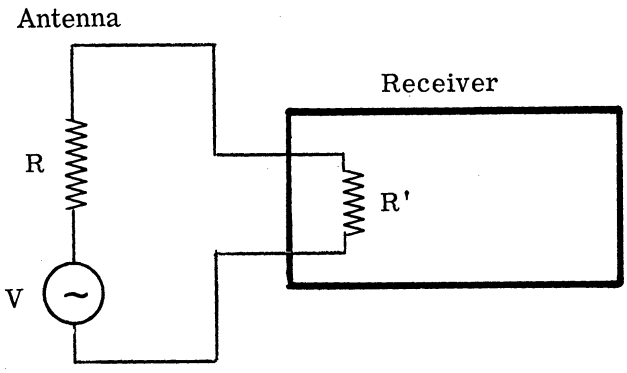


FIGURE 1-3a. ELEMENTARY DIAGRAM OF A MICROWAVE RADIOMETER

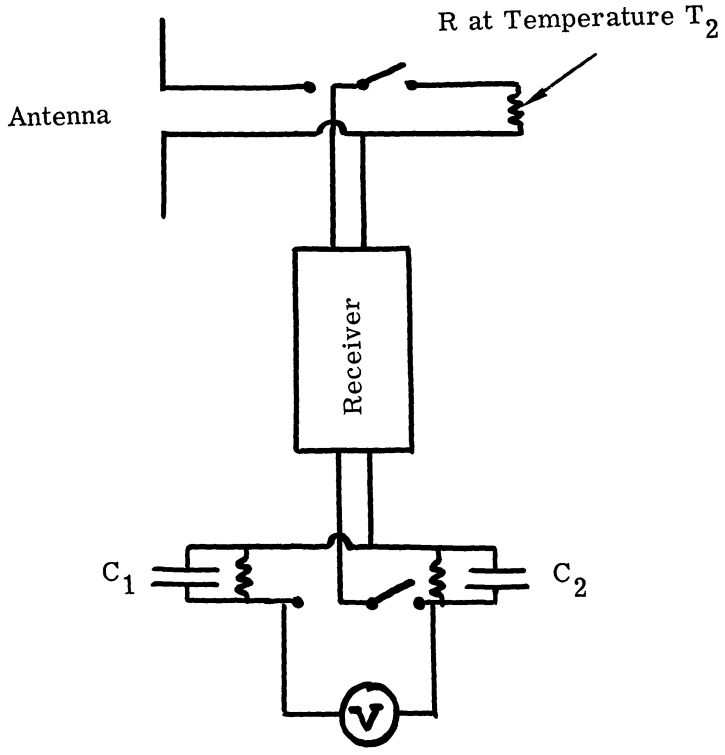


FIGURE 1-3b. SCHEMATIC OF MICROWAVE RADIOMETER

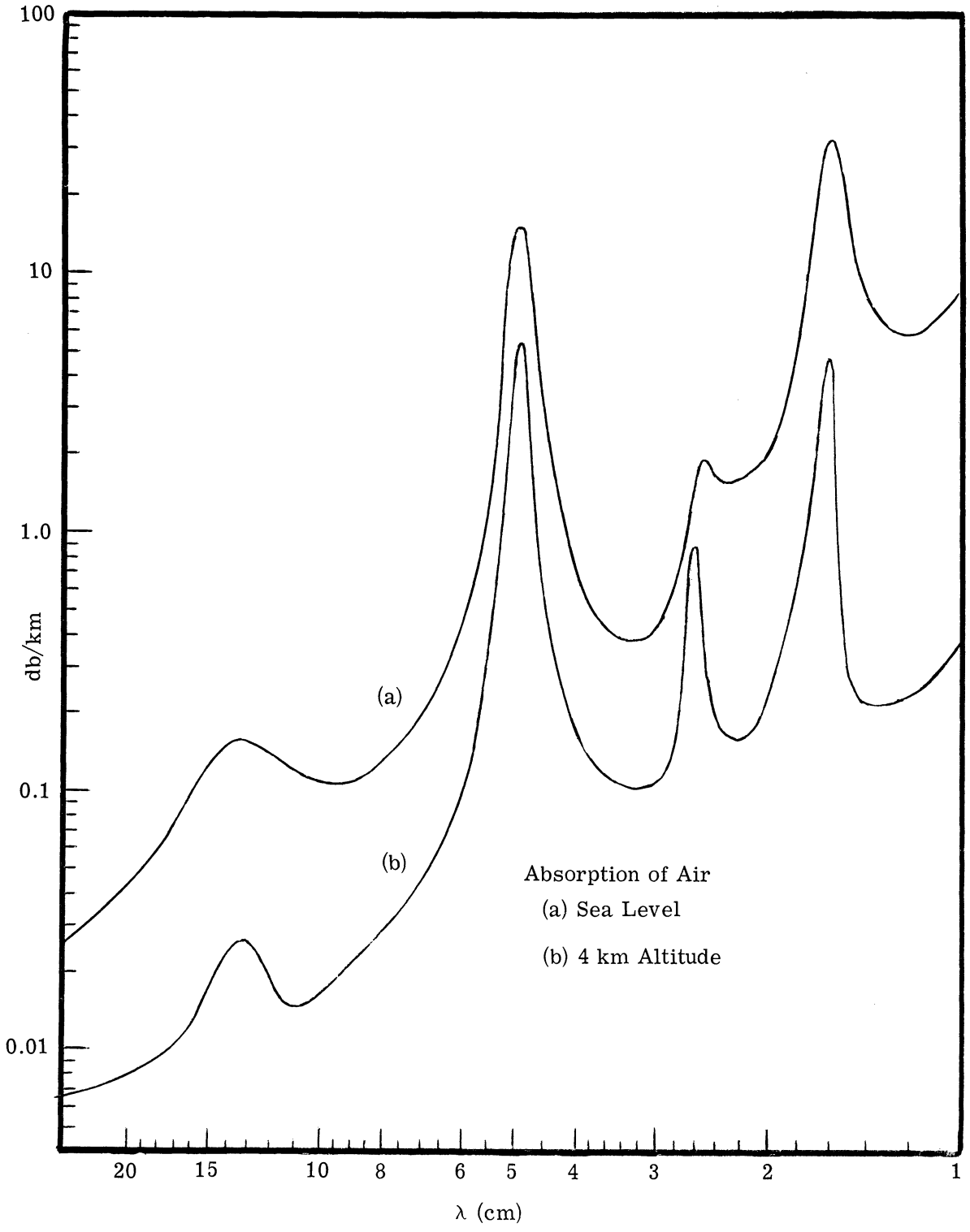


FIGURE 1-4. ABSORPTION OF AIR

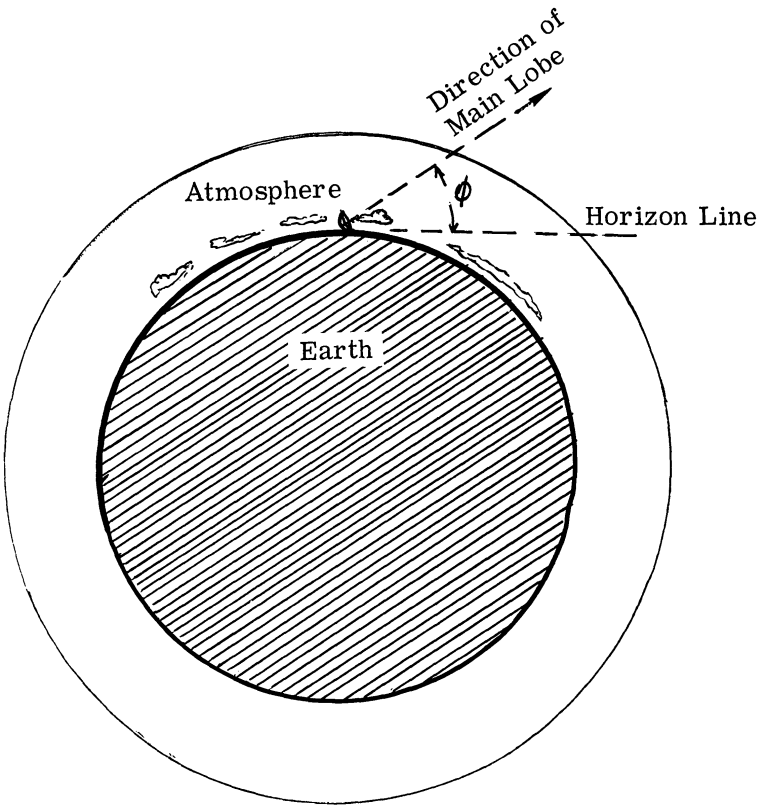


FIGURE 1-5. CHANGE IN ATMOSPHERIC PATH WITH DIRECTION OF VIEW

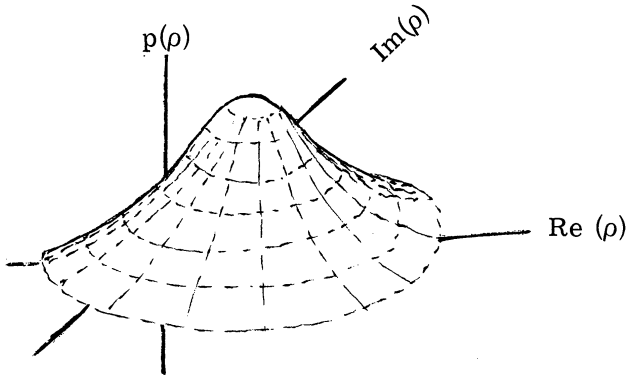


FIGURE 2-1. BIVARIATE DISTRIBUTION OF SIGNAL AND NOISE VOLTAGES

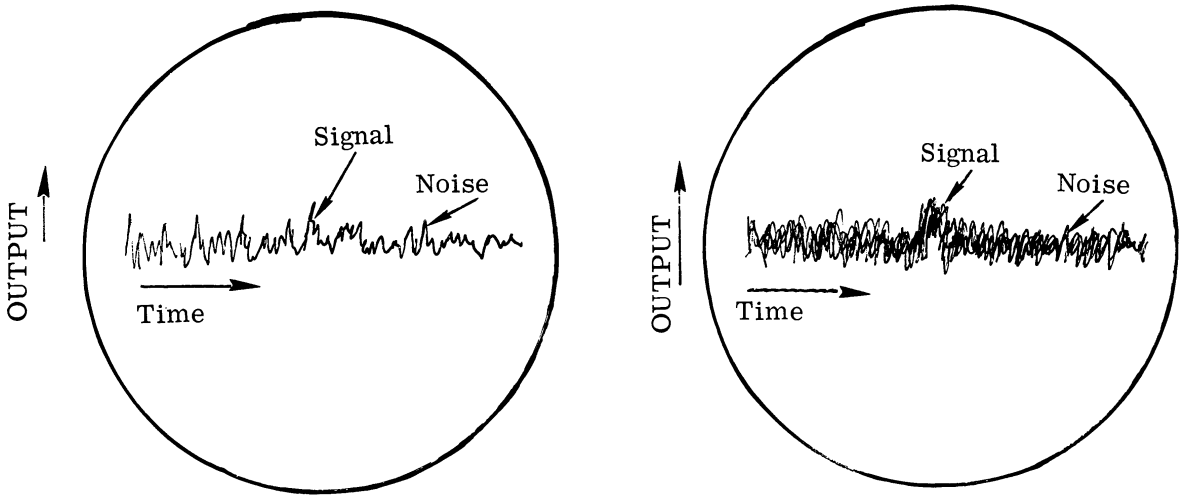


FIGURE 2-2. OSCILLOSCOPE TRACE OF RADAR SIGNALS

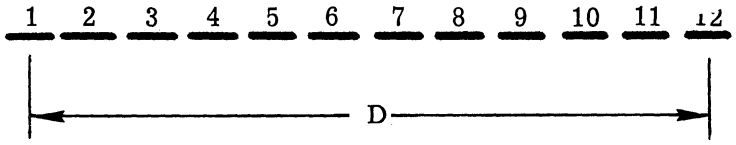
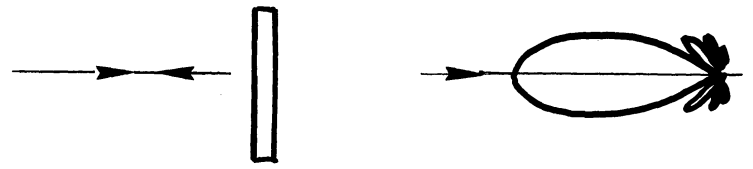
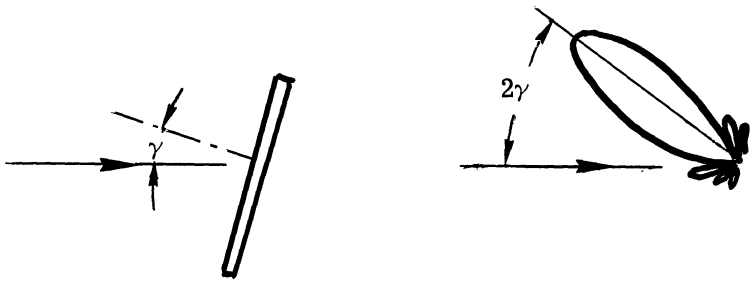


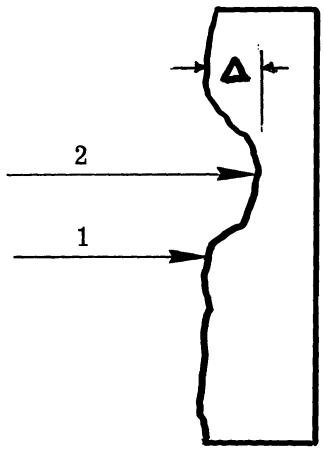
FIGURE 2-3. PRINCIPLE OF SYNTHETIC RESOLUTION



(a)



(b)



(c)

FIGURE 2-4. REFLECTION FROM A PLATE

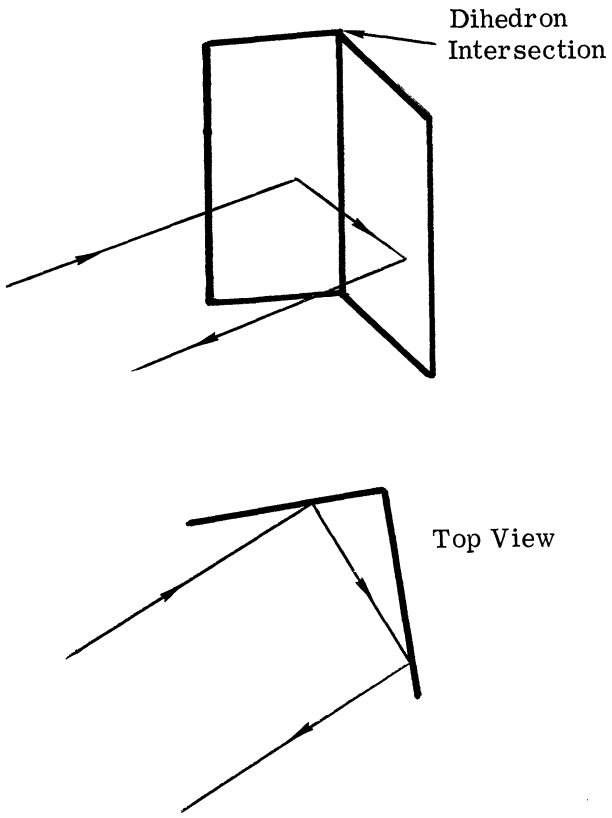


FIGURE 2-5. REFLECTION FROM A DIHEDRON

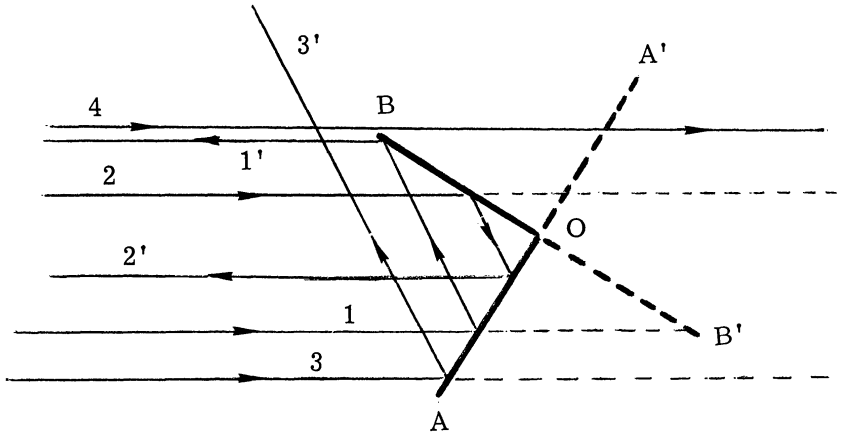


FIGURE 2-6. RAY TRACING FOR A DIHEDRON

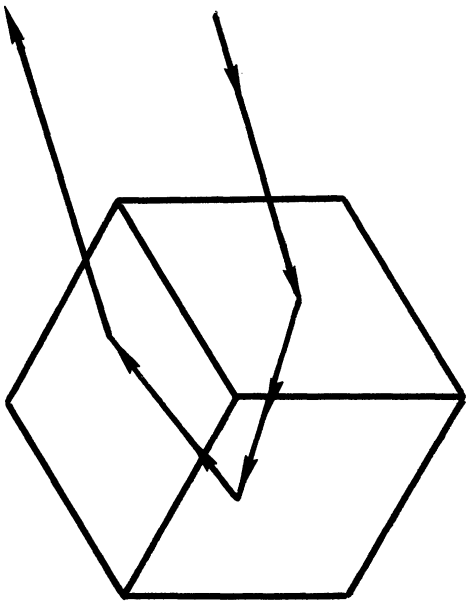


FIGURE 2-7. REFLECTION FROM CORNER REFLECTOR

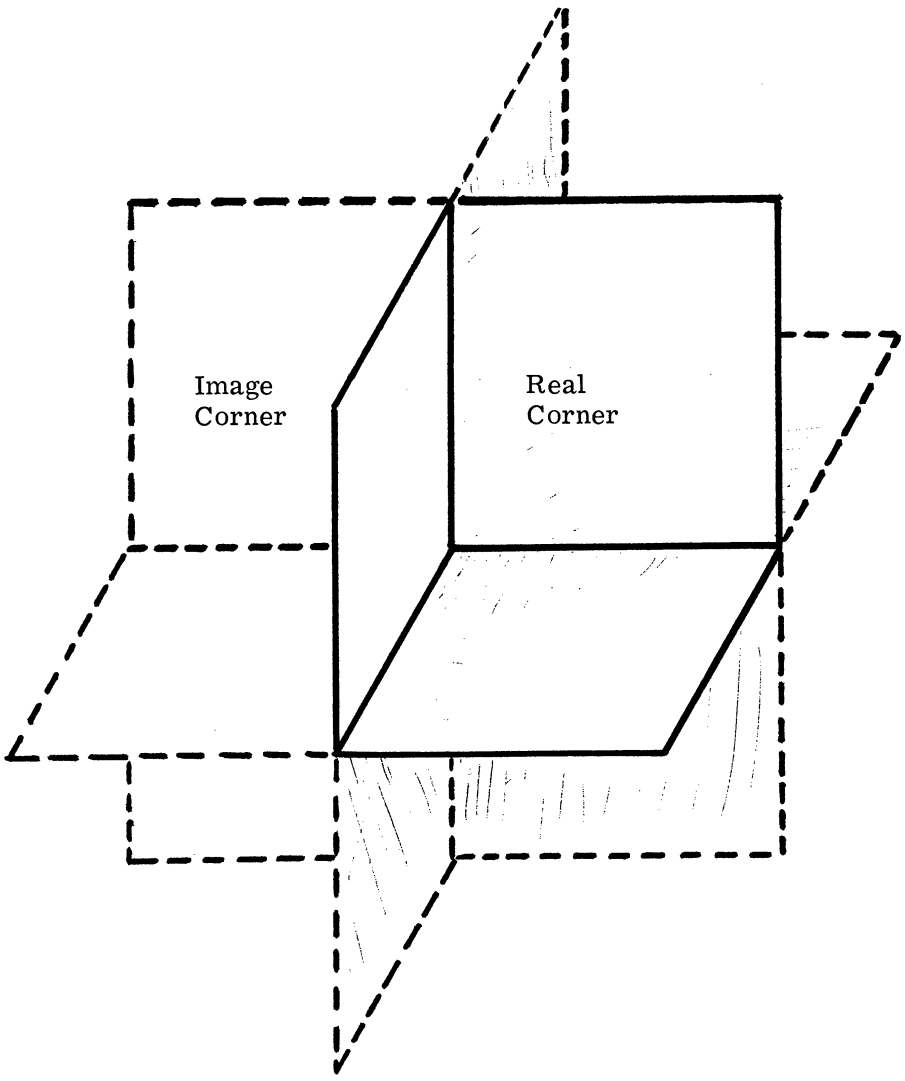


FIGURE 2-8. IMAGE CORNERS

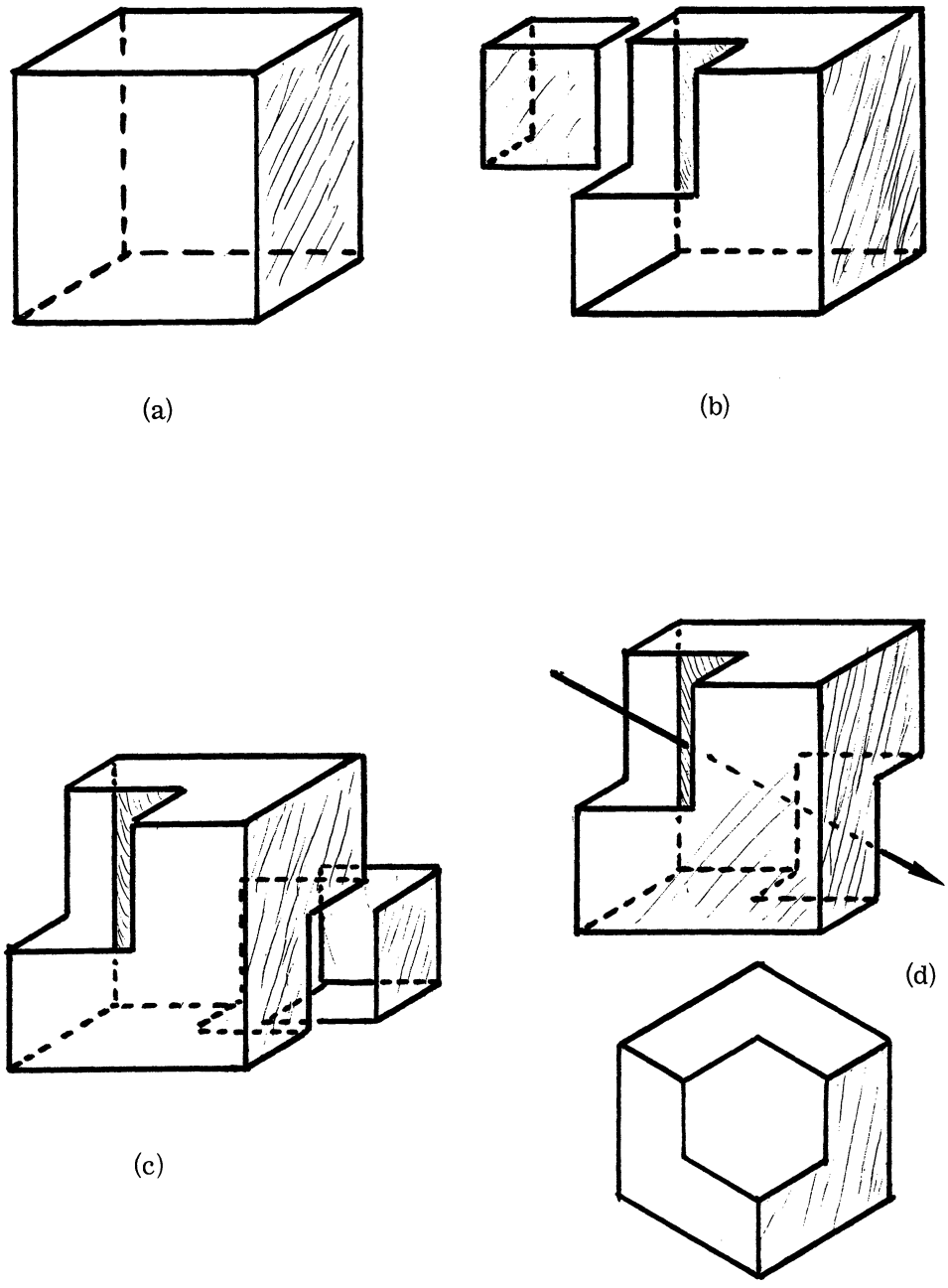


FIGURE 2-9. EFFECTIVE AREA OF CORNER REFLECTOR

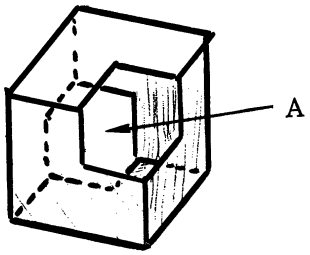


FIGURE 2-10. REDUCTION OF EFFECTIVE AREA WITH ASPECT

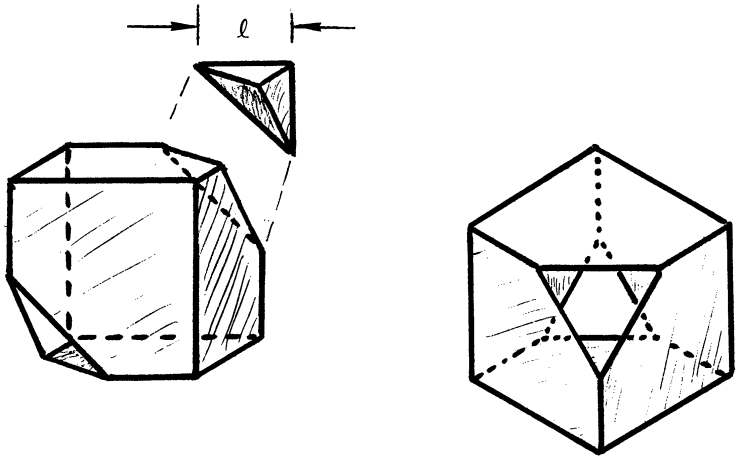


FIGURE 2-11. CORNER REFLECTOR WITH TRIANGULAR SIDES

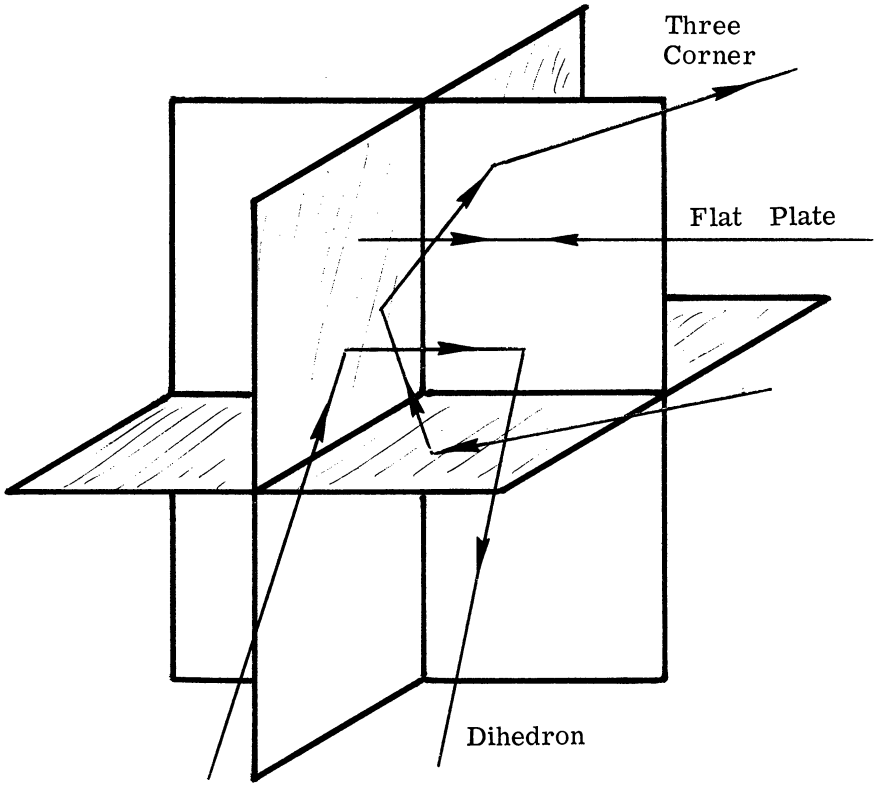


FIGURE 2-12. PRACTICAL RETROREFLECTOR. Reflections occur in all three ways; corner, dihedron, and flat plate. It is retroreflective for all incident angles.

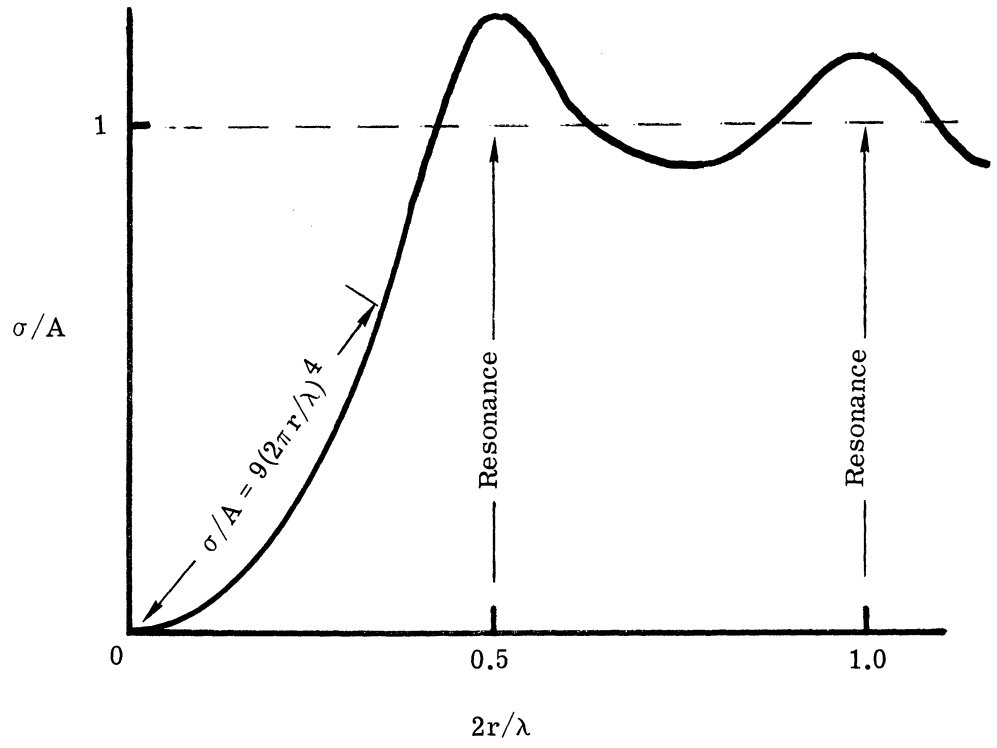


FIGURE 2-13. RADAR CROSS SECTION OF A SPHERE

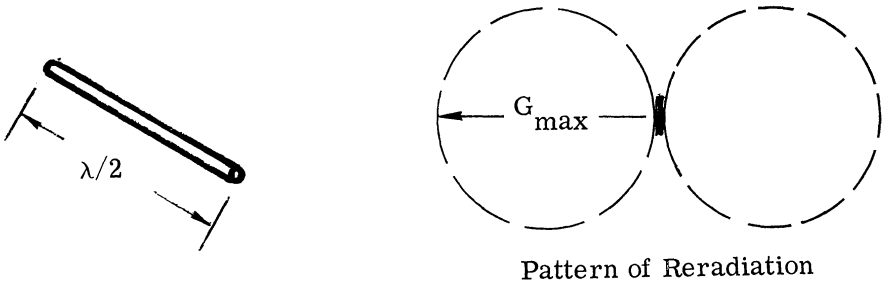
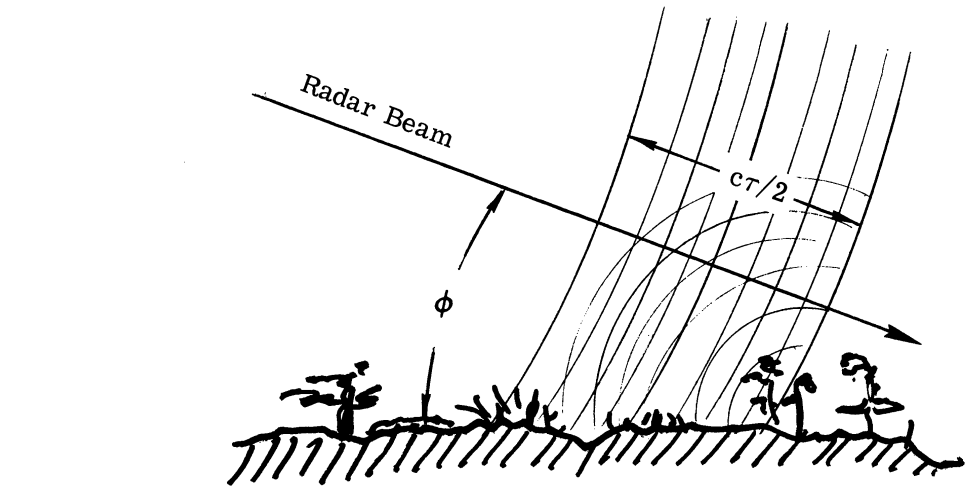
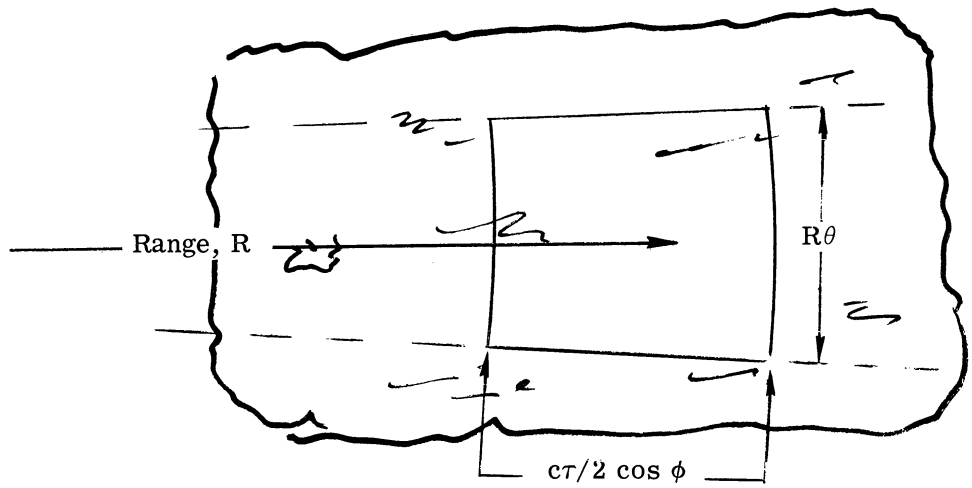


FIGURE 2-14. RE-RADIATION PATTERN OF A RESONANT HALF WAVELENGTH DIPOLE



Side View



Top View

FIGURE 2-15. GEOMETRY OF REFLECTION FROM A GROUND PATCH

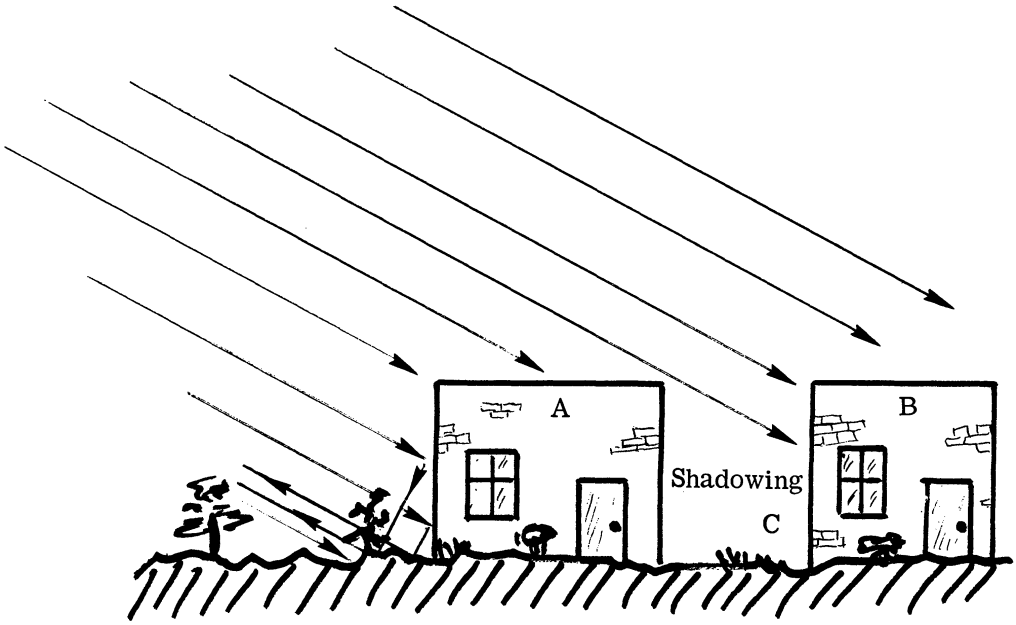


FIGURE 2-16. RADAR SHADOWS

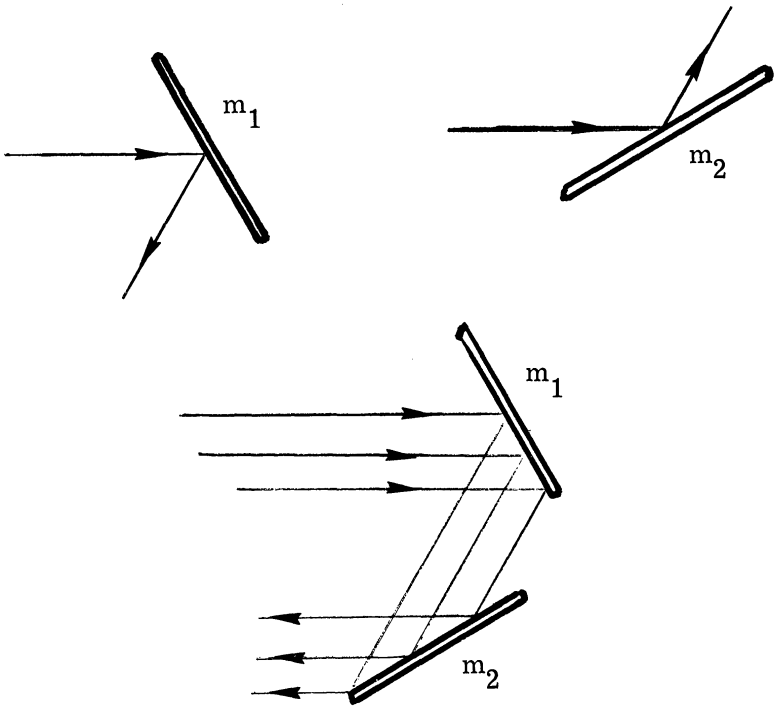


FIGURE 2-17. MULTIPLE REFLECTION

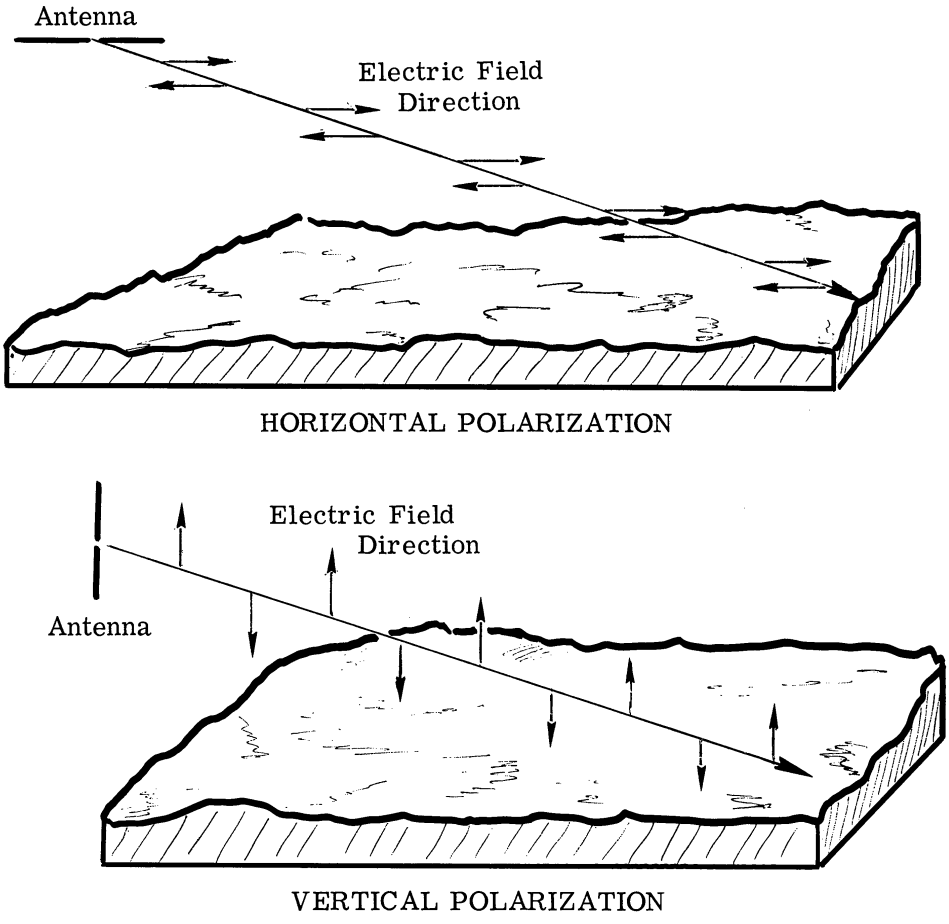
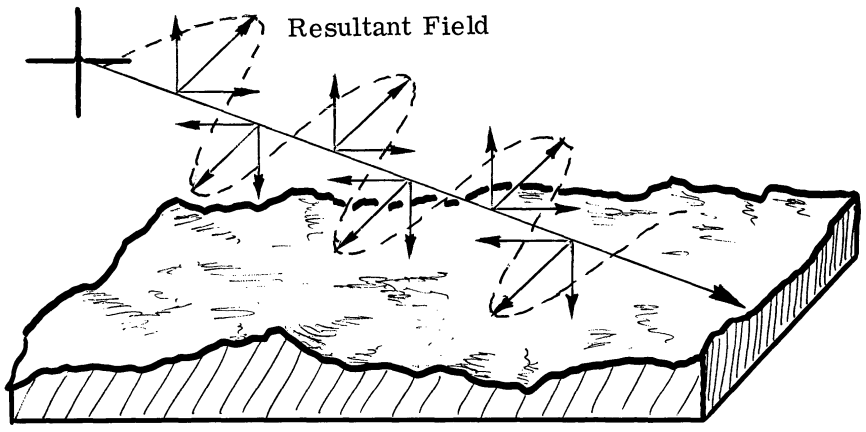
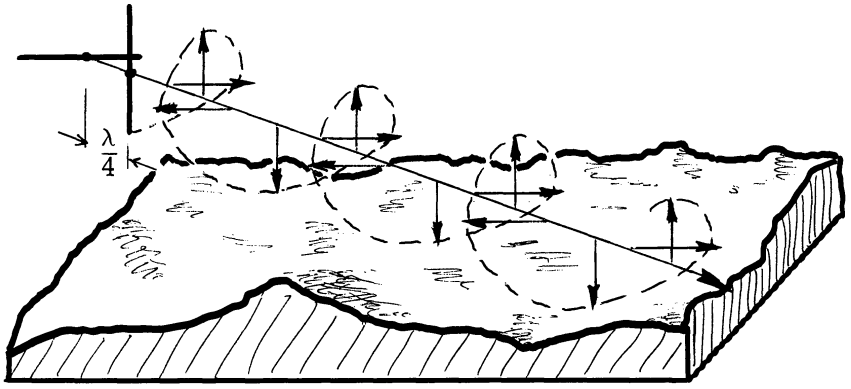


FIGURE 2-18. DIRECTION OF POLARIZATION



45 DEGREE LINEAR POLARIZATION



CIRCULAR POLARIZATION

FIGURE 2-19. OTHER USEFUL POLARIZATIONS

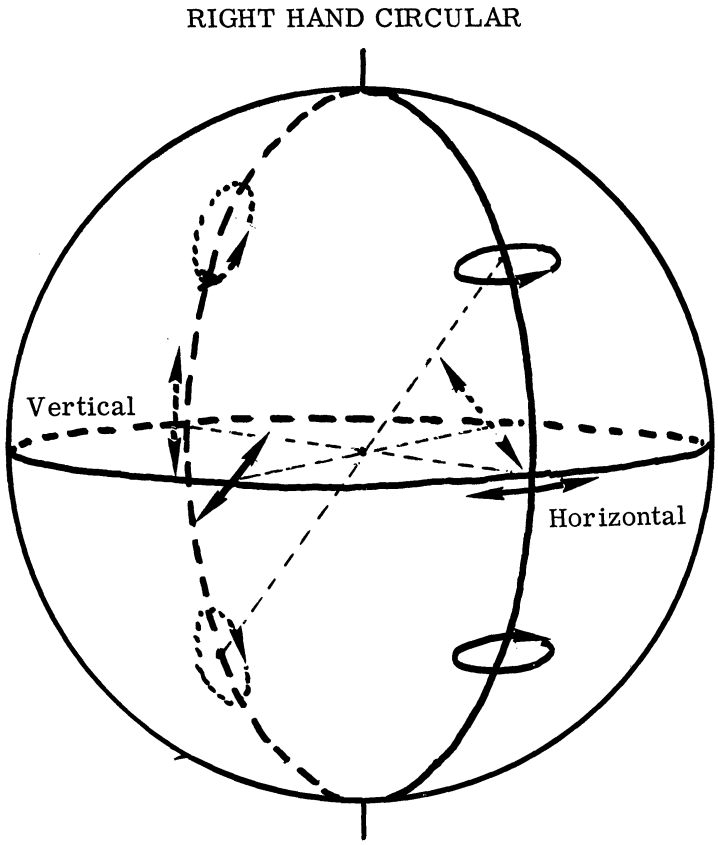


FIGURE 2-20. POINCAIRE SPHERE. All polarization types are represented by the points on the surface of a sphere.

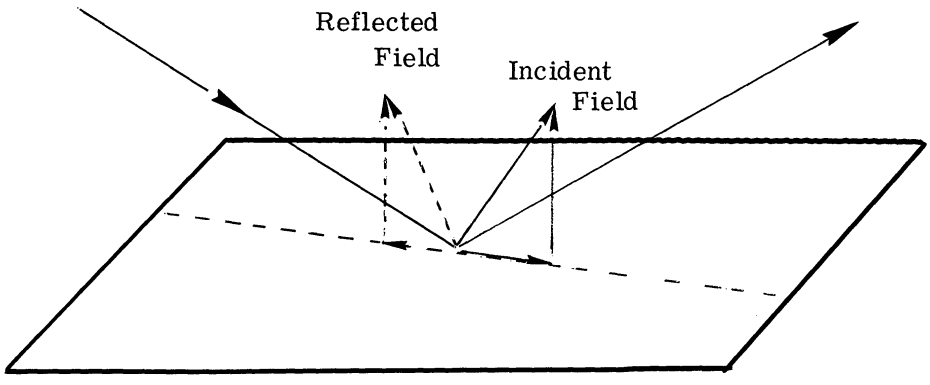
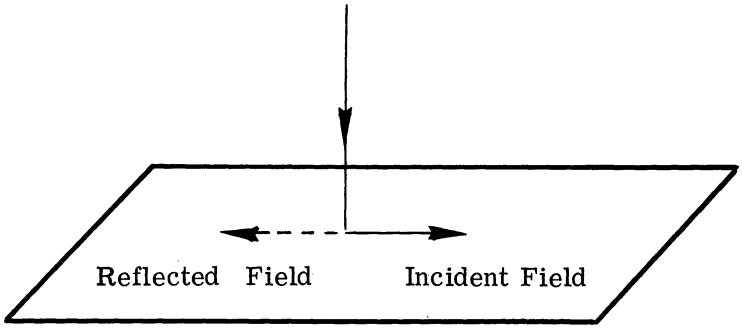
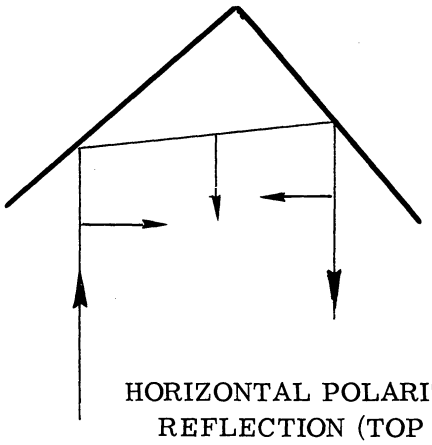


FIGURE 2-21. REFLECTION OF ELECTRIC FIELD COMPONENT FROM PLATE

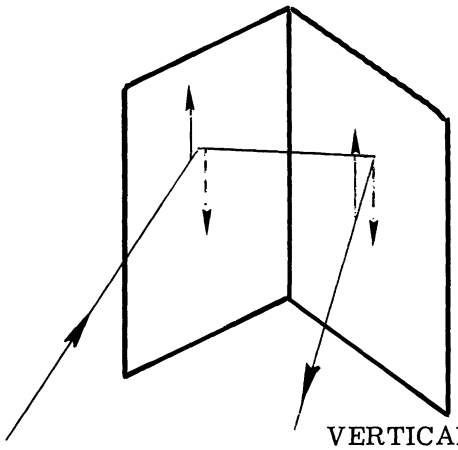


RETRO REFLECTION FROM PLATE

FIGURE 2-22. RETROREFLECTION FROM PLATE



HORIZONTAL POLARIZED WAVE
REFLECTION (TOP VIEW)



VERTICAL POLARIZED
WAVE

FIGURE 2-23. RETROREFLECTION FROM A VERTICAL DIHEDRON

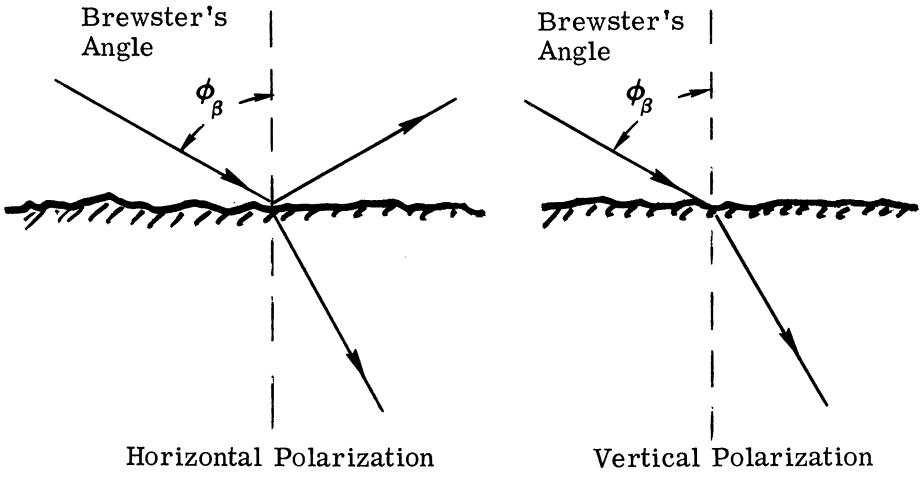


FIGURE 2-24. BREWSTER'S ANGLE

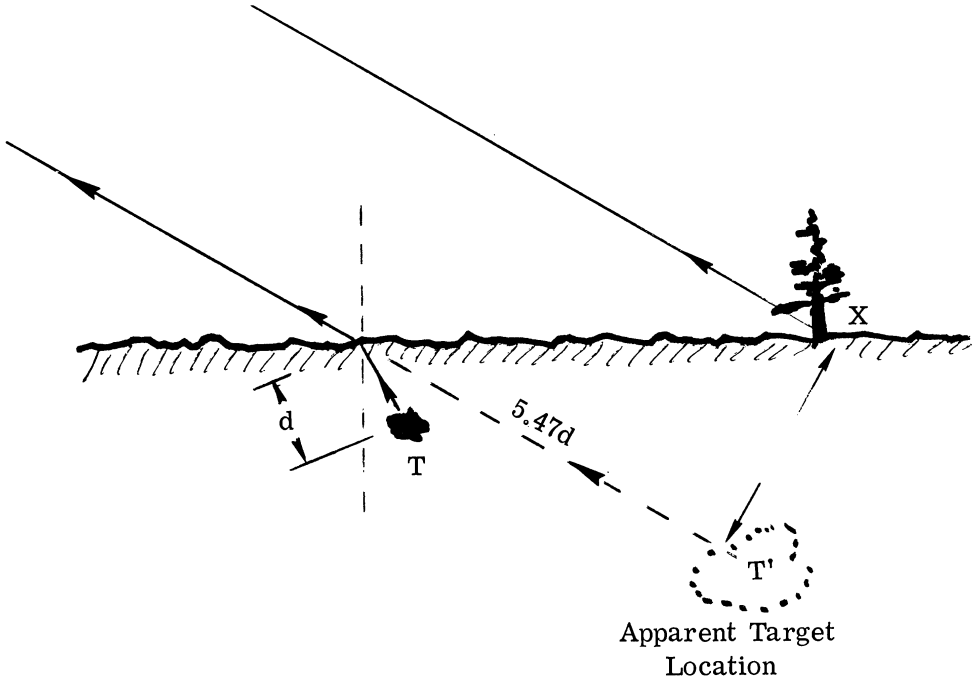


FIGURE 2-25. RAY TRACE BELOW GROUND

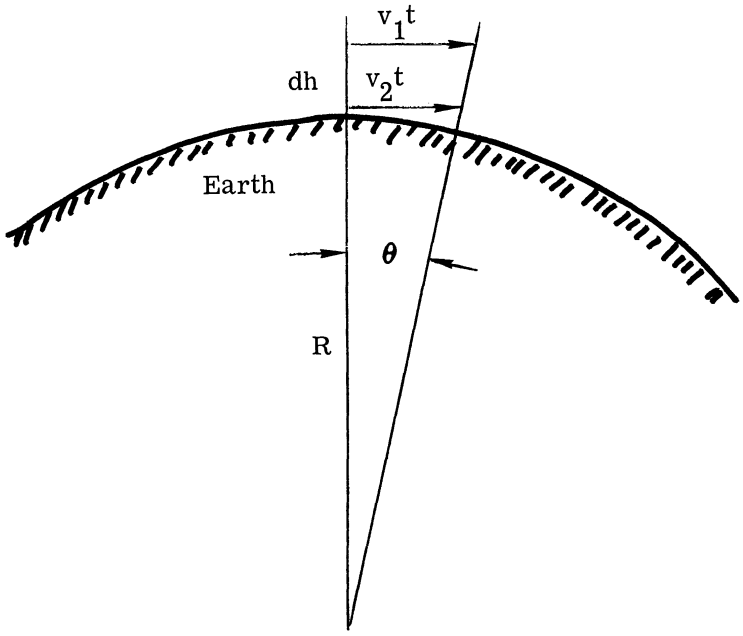


FIGURE 2-26. DIFFERENTIAL VELOCITY OF PROPAGATION

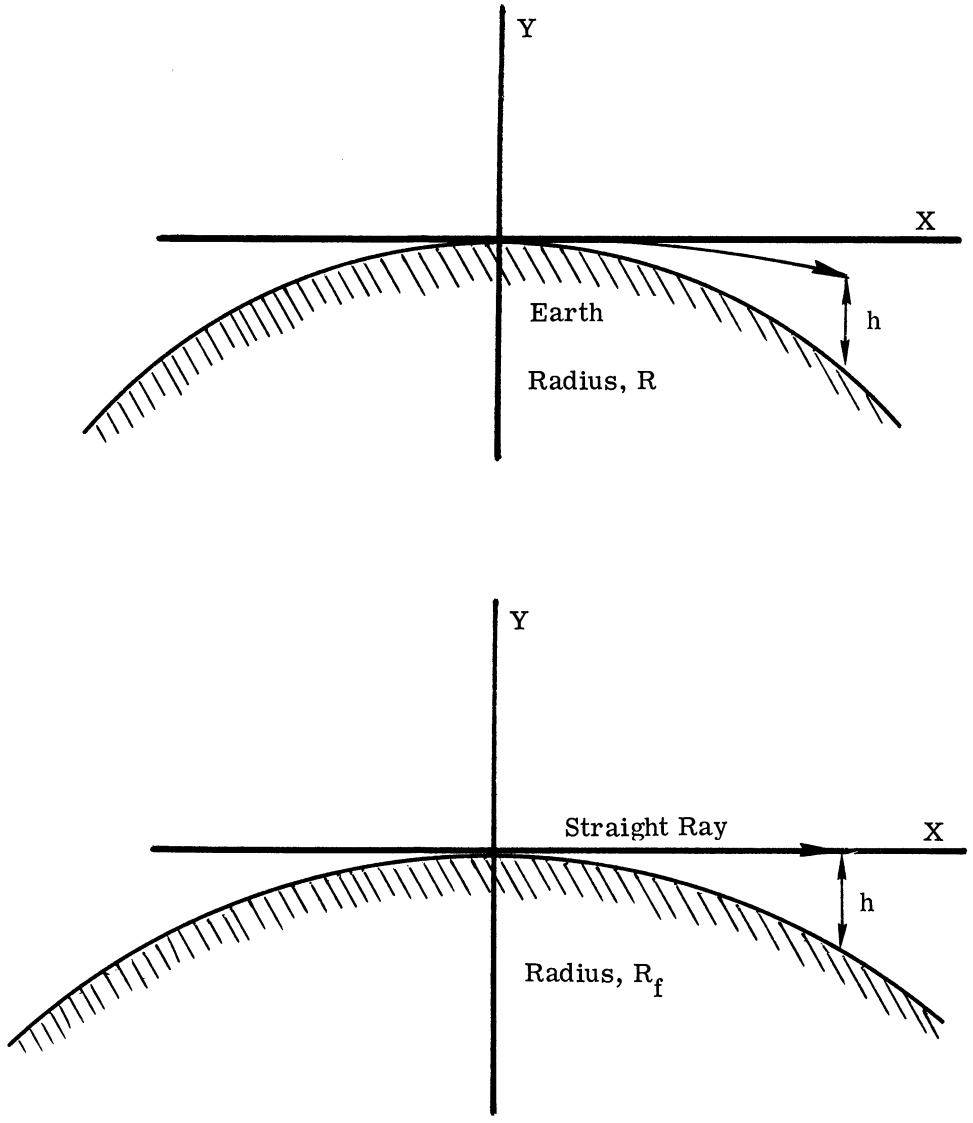


FIGURE 2-27. USE OF FICTICIOUS EARTH RADIUS TO VISUALIZE REFRACTION DUE TO THE ATMOSPHERE

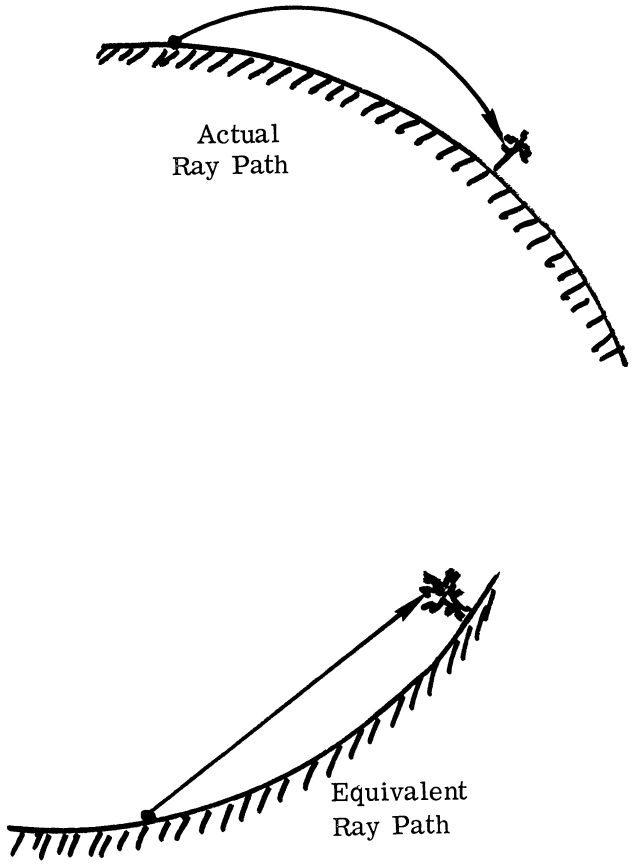
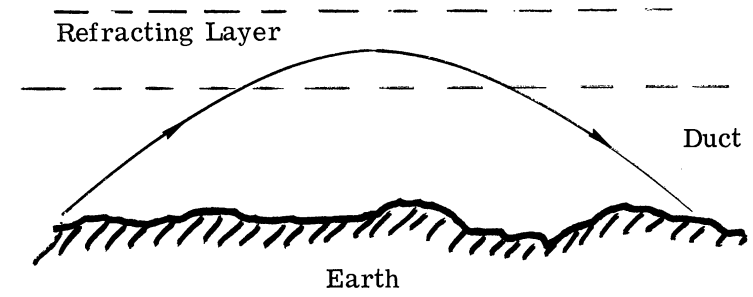


FIGURE 2-28. RADAR IMAGE



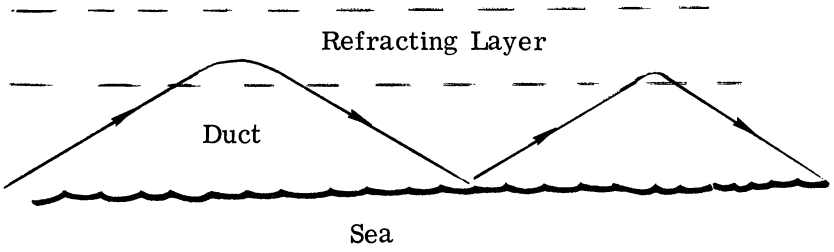


FIGURE 2-29. DUCTING

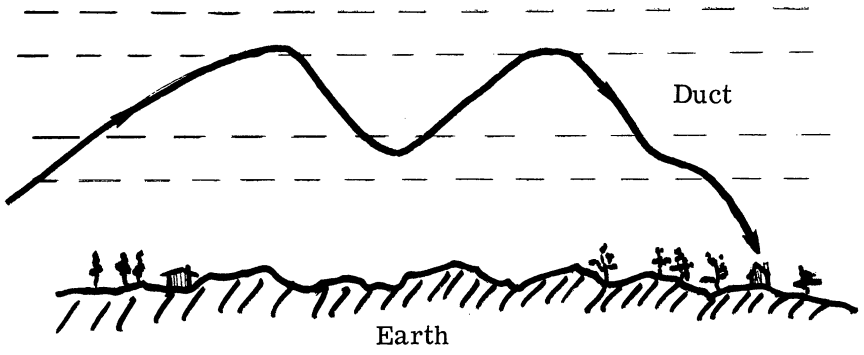


FIGURE 2-30. DUCTING ALOFT

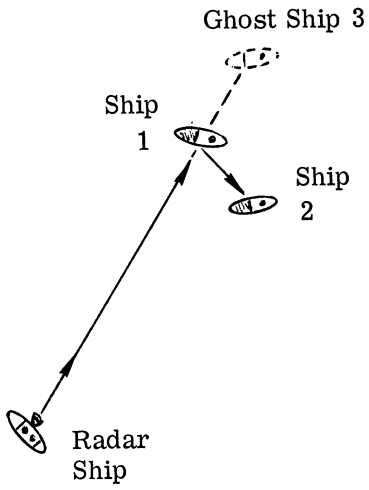


FIGURE 2-31. RADAR GHOSTS

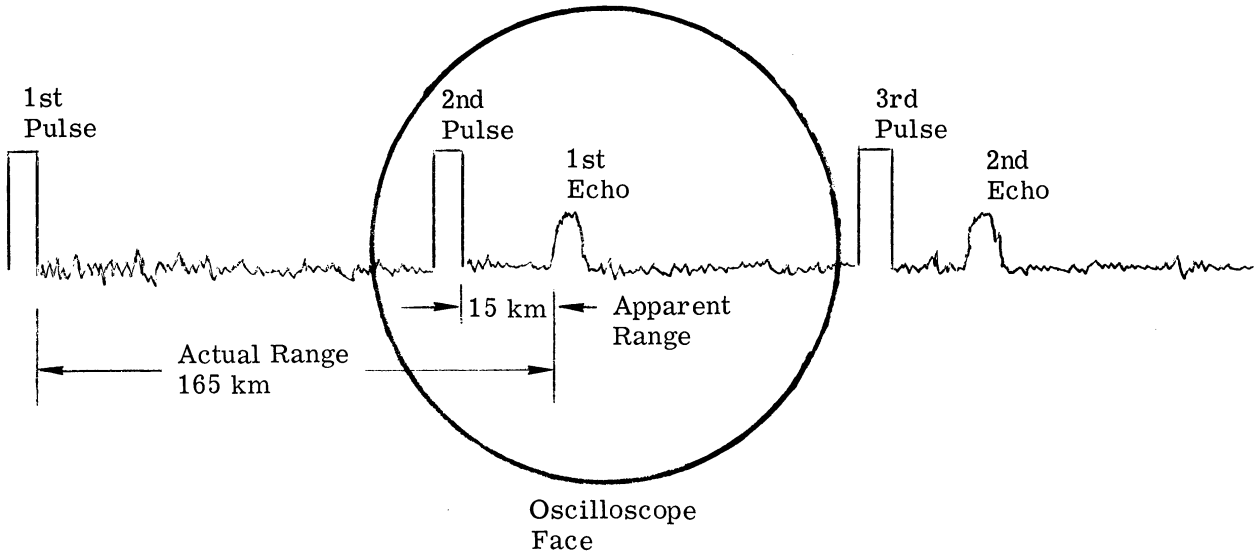


FIGURE 2-32. SECOND GO-AROUND ECHOES

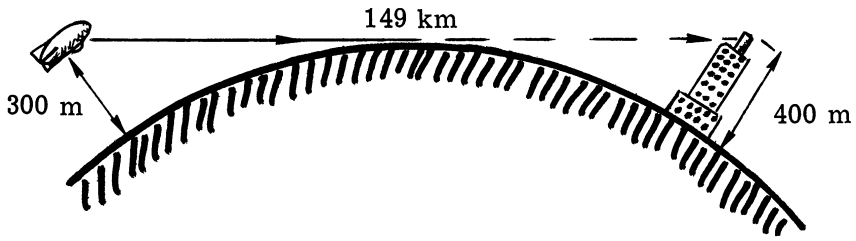


FIGURE 2-33. POSSIBLE CAUSE OF SECOND GO-AROUND ECHO

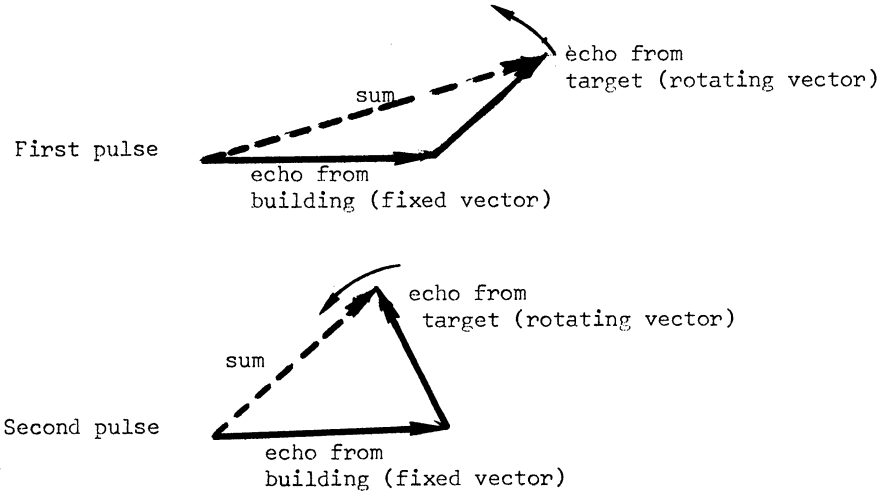


FIGURE 2-34. PRINCIPLE OF NONCOHERENT MTI

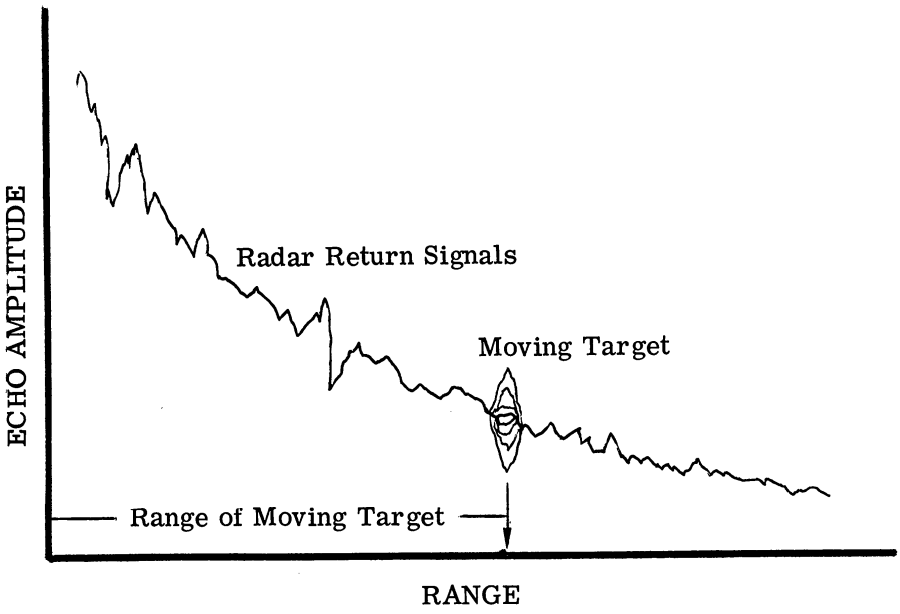


FIGURE 2-35. A-SCOPE MTI

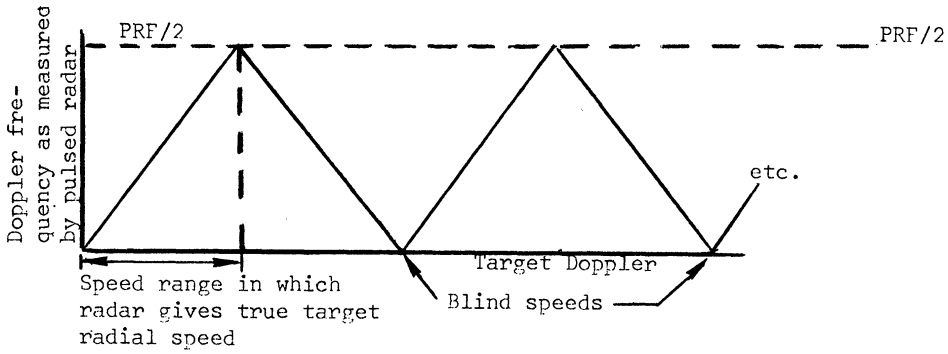


FIGURE 2-36. DOPPLER FREQUENCIES MEASURED BY PULSED RADAR

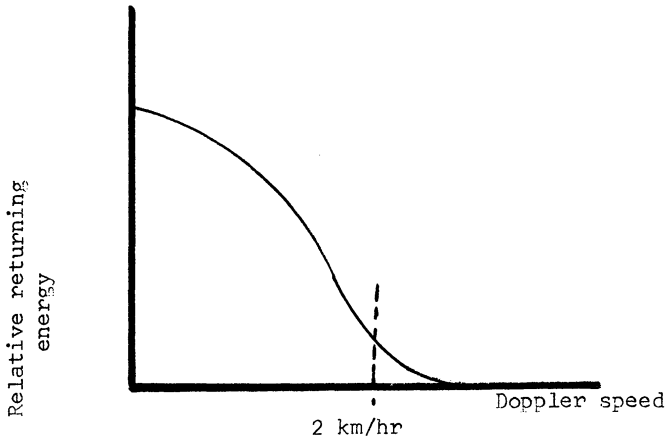


FIGURE 2-37. DOPPLER RETURNS FROM MOVING TREES

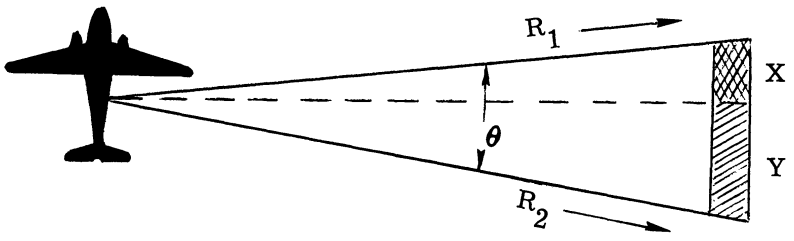


FIGURE 2-38. SIDELOOKING AIRBORNE RADAR

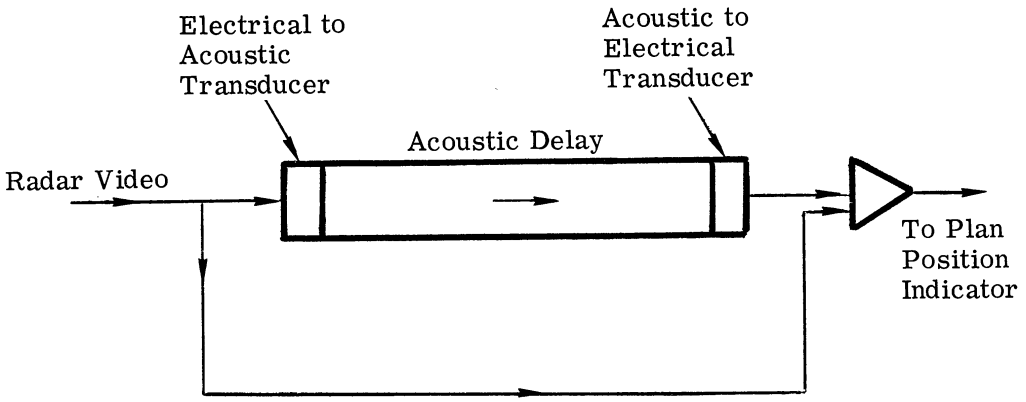


FIGURE 2-39. BASIC ELEMENTS OF DELAY-LINE MTI

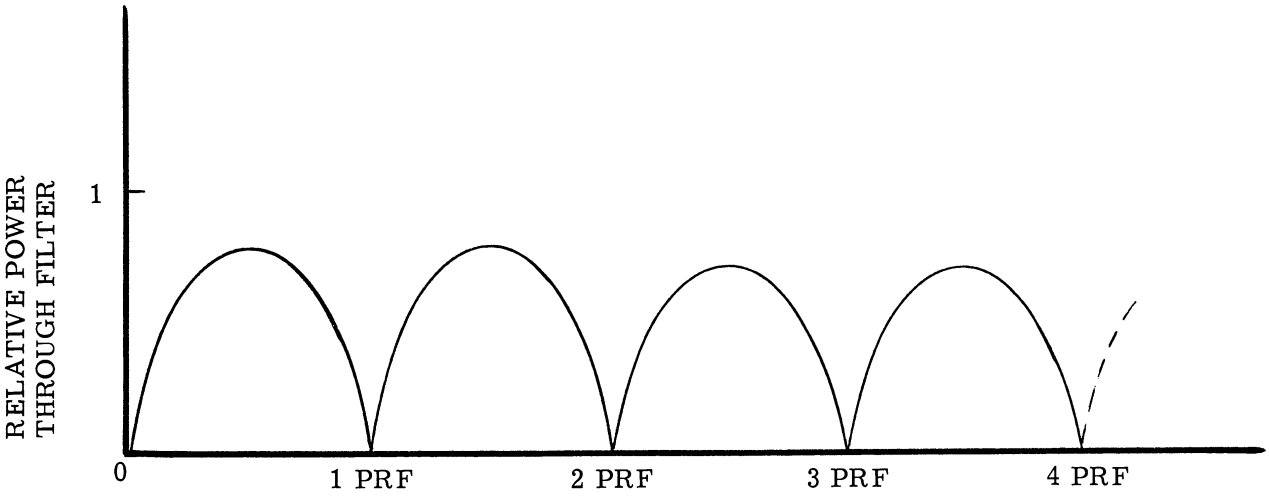


FIGURE 2-40. FREQUENCY RESPONSE (AMPLITUDE) OF DELAY-LINE CANCELLER

SECTION VII
DATA HANDLING AND PROCESSING*

* This section was composed by G. Suits guided by the original notes.

1.0. INTRODUCTION

The term data processing refers to the procedures, recipes, and computations which are to be applied to the original data acquired by remote sensing device in order to provide data concerning how much of what is where at what time. In the case of photogrammetry used for the purpose of making annotated contour maps, the data being processed is the collection of photographic prints with the ancillary data concerning altitude, sun angle, camera focal length, photograph number, date, general locality of the flight line, and location of the survey points relative to the map grid as well as some photo interpretation and documented information permitting the classification of major features to be annotated. The map is the end product of processed data. The end product of processed data is itself data in some new combined form. The use of this processed data to arrive at decisions or conclusions having economic or scientific benefits requires analysis of this data by the user. The design of a remote sensing mission is best guided by the expected analysis in which the processed data is to be used. The nature and purpose of the analysis can best dictate the desired data processing procedures, recipes, and computations which in turn establish the criteria for the form of unprocessed data, the remote sensing apparatus, and the acquisition plan.

Since the final data resulting from data processing can always, in principle, be verified as to its accuracy by detailed tests or observations at the site being sensed, the performance of a complete remote sensing system can be measured by comparing the final processed data with the "ground truth" independently of the particular analysis which is to follow. The criterion establishing whether or not such performance is sufficient will be governed by the requirements of the analysis.

In practice, it is not feasible to test the performance of a remote sensing system for every possible region of the earth where it might be used. Nevertheless, some measure of performance is essential in order that subsequent analysis may be performed with confidence. The accepted procedure is to establish "representative test sites" selected in such a way that one can believe that the tests sites are a realistic microcosm of all sites of interest. Then we will assume that the performance of the remote sensing system used over the test sites will be the measure of performance over any site of interest. Such a measure of performance will involve statistical concepts, notably the concept of probability, probability densities, and variance.

In the following sections a review of the terminology of probability is given. A probabilistic definition of system performance is presented. The manner in which multi-attribute data in remote sensing systems can be processed is explained in more detail using multispectral attributes

as an example. A contrast ratio study illustrates a simplified approach using the photographic process. Radar polarization and aspect diversity systems are discussed. The problems of data storage and retrieval are reviewed.

2.0. REVIEW OF PROBABILITY

The feeling of people about the likelihood of a hypothesis being correct or of the likelihood of an event to occur in the future is formalized by the mathematics of probability. In remote sensing, the only indications which we have concerning the presence of something of interest at a remote location are the signals received by our remote sensing devices. Along with these signals we attempt by means of data processing to add the contextual knowledge about the area under investigation in order to establish with some degree of probability the identity of the materials which reflected or emitted the detected radiation. In order to make maximum use of the sensed data, the formalized logic of probability is used extensively.

A review of this formalization is presented in order to establish the vocabulary, nomenclature and concepts which will be needed in the sections which follow.

The term "event" is given to any outcome of a trial or experiment. In tossing a coin one expects one of two possible events to occur — heads up or tails up. An "event" is any one of the distinguishable outcomes. An event may be either the concurrent occurrence of a set of radiative attributes of a material in the field of view of a sensor, the concurrent occurrence of a set of radiation field properties at the aperture of a sensor, the concurrent occurrence of a set of signal outputs of a sensor, or the classification of such output signals in conjunction with ancillary data to identify the nature of the substance in the field of view of the sensor.

Mutually exclusive events are a set of possible outcomes such that the occurrence of one of this set precludes the occurrence of any other of that set on a given trial. For example, the events (heads up and tails up) of a coin are considered to be mutually exclusive because a tossed coin cannot land heads up and tails up at the same time.

Wholly exhaustive events are a set of events such that on any given trial at least one of the events must occur. If one precludes the possibility that a coin may land on its edge and stay there, then the events heads up and tails up are wholly exhaustive events for coin tossing trials.

Expectation ratio for an event to occur is the number of times the event did occur in a given number of past trials divided by the number of those trials. Clearly the expectation ratio of any event on the first trial must be either one or zero, on the second trial it must be one, one-half or zero, etc. The limit of the expectation ratio as the number of trials becomes infinitely large is the probability of the event to occur. In any experiment in remote sensing we will be making only a small number of trials over test sites. These trials should be made so that the expectation ratio will be a good enough measure of the probability for practical purposes.

The symbol, P(A), stands for the probability that an event labeled A will occur.

If two or more distinguishable events can occur on the same trial, we can ask with what probability both events occur together. This would be called the joint probability and would be symbolized P(AB). Two events A and B are said to be independent if and only if P(AB) = P(A)P(B). We may also ask what is the probability of B to occur only in cases where A also occurs. Such a probability is called a conditional probability and is symbolized by $P(B|A)$ which is read "the probability of B given A." The relation between joint probability and conditional probability is

$$P(A)P(B|A) = P(AB)$$

Given an event A, it is usually possible to define another event B which is the nonoccurrence of A. This can be symbolized as $B = \bar{A}$ (read as B equivalent to "not A"). The two events A and $\bar{A}$ are always taken to be mutually exclusive and wholly exhaustive. Either A occurs or it does not occur. No other possibilities exist. The negation symbol (a line over the symbol for the event) can be extended to joint events so that the joint event of A and B is AB and the failure of AB to occur is $\overline{AB}$.

For two mutually exclusive and wholly exhaustive events such as A and $\bar{A}$ we can write

$$P(\bar{A}) = 1 - P(A).$$

When making measurements of radiance and irradiance, we may wish to express the probability of occurrence of certain radiance or irradiance values under different circumstances. These values have a continuous range. However, one may consider a range of values to be divided into finite intervals so that the occurrence of a certain radiance value that falls within one of the intervals of values is the recognized event. Naturally, as the size of the interval becomes smaller and the number of possible intervals become proportionally greater, the probability that a radiance value will fall into a particular interval becomes less. Thus the probability that a continuous variable will be found to fall in a particular interval is expressed in terms of a probability density, which is the probability per unit interval size that a reading of a continuous quantity will have a specified value.

The probability density is expressed by lower case p. For example, $p(v)\,dv$ symbolizes the probability that a voltage will have a value within a small interval, dv, centered at the value v. Clearly, if all possible voltage intervals are considered, then

$$\int_{-\infty}^{\infty} p(v)\,dv = 1$$

The probability of the voltage value falling in one interval per trial precludes it from falling into any other and all possible intervals are considered so that the occurrence of different voltage values must be mutually exclusive and wholly exhaustive events. For that reason the sum of probabilities for all occurrences must be equal unity.

Parallel to the finite event case we may wish to consider the joint probability density of two voltage readings from two separate channels. Thus

$$p(v_1, v_2)\, dv_1\, dv_2$$

is the joint or "bivariate" probability that the voltage in channel one will fall in the interval dv_1 centered at v_1 and also that the voltage in channel two will fall in the interval dv_2 centered at v_2 at the same time.

Another useful concept to quantify the relative likelihood of two events to occur is called the likelihood ratio. The relative likelihood that event A will occur as opposed to B occurring on a given trial is P(A)/P(B). The two events A and B need not be wholly exhaustive events. Another event C might also be possible but when event C occurs no importance is attached to it. The likelihood ratio is not a probability because it is not the limit of an expectation ratio for increasing numbers of trials.

3.0. DISCRIMINATION AND SYSTEM PERFORMANCE

Every remote sensing system which begins by remotely sensing either a static or dynamic force field or both is not complete until some claim is made from the data as to how much of what is where. A remote sensing device by itself does nothing but receive and record. The data processing part of the remote sensing system must produce the necessary annotations. The act of establishing the correct annotations from the set of incoming signals is the process of discrimination.

Discrimination may be defined as the act of sorting by designating with sufficient accuracy the set membership of the elements of a population. In this case the elements are the objects to be sensed and the set membership is the annotation to be applied to the signals from the objects.

This definition implies four supporting assumptions.

1. The elements of the population are defined as separable units.
2. There exist at least two sets defined by certain attributes of the elements of each set (otherwise there can be no sorting). These attributes are called the defining attributes of the sets.
3. There is at least one infallible means of determining the set membership of every element from certain attributes of the element—the defining attributes if no other.
4. A criterion for sufficient accuracy is defined.

The problem most often required of a discrimination system is to perform discrimination using attributes of elements which are not the defining attributes. A quality-control method which requires destructive testing of every element produced to obtain defining attributes of the units is impractical. In particular, testing for lethally poisonous mushrooms by using the defining attribute of "causing death when eaten" would not only provide no mushrooms for market, since all must be eaten for testing, but would also produce a high turnover and low morale in the personnel.

If the population of elements is sufficiently large so that the concept of probability can be applied meaningfully, one can define a criterion for sufficiently accurate sorting by statistical measures. Suppose one chooses two mutually exclusive sets—target and not-target. There is presumably some infallible means using defining attributes for establishing set membership, target or not-target, of every element. The discrimination system is to use other attributes of the elements which may not all be defining attributes to establish set membership. Thus, it is possible for the system to designate a target element as either a target element or a not-target

element, or to designate a not-target element as a target or not-target element. By recourse to the defining attributes one can always check whether the designation of set membership was correct or not. The results of a lengthy test of a discrimination system could be displayed in chart form as shown in figure 3-1.

Assume that we were interested in oak trees as the only item of interest. Every oak tree would be a target. Every other item would be a "not"-oak tree or a not-target. The final processed data showing the annotation of oak trees using the remote sensing system over the test sites can be taken to the test sites and checked for accuracy using the defining attributes for the oak trees. When an oak tree was correctly annotated, a scoring mark is entered into the upper left hand corner of figure 3-1. When an item was correctly annotated as a not-oak tree or simply left unmarked, a mark is entered in the lower right hand corner. There are two other scores which are errors. The upper right hand scores the annotations for oak trees when an oak tree did not exist and the lower left hand scores the oak trees which failed to be annotated as an oak tree.

This score sheet is sufficient data to specify the accuracy of a complete remote sensing system for oak trees. If the test sites do constitute a microcosm of the oak tree world, then there is good reason to expect the same statistical behavior over unknown areas for the remote sensing system.

It is convenient to present these scores as ratios. Clearly, from the score sheet, there were 23 + 10 = 33 oak trees in the test site because ground truth lists these as being either correctly or incorrectly annotated. Also there were 8 + 1584 = 1592 other items which were not oak trees. We may say from this data that, given a true oak tree (by ground truth), the probability that it was annotated correctly is 23/33. Therefore $P(T_A \mid T) = 23/33$. Here, T_A, is the target annotation event; T is the oak tree occurrence event. Also the probability that, given an oak tree, the annotation would be incorrect is 10/33. Therefore, $P(\overline{T}_A \mid T) = 10/33$. In similar manner $P(\overline{T}_A \mid \overline{T}) = 1584/1592$ and $P(T_A \mid \overline{T}) = 8/1592$.

It is not difficult to see that by definition $P(T_A \mid T) + P(\overline{T}_A \mid T) = 1$, and $P(T_A \mid \overline{T}) + P(\overline{T}_A \mid \overline{T}) = 1$. Thus, two independent probabilities are required to describe the statistical behavior of the remote sensing system. We may choose $P(T_A \mid T)$ and $P(\overline{T}_A \mid \overline{T})$ as the two to use. We may graph these probability values as a system score point on a graph as shown in figure 3-2. Perfect discrimination would be plotted at (1, 1). Completely the reverse (all wrong) is (0, 0). Chance scores would lie on the line between (1, 0) and (0, 1). A sufficient region may be defined as shown. The measure of sufficiency can be extended to apply to any number of mutually exclusive, wholly exhaustive, sets so that for N sets A, B, C, . . . K,

$$P(A|A) + P(B|A) + P(C|A) + \ldots + P(K|A) = 1.$$

$$P(A|B) + P(B|B) + P(C|B) + \ldots + P(K|B) = 1$$

$$P(A|K) + P(B|K) + P(C|K) + \ldots + P(K|K) = 1$$

The performance of a discrimination device is specified by N(N - 1) numbers. A region of sufficiency can be defined by a region in a space of (N)(N - 1) dimensions.

3.1. ILLUSTRATION OF USE OF TEST SITE RESULTS

Suppose that over a test site which represents a microcosm of unknown territory the remote sensing system were tested and the data tabulated as shown in figure 3-3. Instead of writing in the usual score numbers it is convenient to write the score normalized to the total number of grids surveyed. Then d would be the number of correctly annotated targets per grid and similarly for m, f, and b so that

$$d = DP(T_A|T),$$

$$m = DP(\bar{T}_A|T),$$

$$f = (1 - D)P(T_A|\bar{T}),$$

$$b = (1 - D)P(\bar{T}_A|\bar{T}).$$

where D is the average number of targets per grid and (1 - D) is the average number of not-targets per grid. (Each specific grid is either a target or not-target containing grid.) The value of D for the test site is known and tests with the remote sensing system provides the values of d, m, f, and b and consequently the statistical behavior of the remote sensing system, $P(T_A|T)$ and $P(\bar{T}_A|\bar{T})$, is now known. The other probabilities are derived from these two.

Now the remote sensing system is used over unknown territory. It is desired to estimate the density D' of targets per grid from the remote sensing data. If this unknown site were an average site and the statistical behavior of the remote sensing system did not change, then we would expect that the new score numbers d', m', f', b' would be

$$d' = D'P(T_A|T),$$

$$m' = D'P(\bar{T}_A|T),$$

$$f' = (1 - D')P(T_A|\bar{T}),$$

$$b' = (1 - D')P(\bar{T}_A|\bar{T}).$$

However, the data which is available from the survey over unknown territory provides only $N' = d' + f'$, the number of target annotations per grid, and $M' = m' + b'$, the number of not-target annotations per grid. Thus, $N' = d' + f' = D'P(T_A|T) + (1 - D')P(T_A|\bar{T})$. Solving for D' the result is

$$D' = \frac{N' - P(T_A|\bar{T})}{P(T_A|T) - P(T_A|\bar{T})}.$$

The value D' is the best estimate of target density over the unknown territory. The confidence which one could have concerning the accuracy of the estimate will depend upon the performance of the remote sensing system. If the system score falls upon the chance line, then $P(T_A|T) - P(T_A|\bar{T}) = 0$ so that D' becomes indeterminant as would be expected. If $P(T_A|T) = 1$ and $P(T_A|\bar{T}) = 0$, as would be the case for perfect discrimination, then $D' = N'$ exactly. For an analysis which uses the estimated target density as one of the factors the accuracy of the estimate that is needed will establish the criterion for the needed score point of the remote sensing system.

3.2. CHOICE OF ATTRIBUTES AND MULTI-ATTRIBUTE SYSTEMS

A remote sensing system rarely has available the defining attributes of a target for the purpose of establishing set membership. Some one or more attributes may be available at the remote location which can be used as reasonably successful substitutes. One of the most important questions which must be answered is which attributes should be measured by the remote sensing device in order to achieve maximum performance. Intuition leads us to believe that the more attributes that are measured, the more likely discrimination will be sufficient, but there are economic and practical timeliness limits on this number. The reasoning which one may apply toward answering this question is best illustrated by the following simple examples.

Let the elements of the population be the faces of a deck of playing cards placed face up in white light on a smooth black background. We may select any of a variety of sets which would be interesting for sorting purposes. The defining attribute of each card is the card label written in the corner of each card. We may define the set of all diamonds and hearts as the set we want and all others are the "not-diamonds and hearts." Intuition might lead us to try a simple color filter system for our discrimination to distinguish the available attribute—the red color as opposed to the black color. A red filter could be made which would make the effective radiance of the filtered light from the red spot the same as the white background of the rest of the card so that all cards from the ace to the 10 of both hearts and diamonds would be the same but the black cards would present less radiance in proportion to the number of spots on the black card. However, we would find upon test that such a scheme may not work on the face

cards, which all have some red, some yellow and some black ink. The jack of hearts could resemble the ace of clubs, for instance, as far as an apparent radiance would indicate. We may argue that we have not found the best filter and should change filters. Indeed, the concept of the "matched filter" which is used extensively in information theory will lead us to search through our possible filter materials for that one. With a full knowledge of the spectral properties of our population we can follow the sophisticated mathematics in "matched filter" theory and arrive at the very best filter one can possibly use, but such a filter may still not provide us with the sufficient accuracy in the discrimination task which we require. Now, this does not mean that filters are no good. It means only that we have come to the end of the line with the simple implementation we had in mind. We have yet to explore two attribute systems with two channels using two different filters or N attribute systems with N channels or some filter channels and some reticle channels, etc. But now we must face a very difficult decision. which attributes that are available at the receiver should we attempt to measure and use. The red filter measured the red radiance of each card. A mask of some shape would obscure some of each card so that an indication of the geometrical distribution of radiance across the card could be measured. That is certainly another attribute of the card. First, disregard specific implementations and consider only the available attributes. Suppose we consider some N available attributes of each of the elements of the population. We can only use $K < N$ because of practical limitations. Let us ask the question, "Can we proceed to test each one of the N attributes separately, select the K attributes which give the best single attribute system performances and then combine these attributes into a multi-attribute system of K attributes to get the best K attribute system?" The answer is: not necessarily.

The combination of the K best single attribute systems doesn't necessarily lead to the best K attribute system. This can be proved by a single example where it does not work.

Take as the elements of the population circles and squares of all sizes. Make 3/4 of the circles green and 1/4 red. Make 1/4 of the squares green and 3/4 red.

Now choose for attributes, 1. The redness, 2. The greenness, 3. The greatest distance across, 4. The distance around.

Choose as the sets for sorting the circles and the not-circles (squares). It is evident that we can tell circles from squares 3/4 of the time by either attribute one or two, but do no better than chance with attribute three or four. If we were to make a two-attribute system with attributes one and two, we would find no better discrimination than with one alone. But using the two worst attributes together, we can obtain perfect discrimination. There is a necessary connection between the greatest distance across and the distance around through the definition of the geometrical figures.

3.3 A GRAPHICAL VIEW OF DISCRIMINATION

One- and two-attribute systems can be plotted in graphical form so that the basic logic of discrimination may be exhibited. It is also easy and very instructive to actually construct an elementary discrimination system using two attributes.

Consider a target population from which a single attribute v_1 is received for each element. These could be plotted along an attribute scale—a point for each element as shown in figure 3-4. If enough elements are measured, it is possible to determine the probability density, $p(v_1)$, for that population. Given an element of the target population taken at random, the probability that its attribute measure will fall in an interval dv_1 centered at v_1 is $p(v_1)$. The knowledge of $p(v_1)$ is insufficient for establishing v_1 as a good attribute choice for discrimination. Discrimination is a sorting process between elements of one set and elements of another. We must obtain a similar distribution for all not-target elements likely to be found in the regions of interest. Assume such a distribution is shown by the dashed line in figure 3-4. It can be seen that some degree of discrimination is possible because targets tend to have low values of attribute v_1 and not-targets tend to have high values.

Now visualize a simple data processing scheme which consists of a very simple decision rule: If $v_1 \leqq A$ annotate target; if $v_1 > A$ annotate not-target. The sensor device operating over the test site will produce values of v_1 for each grid. When a target grid is examined we will not always annotate correctly since some targets produce a value $v_1 > A$. Thus $P(T_A \mid T) < 1$. In particular, $P(T_A \mid T) = \int_0^A p(v_1, \text{target})\, dv_1$. However, when a not-target grid is examined, a not-target will be annotated because all not-targets have a value of $v_1 > A$. Thus,

$$P(\overline{T}_A \mid \overline{T}) = \int_A^\infty p(v_1, \text{not target})\, dv_1 = 1.$$

The system score point is determined.

One could choose a different decision rule and the result would be a different system score point. If instead of A, one chose C as the decision point, then $P(T_A \mid T) = 1$ and $P(\overline{T}_A \mid \overline{T}) < 1$ would result. At some intermediate point, B, a compromise score point would result. Recall that the sufficient region for score points depends upon the nature of the following analysis. It may be possible to alter the data processing in order to bring the system performance within the sufficient region when another processing would be insufficient. In some cases it is very important to annotate only valid targets even at the expense of missing some. For that purpose, the decision point A would be chosen even though point B would result in many more targets being annotated correctly.

Now suppose that no matter what data processing is done, the system scorepoint will not fall within the sufficient region. It can be concluded that the attribute measure, v_1 contains

insufficient information. No amount of changing variables in plotting such as plotting v_1^2 or e^{v_1} etc. instead of v_1 will change the fact that too many targets and not-targets exhibit the same attribute value and are therefore logically indistinguishable by that attribute alone.

Suppose that the remote sensing device presents two attribute values for each element instead of just one. We can plot the pairs of values v_1 and v_2 for each element as a point in a two dimensional attribute space. In the analogous way we can find for target and not-target elements the joint probability density $p(v_1, v_2)$ for each set. Since two of the dimensions are used for the attribute measures, the values of $p(v_1, v_2)$ will be plotted as the third dimension in contour map form. The result of adding a second attribute measure is shown in figure 3-5. If v_2 were ignored in the data, the decision rule involving A, B, and C would produce the same system score points as before. By using both v_1 and v_2 together we have the possibility of using a new decision rule by drawing a line separating the target and not-target regions. The line need not be straight but is drawn to divide these two regions in the compromise manner to obtain the best values for $P(T_A \mid T)$ and $P(\bar{T}_A \mid \bar{T})$. The rule becomes: If v_1, v_2 is a point above the line, annotate not-target. If v_1, v_2 is a point below the line, annotate target. Analogous to the one attribute case $P(T_A \mid T)$ is the volume below the $p(v_1, v_2, \text{target})$ surface bounded by a surface normal to the attribute plane and passing through the descision line. Thus, changing the decision line changes the system score point. Notice that the inclusion of the second attribute can never make discrimination worse than it was with only one attribute alone because we can always use the decision line, $v_1 = B$, and achieve the same results as before. There is always the chance that the overlapping curves of target and not-target sets with one attribute will be more nearly separated in two attribute space. There is nothing to lose but the cost of trying.

If two attributes permit the possibility of improvement, then certainly three attribute measures could improve the system score further. Therefore, the use of multiple attributes for the purpose of improving remote sensing system performance is likely to be a profitable venture.

The procedure for N attributes is clear enough, however, some very important practical problems arise. The simple forms of $p(v_1, v_2)$ used in the illustration are not likely to be simple in an actual case. The surfaces defined by $p(v_1, v_2)$ may have many maxima and minima. The nature of the surfaces for the two categories, target and not-target could be very complicated. Hence, the optimum decision rule could be very complicated. The second problem is that for more than two attributes, the easy graphical approach to locate the target and not-target regions can no longer be used. Some kind of computer search is required to see where the clusters of points lie. Third, the more attributes one introduces, the more sample elements of each set must be used to establish statistical significance to the probability densities. One may rapidly run out of test sites and computer capacity.

3.4. THE ADAPTIVE CATEGORIZER

A popular and very interesting implementation of a two-attribute system is called the "Adaptive Categorizer."* It is a particular method of distinguishing between two sets of elements by the adjustment of thresholds.

Let X_1 and X_2 be the attribute values introduced into the categorizer from two different attribute measurement channels. Let C be some constant value forming a third input. If we amplify or attenuate each of these separately by some means by factors W_1, W_2, W_3, respectively, and add the resulting terms, we obtain the sum, S, where

$$W_1X_1 + W_2X_2 + W_3C = S.$$

This is the equation for a family straight lines in two attribute space. The sum S may be considered to be a parameter. Consider the line for which $S = 0$. Substitution of points on one side of the line provides a sum $S > 0$, points on the other side provide a sum $S < 0$. Now the object is to adjust W_1, W_2, and W_3 in such a way that the line for $S = 0$ divides the two sets of elements so that attributes of a target element give $S > 0$, attributes of not-target elements give $S < 0$. The adjustment is done by training sequences where the elements are identified for the machine and wrong responses require small adjustments of the weights W_1, W_2, and W_3 in such a direction that the wrong response tends to be corrected. For instance, if the sum were positive and it should have been negative and X_1 was positive for that element and X_2 was negative, then W_1 would be reduced slightly and W_2 would be increased slightly. W_3 is adjusted in the same way.

Such an implementation does not provide for the maximum accuracy which is possible unless the grouping of the elements happen to fall into regions which can be separated by a single straight line. What determines whether or not elements of a population can be sorted into sets depends upon the attributes which are used and the sets which are chosen. If the sets overlap as plotted in attribute space, then no elaborate data processing scheme of any kind will be able to discriminate perfectly with these attributes. On the other hand, if the sets can be separated into distinct regions, simply connected, or otherwise, then perfect discrimination is logically possible although a discrimination system may be costly and complex in its implemented form.

---

*See review article by Charles A. Rosen, "Pattern Classification By Adaptive Machines," Science, Vol. 156; 7 April 1967.

4.0. SPECTRAL DISCRIMINATION

A special form of multi-attribute remote sensing system is the multispectral system. The attributes to be used to annotate the set membership of various materials will be measures of the spectral reflectances and emissances which are properties of the materials. These measures of reflectances and emissances are not defining attributes but there is good reason to believe that in many useful situations these properties may be closely allied with the defining attributes and would be a versatile set of attributes to choose.

The remote sensor measures the spectral irradiance in the image plane of the sensor located remotely from the material. One must attempt to infer the spectral reflectance and emissance of the material from these data. Thus one must account for the atmospheric effects introduced between the material and the sensor, the aspect dependent properties of the reflectance and emissance (to take into account the relative location of the sensor) and the nature of the driving power that produces the radiance from the material e.g., the spectral irradiance at the surface of the material and the direction of flow of this power, and the temperature of the material causing the emission. In the case of fluorescence under violet and ultraviolet irradiance, the internal conversion of power to longer wavelengths should also be considered. Failure to account for these factors to some reasonable degree of accuracy introduces an apparent variability of attributes which has no useful significance in the annotation process and will tend to degrade the system performance.

Imaging sensors making use of many spectral bands will provide much additional information but it will not be useful unless improved methods of handling and displaying the information for final data processing by the human interpreter are developed. The three most common methods of displaying multispectral image data are demonstrably deficient; they are the adjacent presentation of the black and white image in each band, the use of color film and the use of superposed images of black and white films through colored filters.

The adjacent presentation of images from a number of different spectral bands is currently being practiced to some extent. It is clear that, in general, this places additional burdens on an already overloaded human interpreter. Furthermore, many kinds of joint use of such information are very difficult for a human being to effect.

The use of color film permits superimposing simultaneous images in different bands and makes the interpreter's task manageable. It, however, is restricted to no more than three simultaneous spectral channels and there are good reasons for believing that the optimum number is much greater than three. For instance, color film cannot simultaneously do full justice to the three dominant visible colors and also display the distinctive behavior of

botanical materials just outside the red end of the visible spectrum. The color film preserves the shape, tone, color, texture, shadow, etc. used by the photointerpreter who "computes" mentally with these multiple attributes in conjunction with extensive contextual data to produce the annotations. To be sure, the human interpreter is a remarkable multi-attribute data processing device. No one, including the interpreter, is quite sure which attributes and what data are being used from moment to moment to produce the resulting annotations.

However, the human data processor is relatively slow in operation, variable in performance, and distinctly limited in the type and format of data which can usefully serve as input data. It would not be difficult to show by experiment with color and black and white photographs that the color information may be the least useful part of the data which the interpreter can use for annotating set membership of an object or material. By presenting a small spot of some color picture to an interpreter devoid of shape, tone, texture, and context, he will nearly always be in error in annotation. Apparently, shape and context data provide the human data processor with the most valuable input. Because of the high information value of shape and context, a multispectral system which provides some degree of spectral discrimination in a form which does not lose shape and context could be processed a second time by the human multi-attribute interpreter so as to take advantage of both spatial and spectral attributes. In some cases, the multispectral attributes may provide enough discrimination to provide sufficient accuracy in system performance without the addition of the human interpreter. In such a case the speed of the system performance is limited only by machine speed limits.

The presentation of spectral attributes can take many forms and still preserve the spatial and contextual data suitable for human data processing. Some of the simple forms make use of available photographic processes mentioned before. More complex forms utilize computers and decision rules. The principle of operation is the same but the system performance can vary greatly.

4.1. SIMPLE LINEAR FILTER SYSTEMS

The most elementary linear filter system is a system where the contrast between the image of the elements of the target set and the image of the elements of the not-target set which are likely to be spatially adjacent is increased by using a passive spectrally selective transmission filter. The output is a black and white image to be viewed by the human interpreter. The increased contrast aids the interpreter in applying his data processing system using primarily spatial and contextual attributes.

Suppose that the spectral irradiance at the remote sensing device due to a target element is E_λ(target) and for a not-target element adjacent to the target element is E_λ(not-target). The transmission filter intercepts the irradiance at the aperture to produce a spectral

irradiance in the image plane proportional to $\tau(\lambda)E_\lambda$(target) and $\tau(\lambda)E_\lambda$(not-target). Assume for simplicity of argument that the detector (or film) responds with an output signal, v, which is proportional to the irradiance incident upon it so that for each element

$$v = K\int_0^\infty \tau(\lambda)E_\lambda \, d\lambda,$$

where K is the constant of proportionality.
Now, the contrast, C, may be defined as

$$C = \frac{v(\text{target}) - v(\text{not-target})}{v(\text{not-target})}$$

The object is to discover the spectral transmittance, $\tau(\lambda)$, which will make C as large as possible. If we have target and not-target elements which have spectral irradiances which are unique to their set membership, then it would be a simple matter to establish the best filter from these spectra. Let the spectra be as shown in figure 3-5. Take a narrow spectral band at λ, only so that $\tau(\lambda_1) = 1$, $\tau(\lambda) = 0$ for all other λ. Then

$$C = \frac{E_\lambda(\lambda_1, \text{target}) - E_\lambda(\lambda_1, \text{not-target})}{E_\lambda(\lambda_1, \text{not-target})}$$

or

$$C = \frac{E_\lambda(\lambda_1, \text{target})}{E_\lambda(\lambda_1, \text{not-target})} - 1$$

Suppose that at λ, the ratio, $E_\lambda(\lambda_1, \text{target})/E_\lambda(\lambda_1, \text{not-target})$ is maximum for all narrow band choices one can make. Then for narrow band choices only

$$1 + C_{max} = \frac{E_\lambda(\lambda_1, \text{target})}{E_\lambda(\lambda_1, \text{not-target})},$$

or

$$E_\lambda(\lambda_1, \text{target}) = (1 + C_{max})E_\lambda(\lambda_1, \text{not-target}).$$

Now consider another narrow band at λ_2 for which the target irradiance is noticeably greater than the not-target irradiance. Because we have assumed that the greatest narrow band contrast was found at λ_1, it follows that

$$\frac{E_\lambda(\lambda_2, \text{target})}{E_\lambda(\lambda_2, \text{not-target})} < (1 + C_{max}),$$

or

$E_\lambda(\lambda_2, \text{target}) < (1 + C_{max})E_\lambda(\lambda_2, \text{not-target})$.

Adding the two contributions from λ_1 and λ_2 together, it follows that $E_\lambda(\lambda_1, \text{target}) + E_\lambda(\lambda_2, \text{target}) < (1 + C_{max})\,[E_\lambda(\lambda_1, \text{not-target}) + E_\lambda(\lambda_2, \text{not-target})]$, so that the contrast using the sum of power in these two bands is less than maximum. Therefore, the filter for achieving maximum contrast would be a very narrow spectral band where the ratio of the spectral irradiances is maximum. No other part of the spectrum should be allowed to contribute since that would reduce the contrast below maximum.

There are a lot of important facts which have been neglected in this simple derivation.

1. A very narrow spectral band filter would require the device to respond to very small amounts of power. The inherent instrument noise of a radiometer or the fog level and granularity of the photographic film place lower limits to this power. The band pass of the filter must be quite broad in reality.
2. The spectral reflectances of the elements of a set are rarely the same. Most frequently the elements of a set are composed of subsets of elements where the spectral reflectances may be similar within subsets but can differ greatly from subset to subset. Such is almost always true of the not-target set which is to contain everything else other than target elements.
3. The spectral irradiances at the instrument aperture are functions of spectral irradiances of sources, aspect of illumination and reflection, the effect of the mixture of other reflectances within a resolution patch containing the object and the atmospheric transmittance and radiance due to scattering and emission. All of these must be accounted for to arrive at the reflectance or emissance property which is the attribute of the elements of the set. Even if the target reflectances were known, the precise value of the spectral irradiance may remain uncertain due to unaccountable variations in these other factors.

At best, some improved contrast using a single passive filter is useful for improving the data input to the human data processor where the overwhelming majority of the data processing is done using non-spectral attributes.

4.1.1. CONTRAST RATIO STUDY. In spite of the fact that field conditions are highly variable because of the effects of dirt, and surface conditions, it is reasonable to expect that certain spectral regions may exist in which there is a high probability that the objects of interest can be sorted out from among others.

A first step toward a systematic approach was taken in 1962 by a man named Langley, when he took two sets of spectral reflectance curves (for disturbed and undisturbed alfalfa)

and computed the normalized reflectance differences (NRD) of paired observations between the two sets, disturbed alfalfa the target set and undisturbed alfalfa the not-target set.

The normalized reflectance difference was defined as

$$\text{NRD} = \frac{\rho(\lambda, \text{(target)} - \rho(\lambda, \text{not-target})}{\rho(\lambda, \text{not-target})}.$$

One of the results of the procedure followed by Langley is illustrated in figure 4-1. This graph shows the normalized reflectance differences and the 95% confidence limits for disturbed and undisturbed alfalfa. The cross-hatched areas indicate those portions of the spectrum that would normally produce a tonal contrast greater than +50% at the 95% confidence level. This information was compared with film-sensitivity and filter-transmission characteristics to determine film-filter combinations which would exhibit a tonal contrast between the two desired features.

Using the results of the computer-aided analysis, film-filter combinations were selected and photographs were made. Microdensitometer analysis of the photos demonstrated a 96% correlation between the predicted and achieved results. A few inconsistencies were observed, but the overall results were encouraging.

A similar effort, the Contrast Ratio Program (CRP), was developed by The University of Michigan using the IBM 7090 digital computer in the analysis of reflectance data collected on various objects (e.g., soils, grasses, trees, etc.) at points throughout the wavelength spectrum. Each object was classified as a target or a background (not-target). Several independent measurements of each object's spectral reflectance were utilized. The ratio of the target reflectance to the background reflectance was calculated at specified wavelength intervals for all combinations of the specified targets and backgrounds, and their logarithms were computed. Then, the frequency distributions, means, and standard deviations were computed for all logarithms in each interval.

A brief justification for the use of logarithms for this problem is in order. For a typical photographic film (which is used for data analysis in many remote sensing applications), the measurements are made in terms of photographic density, which is given over the straight line region of the characteristic curve by:

$D = \gamma \log (q/q_0)$ where q_0 is the straight line intercept on the log q axis.

The contrast C' between two scene points was defined as the difference in the densities, or:

$C' = D_2 - D_1$ so that

$C' = \gamma (\log q_2/q_0 - \log q_1/q_0)$ or

$C' = \gamma \log (q_2/q_1)$.

In a given wavelength interval, the exposure, q is proportional to the reflectance ρ for wavelengths where emission is negligible, so the contrast may be written as:

$$C' = \gamma \log (\rho_2/\rho_1)$$

A typical result of the CRP is illustrated in figure 4-2, where 13 kinds of green painted surfaces (T) were compared to 13 species of Northern Michigan green tree leaves ($\overline{T}$). The mean contrast is denoted by the solid curve, whereas the standard deviations are represented by the shaded areas.

It will be noted that the standard deviations appear to be large in all cases. Much of the variation is due to wide variations among the elements within a given set. The intra-set variation is illustrated in figure 4-3, where green leaves were contrasted with green leaves. Obviously, the mean contrast is zero throughout, while the standard deviations indicate the widespread reflectance variations in the data.

Many of the reflectance variations are real, but some are the result of a combination of uncertainties and inconsistencies in measurement techniques. Some improvement could be obtained by limiting the data to sets of data obtained by the same researcher under constant conditions, but as a rule there is not enough data from any given researcher.

As indicated in figure 4-2, the wavelengths for maximum contrast for the case at hand are at 530 nm and from about 700 nm into the near infrared. Using the frequency distribution tables along with the figures, it is seen that about 95% confidence could be obtained by using 540 nm to 560 nm and 760 nm to 800 nm. One can plot the probability densities p(C', target) and p(C', not-target) from the frequency distributions if C' is determined from the photographic density difference in two grid squares where at least one contains a tree with green leaves.

Thus at 1.140 μm from figure 4-2 and figure 4-3 the approximate probability densities are derived and shown in figure 4-4. The relatively few samples may not be enough but they indicate that a decision rule at C' = -0.2 would provide a perfect system score. Notice that in the range 0.540 to 0.560 μm the contrast is high as shown in figure 4-2 but the intra-element variability of the not-target set is also high. The plot similar to figure 4-4 of the two probability densities for this visible wavelength region would show significant overlap so that a less than perfect system score would result. The added variability due to environmental factors which would appear in practice would be significant in both cases. The assumption has been made that exposure q is proportional to ρ and independent of all other factors. This assumption will not be valid in general.

4.1.2. TRICOLOR SYSTEMS. Mention has already been made of ordinary Ektachrome color film, infrared Ektachrome, and other possible color translation schemes. It is

instructive to examine these systems again in the formal manner of a discrimination system. Each resolved element of these tricolor systems represents three spectral attribute measures, v_1, v_2, and v_3 where

$$v_1 = K_1 \int_0^\infty \tau_1(\lambda) E_\lambda(\text{aperture}) R_1(\lambda)\, d\lambda,$$

$$v_2 = K_2 \int_0^\infty \tau_2(\lambda) E_\lambda(\text{aperture}) R_2(\lambda)\, d\lambda,$$

$$v_3 = K_3 \int_0^\infty \tau_3(\lambda) E_\lambda(\text{aperture}) R_3(\lambda)\, d\lambda.$$

The constants, K_1, K_2, and K_3 are presumably known sensing device properties. The functions $R_1(\lambda)$, $R_2(\lambda)$, and $R_3(\lambda)$ are the known responsivities of the detection components. The functions, $\tau_1(\lambda)$, $\tau_2(\lambda)$, and $\tau_3(\lambda)$ are the transmittances of the filters used to separate the spectral power into the three portions to be measured. In normal Ektachrome film, the three quantities, V_1, V_2, and V_3 are proportional to the blue, green, and red exitances of the film or print. Instead of viewing the film by eye, let us plot the three V values in three dimensional spectral attribute space as shown in figure 4-5. It is apparent that each triple of values belonging to the same resolution element corresponds to a vector, $\vec{V}$, in three dimensional attribute space. Every resolved element in the color image has a corresponding point (or vector) to go with it. If we consider that the three quantities represent the blue, green, and red exitance, then the diagonal line from the origin and making equal angles with the axes is the locus of all grey values from black at the origin to white at some maximum distance from the origin along that line. Thus, in figure 4-5, the point $\vec{V}_A$ would represent some grey level while the point $\vec{V}_B$ would represent the color dominated by the V_1 exitance. If V represents film density, the locus of constant density would be the surface of a quarter sphere with center at the origin. The locus of points having the same color but different density would be a straight line or ray beginning at the origin. Now, it is possible to plot the tri-spectral attributes of the resolved grid covering a region containing some selected target and not-target set. The clouds of plotted points of one set may likely overlap the cloud of the other set to some degree. The function of camouflage is to make a target class cloud overlap the not-target cloud. Many examples of this effect can be found in wildlife. Against predators equipped with the same remote sensing system, the ability to present attribute values, $\vec{V}(T)$, wholly within the cloud of not-targets, $\vec{V}(\overline{T})$, has obvious survival value in spite of the fact that the predator is experienced from birth with his particular attribute system.

A densitometer system using appropriate filters can provide the coordinates of the three attribute points as fast as it can be scanned across a color image. A human observer requires the use of an elaborate set of comparative color chips to achieve modest accuracy of the location of the point. Quantitative use of spectral attributes is not the strong part of the human sensing system but the radiometric system is outstanding in both accuracy and speed in quantitative spectral measures.

Now it is easy to see graphically the function of the color translation systems. If the filters and responsivities were changed so that instead of V_1, V_2, and V_3 corresponding to red, green, and blue visual responses, they corresponded to infrared, red, and green equivalents then resolved elements of healthy foliage would plot along rays close to the V_1 axis while unhealthy vegetation would plot close to the V_2, V_3 plane near the V_3 axis. See figure 4-6. One can visualize a decision plane separating these two clouds. If the separation can be done by human eye, when V_1, V_2, and V_3 are displayed as visual values, then nothing more is required than infrared Ektachrome film.

The Itek and Fairchild multispectral cameras offer a generalization of the three spectral attribute system. In these, the three spectral attribute values are recorded as three separate images and are recomposed for human data processing by superposition through color filters. Such a system is equivalent to a mapping of tri-attribute points in one three dimensional space into the visual three-dimensional color space. Each spectral attribute point $\vec{U}$ in the first is represented by a corresponding point $\vec{V}$ in the visual space. This mapping can be expressed as,

$$V_1 = A_{11}U_1 + A_{12}U_2 + A_{13}U_3$$

$$V_2 = A_{21}U_1 + A_{22}U_2 + A_{23}U_3$$

$$V_3 = A_{31}U_1 + A_{32}U_2 + A_{33}U_3$$

where the values of the constants, A_{ij}, are under the control of the operator of the projector. However, only positive values for A_{ij} can be used in this system because the transmittance values for the filter in front of image number one, for example, are proportional to A_{11}, A_{21}, and A_{31}. Negative transmittances and minus densities are not defined.

The great advantage of such a tri-attribute system is that the choice of spectral attributes U_1, U_2, and U_3 can be chosen to separate the attribute point clouds of the T and $\overline{T}$ sets regardless of the visual properties and then replotted by relatively inexpensive optical means in such a way that the more qualitative eye of the human data processor can recognize where the decision plane should be in the visual color space without loss of shape and context data.

There are some serious disadvantages to this system which are commonly overlooked. If the target and not-target sets are each spectrally homogeneous and only a two category

system is considered, then this system may work well. However, if one of the sets is spectrally inhomogeneous so as to consist of a number of subsets which are characterized by distinctly different mean values of U_1, U_2, and U_3, then the best processing filters, A_{ij}, for one subset is not likely to be satisfactory for another of the subset so that the complete set membership can not be found in a single process. For example, the set of all rose blossoms constitute a spectrally inhomogeneous set. Each rose blossom subset may plot in widely different regions of the spectral attribute space and therefore map into widely separated parts of the visual color space in such a way as to cause the decision plane between some subsets and the not-rose blossom set to be indistinct to the human processor. Thus, the small amount of useful inferential connection between visual color and set membership which the human already had is sacrificed for little advantage. Indeed, it is quite possible to find a net loss in system performance because of the spectral attribute mapping.

Examples of several mappings can be found in the Manual for Color Aerial Photography using the Fairchild four lens camera (p. 461). In the case of the Fairchild camera, four-dimensional spectral attribute space is mapped into a three-dimensional visual color space so that the mathematical relation is

$$V_1 = A_{11}U_1 + A_{12}U_2 + A_{13}U_3 + A_{14}U_4$$

$$V_2 = A_{21}U_1 + A_{22}U_2 + A_{23}U_3 + A_{24}U_4$$

$$V_3 = A_{31}U_1 + A_{32}U_2 + A_{33}U_3 + A_{34}U_4$$

Such a mapping is called a many to one mapping where a certain set of points in U space corresponds to a single point in V space. It is easy to see what that set of points is in U space for any given point in V space. First consider mapping a three attribute space into a two attribute space so that

$$V_1 = A_{11}U_1 + A_{12}U_2 + A_{13}U_3$$

$$V_2 = A_{21}U_1 + A_{22}U_2 + A_{23}U_3$$

For any particular point (V_1, V_2) specified, the equations above are equations of planes in three dimensional U space. The value of V_1 is proportional to the distance of the first plane from the origin and V_2 is proportional to the distance of the second plane from the origin. The direction cosines of the normals to the planes are proportional to the three filter values A_{11}, A_{12}, A_{13} and A_{21}, A_{22}, A_{23} which must be positive numbers. Therefore, the only planes which can be considered must be planes which will fit into the corner or along the walls and floor of the three-dimensional space as shown in figure 4-7. They cannot tilt in any other way. Therefore, the line of intersection can never be a ray through the

origin extending outward in the "constant color" direction. All points along the line of intersection correspond to the single point (V_1, V_2) in two dimensional attribute space. Thus, points A and B which represent spectrally different attributes will be indistinguishable in the two dimensional space. Now, the most likely intraset spectral variability will be due to unaccounted variations in the magnitude of the illuminating irradiance of the element in situ. Hence, the most likely direction of the cloud of attribute points is along a ray. The most unlikely direction is along the intersection line of the two planes. Therefore, what is most likely to happen in this processing method is to make an element with attribute A and an element with different attribute B become indistinguishable when they should remain distinguishable. Having removed the small but non-zero value of normal color inference it is possible to reduce system performance by unintentional confusion with confusing processed colors.

This argument can be extended directly to four attributes mapped into three visual attributes. The three relations for V_1, V_2, and V_3 represent three planes in a four-dimensional space for each choice of point $\vec{V}$. The intersection of these three planes forms a line in four-dimensional space which cannot be in a direction of a ray through the origin. Unless the intraset variability happens to extend along this line, some not-targets may likely be made indistinguishable from a target. Certainly, the expected trend of a spectrally homogeneous set is in the direction of a ray through the origin. The method used by the Fairchild and Itek cameras happens to be available and relatively economical. But the choice of decision boundaries which can be implemented are so restricted that they may not apply to the more general intraset variability that is likely to occur. One can see that the form of the relationships between $\vec{V}$ and $\vec{U}$ points are linear equations involving the filter factors as coefficients. Hence, these processing systems are all classed as linear filter systems.

4.2. THE MULTISPECTRAL SCANNING SYSTEM

A more general implementation of multispectral attribute discrimination is the multispectral scanning system. Although the irradiance at the aperture of a multispectral camera can be measured by measuring film density, it was pointed out in the photographic section that accurate quantitative radiometry was not the strong point of the photographic process. A calibrated radiometer using solid state detectors can generally out perform the photographic system in accuracy and precision. It is, after all, the inherent spectral attribute of the remotely located element which is needed and additional variability due to instrument error can only make this attribute more obscure.

In addition, non-linear processing may be required in order to locate and use a decision boundary by computer techniques so that the attribute measures must be supplied to the

computer in electrical signal form. If the data is in film format, it must be scanned by a densitometer to put the data into electrical signal form. An optical mechanical scanner does the scanning in the first place with radiometric accuracy.

The use of photographic film as the responsive element in the remote sensing device precludes the measurement of irradiance at wavelengths other than the narrow spectral range near the visible. An optical mechanical scanner can make radiometric measurements of radiation at any wavelength as long as it can be brought to focus by reflective optics.

In spite of the fact that optical mechanical scanners are more complex, less available, and more expensive, they appear to be ideally suited to multispectral discrimination applications. The processing of data where each element presents twelve or more spectral attributes offers its own problems. Extending the reasoning as presented above, one must now consider clouds of points in twelve-dimensional space. It is extremely difficult to visualize locations in such a space so that one is reduced to the formalisms of mathematics and computer technology.

The objectives are exactly the same as ever. Given the attribute points of the target and not-target what are the probability densities, $p(U_1, U_2, U_3, \ldots, U_{12})$, for each set and hence what will be the decision rules or where will the decision boundaries be?

4.2.1. EVIDENCE OF INTRASET VARIABILITY. One of the important causes of intraset variability is the variation of spectral attributes with aspect angle of view.

Figures 4-8 through 4-13 show the variation of signal output as a function of scan angle from The University of Michigan twelve channel system. The scan covers essentially the same material on the ground but the signal output relative to its value at the nadir changes significantly and in different ways in the different spectral channels. The variation at large angles is undoubtedly instrument variation but the majority of the variation between $\pm 40^{\circ}$ is characteristic of the material being viewed as well as the angle of illumination. Figure 4-14 shows the variation due to direction of illumination. Figures 4-15 through 4-17 show a similar effect over different material for twelve channels of spectral data. All channels represent apparent radiance due to the reflected solar irradiance in the twelve narrow spectral bands shown in figure 4-18. If the intraset variability were not taken into account in some fashion, the cloud of points in attribute would be widely spread. A lower system performance can be expected. However, the variability with aspect could be an inherent characteristic of the material itself and could be the source of a valuable attribute for establishing set membership.

4.2.2. SIMPLE DECISION RULES. The problem of finding the best decision rules lies with the mathematician and the computer. It is instructive to discuss some simple procedures which come to mind. Suppose that one of the sets (say the target set) was spectrally homogeneous so that the spectral attribute point cloud for that set were clustered about a central attribute point. One could consider a simple decision rule such as:

$$C^2 \geqq (U_1 - \bar{U}_1)^2 + (U_2 - \bar{U}_2)^2 + \ldots + (U_{12} - \bar{U}_{12})^2$$

annotate T otherwise annotate $\bar{T}$

Here C is the decision rule parameter,

$\bar{U}_i$ are the mean values of the respective spectral attributes,

U_i are the values for the specific element being classified.

This decision rule is equivalent to a decision boundary surface in the form of a spheroid in twelve-dimensional space centered on the mean value of the cloud.

As the value of C is changed the radius of the spheroid is changed. The system performance will change accordingly. As the size of the spheroid grows, more target points will be included within the target decision region but also more not-target points may also be included. Notice that this decision rule requires non-linear processing of the spectral attributes and cannot be accomplished by any simple filter system. One could make a similar decision rule slightly more general to take into account the expected distribution of cloud points in the ray direction from the origin due to unaccounted changes in illumination. Consider the decision rule:

$$C^2 \geqq \sum_{i,j=1}^{12} (U_i - \bar{U}_i)(U_j - \bar{U}_j)b_{ij}$$

where b_{ij} are constants which make the general quadratic form represent an ellipsoid in twelve-dimensional space with the longest axis along a ray direction from the origin. The determination of the constants, b_{ij}, can be established through formal mathematical methods which relate b_{ij} to the observed variances of the spectral attributes so that the ellipsoid will best enclose the points.

There are a number of other ways of searching the twelve-dimensional space to locate the target and not-target regions which are more complex but more general and should result in superior decision rules and make successful multiple category discrimination systems possible.

Once the statistical properties of the spectral attributes of resolved grid areas are related to the respective set memberships by tests over test sites, the probable set

membership of spectral attributes over unknown territory may be obtained. Since the process utilizes only spectral attributes, the data should be further processed by a human processor. A display in the form of an image of the area can be prepared from the output of the spectral attribute processor whereby the probable set membership of each resolved grid area is color coded in any coding system which will be most suitable for the human to use in familiar map form.

5.0. MULTI-ATTRIBUTE RADAR SYSTEMS

What can be done with spectral attributes will also apply to any remotely sensed attribute. The principle is unchanged, only the nature of the attribute and the corresponding instrumentation required to measure it are changed.

The radar receiver measures the irradiance at the aperture of the radar antenna which is received from a remotely located grid area. The grid size is determined by the properties of the radar. The attribute of the element to be used to determine set membership is the reflecting properties of the element for radar power.

5.1. POLARIZATION DIVERSITY SYSTEM

In a radar system, the source of illumination is under the control of the remote sensing system user. He can cause a particular type of polarized radiation to be transmitted. The reflection of that power can result in both a change in magnitude and a change in polarization in the received power which is caused by the electromagnetic and geometrical character of the grid area being viewed in combination with the known properties and location of the sensor.

Therefore, this change in polarization can ultimately be attributed to inherent characteristics of the grid area and be a possible measure of set membership. Two attributes can be obtained for each grid element.

$$U_1 = \frac{P(\text{horizontal, received})}{P(\text{horizontal, transmitted})}, \text{ and}$$

$$U_2 = \frac{P(\text{vertical, received})}{P(\text{horizontal, transmitted})}$$

where the notation, horizontal and vertical, refers to the direction of polarization.

Calculations can be applied to U_1 and U_2 to remove the variability of known factors so that the two derived quantities σ_1 and σ_2, the apparent radar cross sections, can be used to infer set membership. Two customary displays showing shape and context are used—the adjacent strip map presentation of these two quantities σ_1 and σ_2 and the more recently the linear filter presentation using superposition of the two maps in a similar manner as is done for the Fairchild camera images.

5.2. ASPECT DIVERSITY SYSTEMS

Both theory and experiment confirm that the radar cross section of a grid area is likely to be highly aspect dependent. Therefore, one can expect that a large part of the intraset variability will be due to aspect variations. However, aspect angle can be known and taken

into account to some degree so that what might be a very disturbing variability may turn out to be a new measurable attribute which characterizes the set membership of the grid element. The scatterometry system used at the University of Kansas is one form of an aspect diversity system. In the Kansas system a line of grid areas along the flight path are illuminated at one time. The limits of the grid area along the flight line is set by the doppler shift of the reflected power—i.e., the grid below the aircraft produces zero doppler shift and the areas far ahead or behind show large doppler shift due to the larger component of radial velocity for these grids. By recording the doppler signals and knowing the ground speed of the aircraft, the radar cross section can be measured for every grid at nearly every aspect angle. A multi-attribute point is then derivable from the data. For each grid area a set of attribute measures,

$$\sigma(\theta_1),\ \sigma(\theta_2),\ \sigma(\theta_3),\ \ldots,\ \sigma(\theta_n),$$

are established. The processing can then be done as previously discussed. In the Kansas system, a downward looking strip map photographic system is also operated so that the combined shape, tone, texture, and context information can be supplied along with the radar data to permit the human processor to utilize his most valuable type of input data along with the radar data.

One can readily see from the foregoing discussion that a large number of multi-attribute systems are possible. The potentialities of each are not easy to visualize and sometimes have to be tried out just to see what utility might be discovered.

6.0. REMOTE SENSING FACILITIES

In general the form and contents of a remote sensing research facility will be defined by the set of missions or problems to which remote sensing is to be applied. Nevertheless, much can be said about a remote sensing facility independent of specialization.

6.1. TEST SITES

Any facility must include appropriate test sites which are to be the microcosm of unknown territory. It is easy to see from the preceding discussion on system performance that much depends upon the astute choice of these test sites. Clearly, much work will be done on the ground in the test site and therefore it must be accessible without undue penalty in time and funds spent. Frequently, test sites are chosen from existing natural science experiment stations because ground truth is already available to the experimenters and funds have already been committed from other budgets for that purpose. However, such sites are not likely to be a convincing microcosm. Recall that the discrimination process is one of sorting elements of at least two sets. One set may be the item of interest but the other set contains everything else that is likely to occur in unknown territory. For example, an agricultural test station is not likely to have poison ivy growing there because it interfers with the health and welfare of the station users while unknown territory may have a considerable amount of poison ivy. Should the remotely sensed attributes of poison ivy resemble the attributes of the items of interest, a reduced system performance can unexpectedly and unaccountably occur over unknown territory. With all factors considered, however, the economics of the matter will undoubtedly prevail and such test stations will be used, hopefully, with the necessary reservations as to the resemblance of the microcosm to the unknown territory.

6.2. PERSONNEL

The highest priority item in establishing a successful remote sensing facility is the selection of personnel. Remote sensing can emphasize sensor research where new sensor devices and techniques are developed or it can emphasize a mission oriented application of state-of-the-art instruments and known techniques to new problems. Either way, it is a multiple disciplinary field involving both physical and natural science. The type of personnel that will be needed will be governed by the emphasis of the research—the physical scientist for the sensor research, the natural scientists for the mission oriented research. In both cases a secondary interest and knowledgeability is required in the counterpart field.

In addition, to the basic discipline requirement is the requirement to innovate. Innovation is usually the result of the capacity of a person to modify his mental set or to view a problem

from "new angles." Much has been written and discussed on the subject of innovation and the "irradiance" on the subject is not much greater for it. Therefore, it would not be out of place to propose a line of consideration on this subject as it bears upon remote sensing. It might be useful. Even a small increment makes a large relative gain when beginning with so little.

The inexperienced student leaving school with his degree has little expertise. When he views a problem, all potential solutions appear equally difficult so that each possible line of reasoning may be taken on an equally likely basis. Because of the more random approach, valuable innovations could be chanced upon. Unfortunately, there are also a large number of useless possibilities which are just as likely to be taken. On the other hand, the expert is an individual who has discovered by experience and learned by extensive practice a few successful lines of reasoning with which he analyzes each problem. He has developed a mental set with regard to these lines of reasoning which worked for previous problems. New lines of reasoning are not likely to be spontaneous because of the natural tendency to use and think about what one knows best. Unless the new problem has an economical solution which happens to fall within the scope of the experts mental set, a new, more economical or useful solution is not likely to be found.

Instead of the strategy of using either all experts or all inexperienced graduates, the strategy of using an appropriate mixture makes good sense. If one were to bring together an assortment of experts whose mental sets were distinctly different, then the chance that an economical problem solution would fall within the scope of at least one expert's mental set is improved in proportion to the number of experts. However, it would be a mistake to believe that hybrid solutions are very likely through intellectual interchange between experts. It is not uncommon to find experts of different mental set viewing each other as misguided zealots pursuing a useless course while expending vital funds which could obviously be put to better use — i.e., their own. If one adds to this view the fact that misguidedness is a common state of the human mind to which experts are not immune, this view becomes credible. It might be true.

A possible means of developing hybrid solutions which may be more successful is to add the inexperienced graduate to this mixture. The inexperienced graduate can accept any one of the experts problem analyses as readily as any other. Moreover, any hybrid possibilities that come to mind are equally acceptable. It may be that approaches to problem solutions formed from a hybridization of approaches which were successful in related problems would provide a higher rate of successful innovation in new problem solutions. Certainly, communication from the experienced to the inexperienced comes naturally. Problems involving self esteem are avoided.

It is the problem of the director of the facility to maintain some modest amount of peace and tranquility within such a mixture and at the same time act as the one individual upon whom all failures can be blamed.

6.3. PHOTOGRAPHIC FACILITIES

A facility for remote sensing should have a photographic facility. The almost universal use of the photographic process in remote sensing either as a recording system for ground truth in the test sites, a direct remote sensing device from an elevated platform, or as a display medium for optical mechanical and radar sensors assures the employment of a photographic facility somewhere. The questions which remain are how much of a facility should be retained locally and how much of the facility should be contracted out to private industries who specialize in providing photographic services.

Much of the answer depends upon the expected amount of usage and the variability of the demands upon the facility. In the case of large scale but sporadic operations using the photographic camera as the prime sensor for mapping purposes in traditional fashion, the use of existing aerial survey companies makes excellent sense. An installation which could handle such a task would be a major investment in equipment and require a standing group of trained technical personnel. If the demands upon the photographic facility are highly variable so that mass production techniques cannot be employed, then one can expect that the motivation of private industry to serve as such a facility will be considerably less and the price for such a service will be very substantially higher than for routine processing. Frequently, it is argued that it is more economical under such circumstances to maintain the photographic facility within the remote sensing facility. Largely, such reasoning rests upon the artifacts of bookkeeping where the extra costs of such a facility are derived from some other budget as a non-photographic cost and are hidden from view. The principal argument rests upon whether or not the needed symbiotic relationship can be established between photographic and remote sensing facilities in order to accomplish the mission. The cost will probably be comparable anyway.

6.4. EQUIPMENT

Equipment in a facility is generally determined by the specific task to be accomplished. In a remote sensing laboratory, there are some basic instruments and components needed in making measurements. A list of typical equipments used in multispectral scanning is given in tables 6-I, 6-II and 6-III. The equipment list is incomplete in that there are no specialized instruments. For example, a geologist may need instruments for measuring hardness, surface roughness, heat conductivity, or chemical analysis. A botanist may wish to measure leaf turgidity and moisture content or to dissect a leaf. None of these types of facilities are listed.

The heart of a facility is its personnel, and any program must be designed around them. The selection of equipment must be made to obtain maximum use for a given expenditure. Thus emphasis should be placed on versatility, low cost, and simplicity. Complex and expensive equipment should be rented or leased when possible. In general, laboratory measurements are simpler and less costly than field operations which in turn are considerably simpler and less costly than airborne operations.

The costs given in the following tables are estimates and are given for general planning purposes only. The symbol K stands for the factor 10^3.

TABLE 6-I: LABORATORY FACILITIES

Spectrometer Laboratory

Optical Alignment and Checkout Laboratory

Imagery Analysis and Interpretation Laboratory

Photographic Laboratory

Tape Playback Room

1	Short wavelength dual beam spectrometer with reflectance attachment (0.3-2.5 μ)	$10.0K
1	Radiometer - $5K and/or scanning radiometer* $30K	
1	Blackbody source	1.2K
1	Calibration lamp	2.0K
1	4" collimator	1.0K
1	Double monochromator	3.0K
1	Long wavelength spectrometer with reflectance attachment (0.3-15 μ)	60.0K
1	Microscope	.3K
1	Optical bench	1.5K
1	Set interference filters	.8K
1	4 × 5" camera	.3K
1	35 mm camera	.1K
1	Polaroid scope camera	.5K
1	Photo dark room	5.0K
1	Light table	.3K
1	Densitometer	1.0K
1	Tape playback and image recording facility	20.0K
1	Stereo-plotter (ranges from $10K to $200K)	
1	Microdensitometer	10.0K

Components and Spares

1	InSb detector	.5K
1	Hg:Ge detector	3.0K
1	Set of dewars for N_2 & He_2	1.0K
1	Vacuum pump	.5K
1	Miscellaneous lenses & mirrors (front surface)	1.0K
1	Volt-ohm meter	.1K
1	Miscellaneous electronics (including detector test equipment)	10.0K

*Can be used in field

TABLE 6-II: FIELD EQUIPMENT

1	Radiometer $5K and/or scanning radiometer $30K*	
1	Multipoint recorder and thermocouple	$2.5K
1	Micrometeorological station	2.0K
1	Field spectrometer (0.3-2.5 μ)	35.0K
1	Field spectrometer (2.5-15 μ)	50.0K
1	Storage oscilloscope	3.0K
1	Digital punched tape recorder	3.0K
1	Instrument van	5.0K

*Can be laboratory equipment

TABLE 6-III: AIRBORNE EQUIPMENT

1	Aircraft (Cessna 185 or larger)	
1	9" × 9" mapping camera	$40.0K
4	70 mm cameras (pan, IR, color & IR color)	12.0K
1	Dual channel scanner with calibration	75.0K
	replaceable filters	
	replaceable detectors	
1	Tape recorder	20.0K
1	Drift meter	
2	Camera mounts	5.0K

Ancillary Equipment for Aircraft*

1	Ground power unit	
1	CRT	2.0K
1	Intercom in aircraft	2.0K
1	Special air to ground communication	3.0K
1	Set of lamps to layout flight path	1.6K
1	Set of calibration panels	1.0K

*Field and laboratory equipment are needed for checking optical alignments, sensitivity, spectral response.

7.0. DATA STORAGE AND RETRIEVAL

The problems of information storage and retrieval are important not only to remote sensing facilities but to the society in general. Libraries and specialized information centers face a common problem. It may be surprising to some to discover that the problems of information and data retrieval are isomorphic to the problem of remote sensing for items of interest in the environment. Although remoteness, natural randomness, and special sensors using non-visual radiations are not necessarily involved, the search and discrimination problem contains all of the same concepts. Generally, the problem of data storage and retrieval is a multi-category problem of major difficulty, however, the defining attributes of the elements being sought are many times available in the form of written titles. The principal problem with data stored in the form of images is that written titles do not cover all of the possible sub-categories which may be chosen by the data bank interrogator.

Various strategies in operating a data bank of remote sensing image and ancillary data can be examined.

7.1. FOLLOW-THE-LEADER STRATEGY

If a systematic categorization system for remote sensing data has already been established—for example, the storage and retrieval methods used by the Federal Government Agencies, then good arguments can be made to simply follow the same categorization systems. The purpose of a data bank is to provide a service to those who will use it. If a widely used system has been learned by a large number of potential users, then the users have adapted their thinking in such a way as to recognize in which data bank categories they can expect to find data fitting their own specialized category of interest. If a wholly new set of categories are presented to them, they will have to learn a new relationship between data bank categories and their same specialized category. It becomes much more difficult to use separate data banks. Transference of data from one data bank to another also requires extra labor. The dominance of the Dewey Decimal and Library of Congress library systems make the use of libraries anywhere a simple and easy process. The categorization systems become common knowledge to users. The strategy of "follow-the-leader" could provide the same common knowledge for remote sensing data.

7.2. THE DEDUCTIVE CATEGORY STRATEGY

In many instances, the systems which have already been established and used may be based upon photographic mapping data and might not be expandable to encompass data from new sensors including the non-imaging sensors.

The deductive category system utilizes those categories which most nearly offer a logical start in locating data within some specialized category of interest.

One of the most useful categorizations is clearly by geographic location. One presumes that an individual with an interest in some specialized category will surely know enough about that category to specify likely geographic locations for it. An individual looking for sensor information on oak trees can hardly be serious in purpose if he has no reasonable idea what regions of the earth are likely to have oak trees. Hence, deductive reasoning using contextual knowledge can eliminate much of the data bank.

The use of the time and date of the data acquisition also provides a useful deductive category. Again, using the example of oak trees, a data bank interrogator should know the season of the year for which he would want oak tree data—or does he want trees with foliage or without foliage. Specification of time and date also allows one to make use of ancillary data provided by weather stations in the same geographical location and relate the sensed data to historical events such as floods and storms. The object is to rely upon the deductive power of the interrogator.

One now comes to consider much more specialized data such as sensor type, scale, spectral band, individual or organization taking the data, the quality and resolution. Following the deductive category strategy, such data might better be presented to the interrogator in image form. The exact scale, quality, resolution, and spectral band are not as important for retrieval purposes as they will be after retrieval. If the data is in the form of an image, then some reasonable microfilm or microfiche facsimile of that image would provide much more general information to the searcher than any costly and specialized annotation system could provide. It permits the exercise of the multi-attribute processing capacity of the interrogator who seeks items of a category which was never considered before.

7.3. RETRIEVAL SYSTEM SCORE

Just as remote sensing system performance can be measured, so also the retrieval system performance can be measured. The equivalent of the test site is the properly selected subset of data to form a statistical microcosm of the entire data bank (the unknown territory). The microcosm can be searched in its entirety. A test interrogator makes his request of the microcosm. The retrieval system is made to function in the same way it would function in the unknown territory. The number of called out relevant items (targets annotations) are presented and the rest of the microcosm items are then the called out irrelevant items (not-targets). The interrogator can then examine the entire contents of the microcosm and present "ground truth." The system score is derived and the expected performance for the entire holding can be estimated. Changes in retrieval methods can be evaluated this way.

There is a difference in the consequences of such tests for remote sensing and for retrieval. The tests of a remote sensing system are expected to lead to operational use of the system to search unknown territory many times for the same target set. A retrieval system is called upon to search for a variety of target sets none of which may ever be searched for again. Clearly the score point depends greatly upon the choice of the target set as much as it does upon the retrieval system.

7.4. COMPUTER VERSUS MANUAL RETRIEVAL

Much of the experience at The University of Michigan with both computer and manual retrieval systems is in document retrieval for information centers. Less experience is available at present with retrieval of images, reflectance and emissance data on a large scale.

In the case of document retrieval it was found that for document holdings below about 50,000, the manual retrieval system using alert and motivated clerks was clearly superior to computer search for most search tasks. Computers are practically manditory in the routine inventory or routine large scale search task required for annotated bibliographies covering a general area of interest. The reflectance and emissance data have been put into digital computer tapes so that retrieval by computer is quite appropriate. However, the reason for putting the data on tape is not principally for retrieval purposes but for possible computational purposes for reflectance and emissance modeling. The philosophy is simple. If you are going to store these spectral reflectance and emissance data, why not store them in the form in which they are most likely to be used.

In the case of multispectral scanner images the data is stored as taken on tape during the flight. Such data contains calibration signals and other ancillary data. The general use of such raw sensor data from data banks does not look attractive since standardization of the tape data format does not exist. A handbook would be required to explain where and how the data can be made into images from the tape. Storage of infrared images on the original tape has been a common practice at The University of Michigan. The volume of such data is not yet excessive.

GROUND TRUTH
(True Set Membership)

REMOTE SENSING SYSTEM ANNOTATION OF SET MEMBERSHIP	T	$\overline{T}$
T	23	8
$\overline{T}$	10	1584

FIGURE 3-1. HYPOTHETICAL RESULT OF SYSTEM TEST OVER TEST SITES

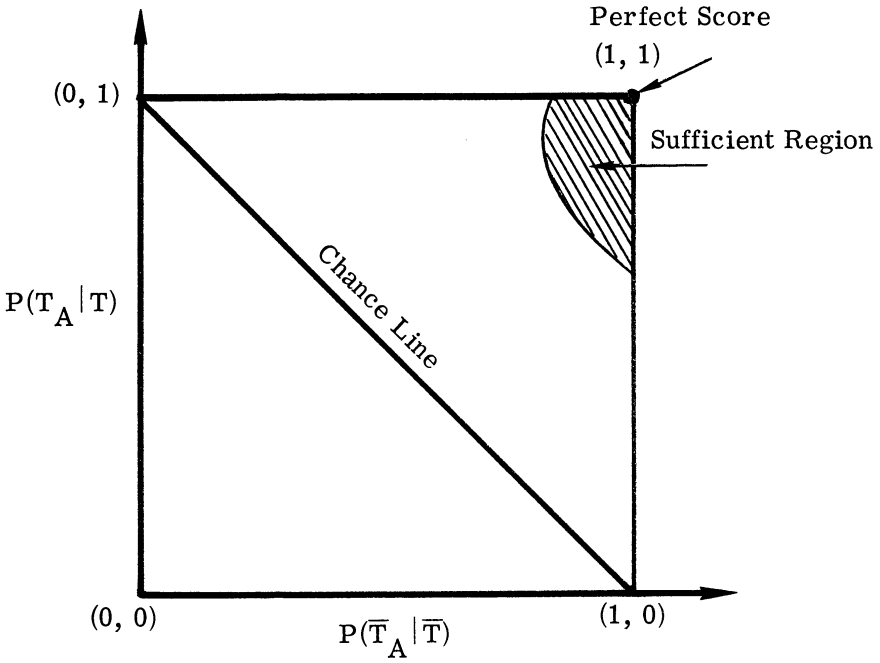


FIGURE 3-2. SYSTEM SCORE GRAPH

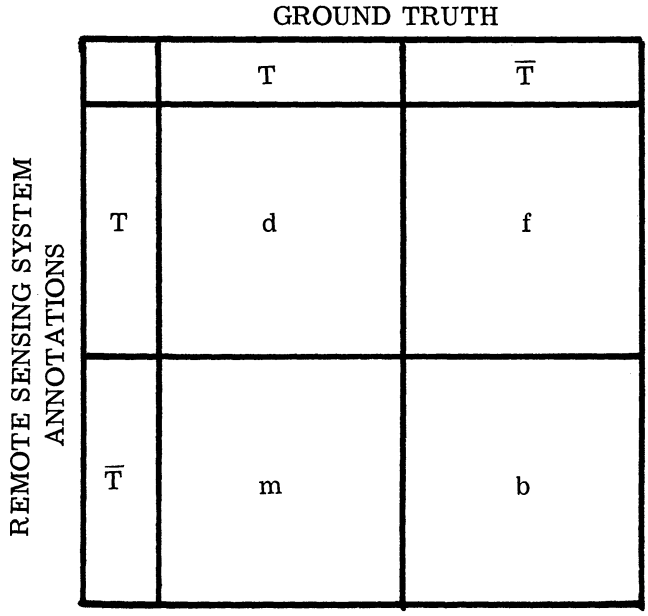


FIGURE 3-3. NORMALIZED TEST SITE DATA FOR THE NUMBER OF TARGETS AND NOT-TARGETS PER GRID. Each grid contains either a target, T, or a not-target, $\overline{T}$.

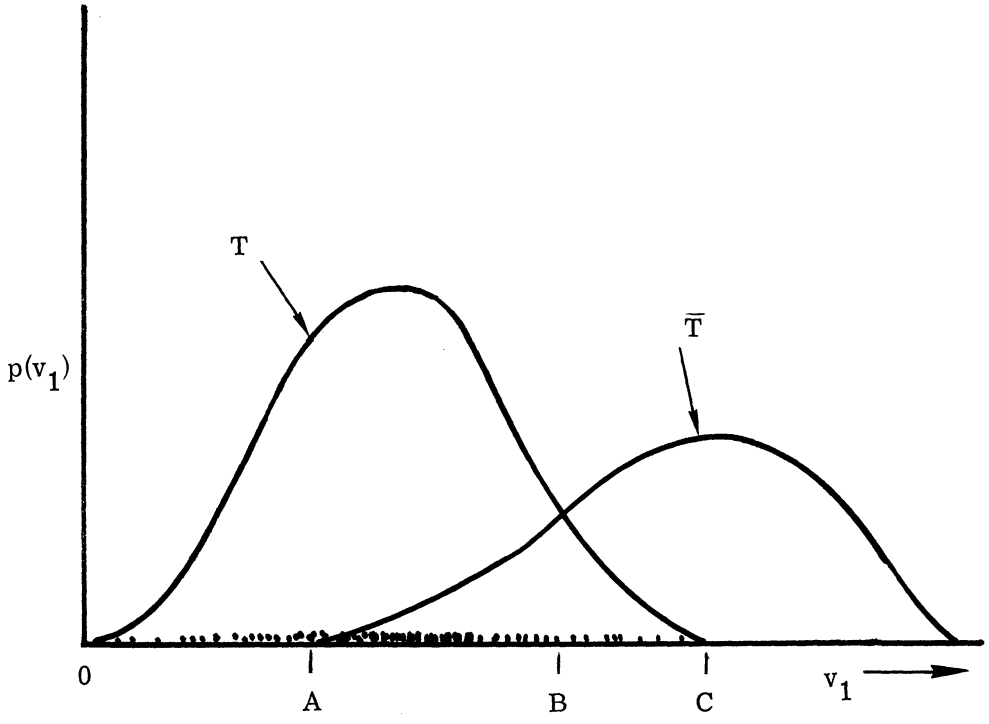


FIGURE 3-4. ONE-ATTRIBUTE SYSTEM. The plot of target elements according to attribute value, V_1, is indicated by dots. The probability density of target attributes $p(V_1)$ is also shown as a function of V_1. The curve for the not-target set is shown as a dashed line.

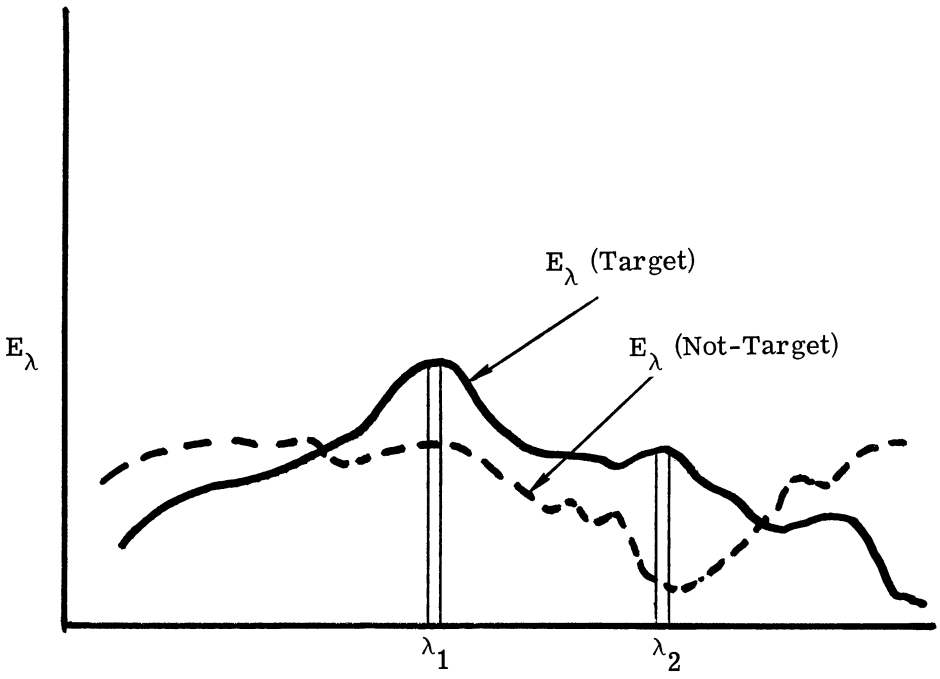


FIGURE 3-5. HYPOTHETICAL SPECTRAL IRRADIANCES AT A REMOTE SENSING DEVICE APERTURE

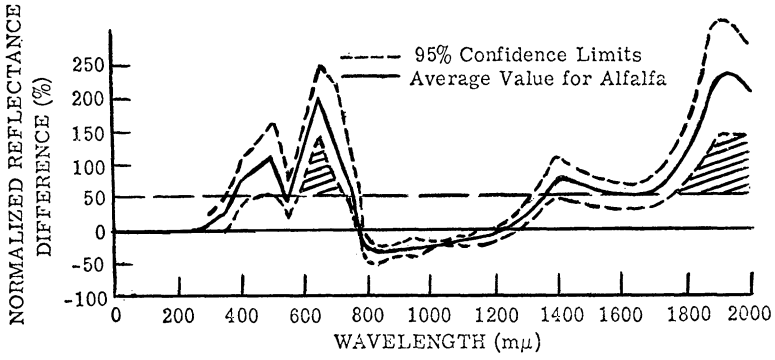
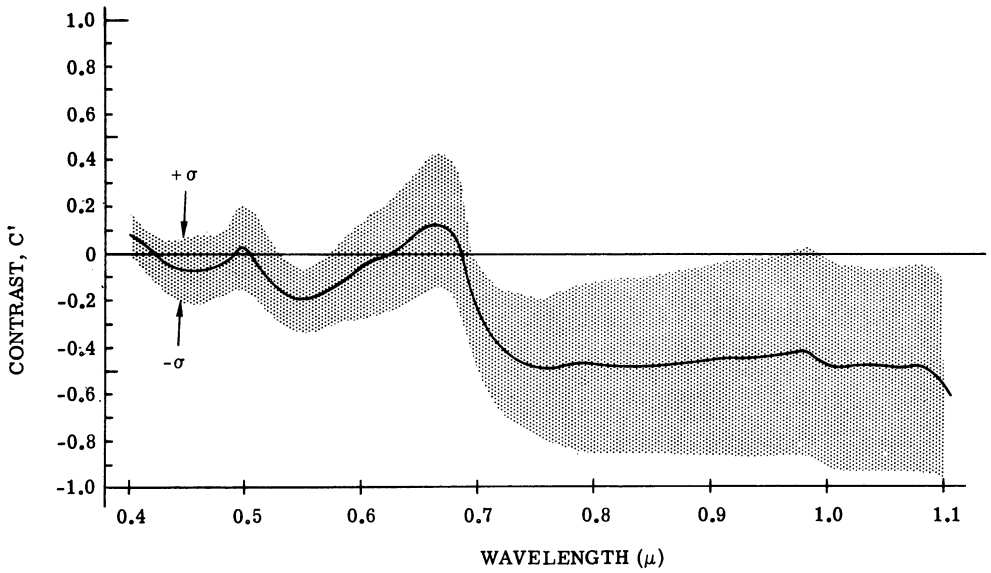
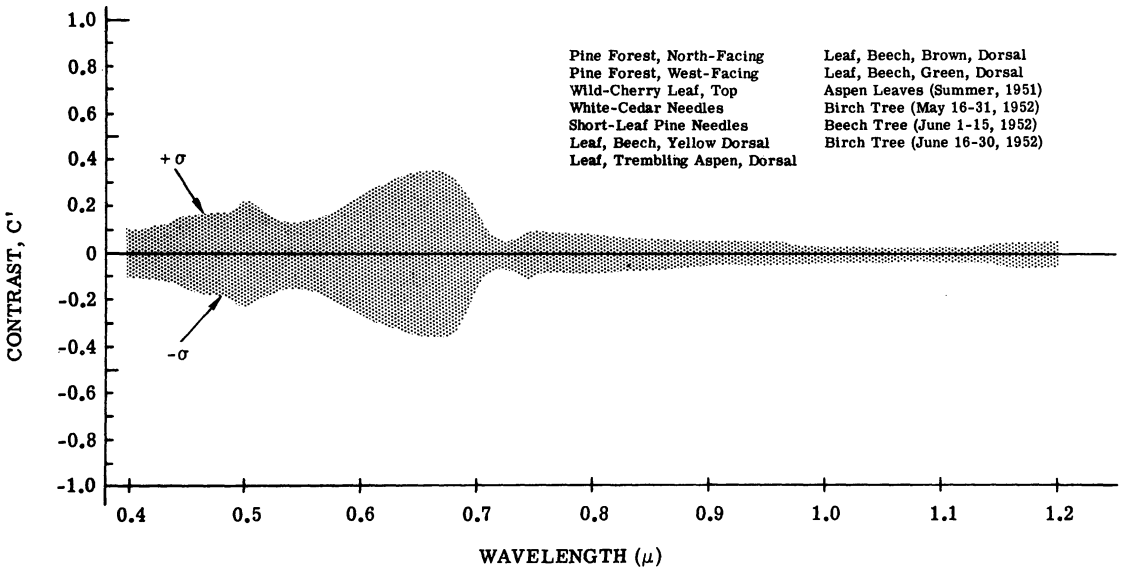

NORMALIZED REFLECTANCE DIFFERENCE (%)
250
200
150
100
50
0
-50
-100
0
200
400
600
800
1000
1200
1400
1600
1800
2000
WAVELENGTH (mμ)
95% Confidence Limits
Average Value for Alfalfa

FIGURE 4-1.



WAVELENGTH (μ)	(<-1)	(-1--.8)	(-.8--.6)	(-.6--.4)	(-.4--.2)	(-.2-0)	(0-.2)	(.2-.4)	(.4-.6)	(.6-.8)	(.8-1)	(>1)
.400	0	0	0	0	0	9	39	6	0	0	0	0
.420	0	0	0	0	0	29	24	1	0	0	0	0
.440	0	0	0	0	6	29	18	1	0	0	0	0
.460	0	0	0	0	9	33	18	3	0	0	0	0
.480	0	0	0	0	8	33	20	2	0	0	0	0
.500	0	0	0	0	11	29	32	18	0	0	0	0
.520	0	0	0	3	18	43	26	0	0	0	0	0
.540	0	0	0	6	35	43	6	0	0	0	0	0
.560	0	0	0	5	37	40	8	0	0	0	0	0
.580	0	0	0	3	33	29	23	2	0	0	0	0
.600	0	0	0	4	25	21	33	7	0	0	0	0
.620	0	0	0	4	21	17	32	16	0	0	0	0
.640	0	0	0	5	15	15	23	30	2	0	0	0
.660	0	0	0	4	8	13	21	30	13	1	0	0
.680	0	0	0	5	6	15	17	20	7	2	0	0
.700	0	0	3	12	25	13	8	2	0	0	0	0
.720	0	0	14	7	18	8	2	0	0	0	0	0
.740	0	6	13	4	15	7	4	0	0	0	0	0
.760	0	5	4	1	6	4	1	0	0	0	0	0
.780	0	5	4	1	5	5	1	0	0	0	0	0
.800	0	7	2	0	6	5	1	0	0	0	0	0
.820	0	7	2	0	5	5	2	0	0	0	0	0
.840	0	8	1	0	5	5	2	0	0	0	0	0
.860	0	9	0	0	5	5	2	0	0	0	0	0
.880	0	9	0	0	3	7	2	0	0	0	0	0
.900	0	18	0	0	2	16	6	0	0	0	0	0
.920	0	18	0	0	2	16	6	0	0	0	0	0
.940	0	18	0	0	2	16	6	0	0	0	0	0
.960	0	18	0	0	1	17	6	0	0	0	0	0
.980	0	15	0	0	0	15	5	0	0	0	0	0
1.000	5	15	5	0	5	15	5	0	0	0	0	0
1.020	5	15	5	0	5	15	5	0	0	0	0	0
1.040	5	15	5	0	5	15	5	0	0	0	0	0
1.060	5	15	5	0	5	15	5	0	0	0	0	0
1.080	5	15	5	0	5	15	5	0	0	0	0	0
1.100	5	15	5	0	5	15	0	0	0	0	0	0
1.120	3	6	3	0	3	0	0	0	0	0	0	0
1.140	3	6	2	1	3	0	0	0	0	0	0	0
1.160	3	6	1	2	1	2	0	0	0	0	0	0
1.180	3	6	1	2	1	2	0	0	0	0	0	0
1.200	3	6	1	2	1	2	0	0	0	0	0	0

FIGURE 4-2.



Wavelength (μ)	(<-1)	(-1--.8)	(-.8--.6)	(-.6--.4)	(-.4--.2)	(-.2-0)	(0-.2)	(.2-.4)	(.4-.6)	(.6-.8)	(.8-1)	(>1)
.400	0	0	0	0	0	12	24	0	0	0	0	0
.420	0	0	0	0	0	14	22	0	0	0	0	0
.440	0	0	0	0	2	12	20	2	0	0	0	0
.460	0	0	0	1	7	12	21	7	1	0	0	0
.480	0	0	0	0	6	15	22	6	0	0	0	0
.500	0	0	0	5	18	21	33	18	5	0	0	0
.520	0	0	0	2	13	30	40	13	2	0	0	0
.540	0	0	0	1	10	34	44	10	1	0	0	0
.560	0	0	0	1	10	34	44	10	1	0	0	0
.580	0	0	0	4	16	25	35	16	4	0	0	0
.600	0	0	0	9	15	20	32	15	9	0	0	0
.620	0	0	2	11	12	20	30	12	11	2	0	0
.640	0	0	4	9	14	18	28	14	9	4	0	0
.660	0	1	5	7	13	19	29	13	7	5	1	0
.680	0	0	3	5	8	12	20	8	5	3	0	0
.700	0	0	0	0	3	18	25	3	0	0	0	0
.720	0	0	0	0	0	21	28	0	0	0	0	0
.740	0	0	0	0	0	21	28	0	0	0	0	0
.760	0	0	0	0	0	3	6	0	0	0	0	0
.780	0	0	0	0	0	3	6	0	0	0	0	0
.800	0	0	0	0	0	3	6	0	0	0	0	0
.820	0	0	0	0	0	3	6	0	0	0	0	0
.840	0	0	0	0	0	3	6	0	0	0	0	0
.860	0	0	0	0	0	3	6	0	0	0	0	0
.880	0	0	0	0	0	3	6	0	0	0	0	0
.900	0	0	0	0	0	15	21	0	0	0	0	0
.920	0	0	0	0	0	15	21	0	0	0	0	0
.940	0	0	0	0	0	15	21	0	0	0	0	0
.960	0	0	0	0	0	15	21	0	0	0	0	0
.980	0	0	0	0	0	10	15	0	0	0	0	0
1.000	0	0	0	0	0	10	15	0	0	0	0	0
1.020	0	0	0	0	0	10	15	0	0	0	0	0
1.040	0	0	0	0	0	10	15	0	0	0	0	0
1.060	0	0	0	0	0	9	16	0	0	0	0	0
1.080	0	0	0	0	0	9	16	0	0	0	0	0
1.100	0	0	0	0	0	9	16	0	0	0	0	0
1.120	0	0	0	0	0	3	6	0	0	0	0	0
1.140	0	0	0	0	0	3	6	0	0	0	0	0
1.160	0	0	0	0	0	3	6	0	0	0	0	0
1.180	0	0	0	0	0	3	6	0	0	0	0	0
1.200	0	0	0	0	0	3	6	0	0	0	0	0

FIGURE 4-3.

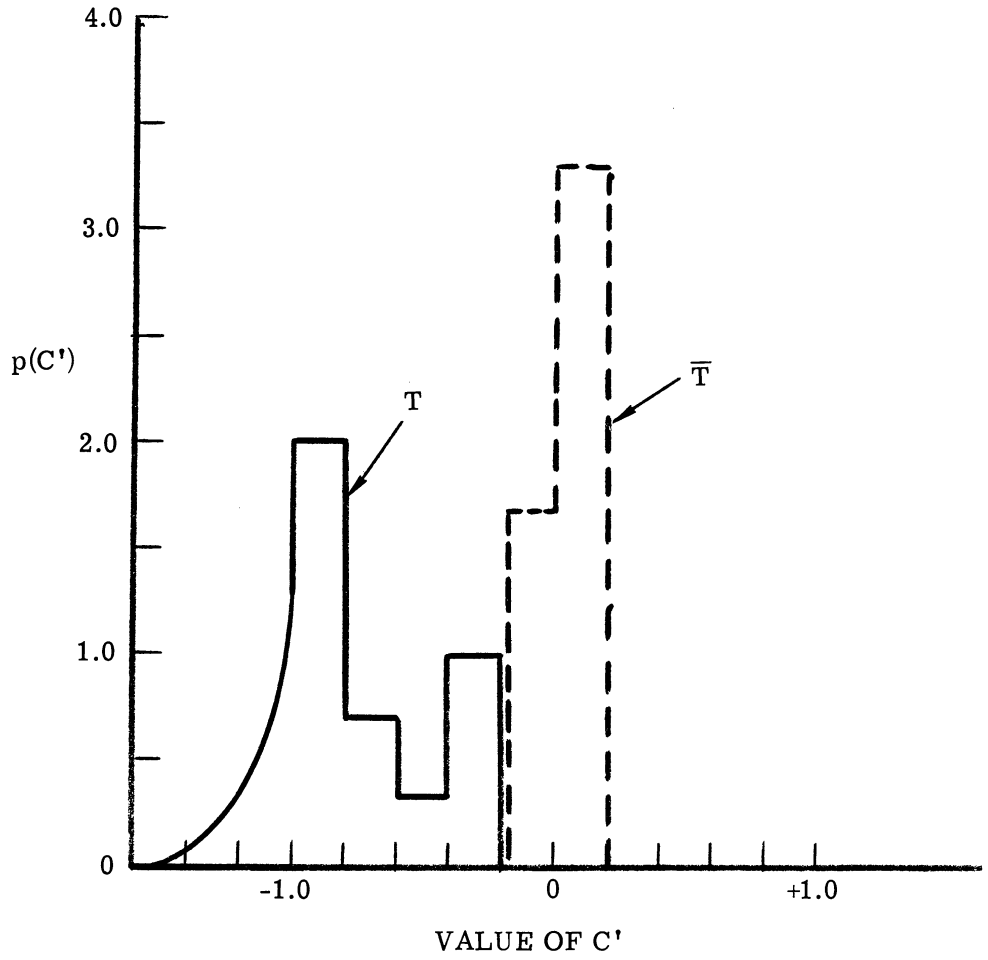


FIGURE 4-4. PROBABILITY DENSITY USING C' AT 1.140 μm AS A SINGLE SPECTRAL ATTRIBUTE

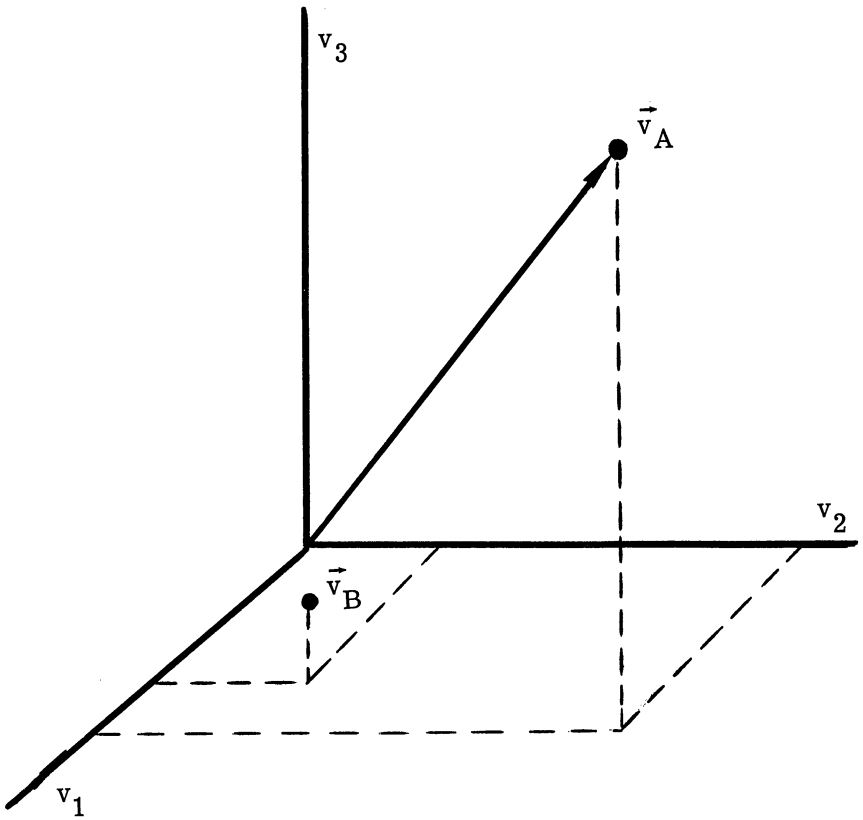


FIGURE 4-5. THREE-DIMENSIONAL ATTRIBUTE SPACE. Each triple-attribute measure corresponds to a three-dimensional vector, $\vec{V}$.

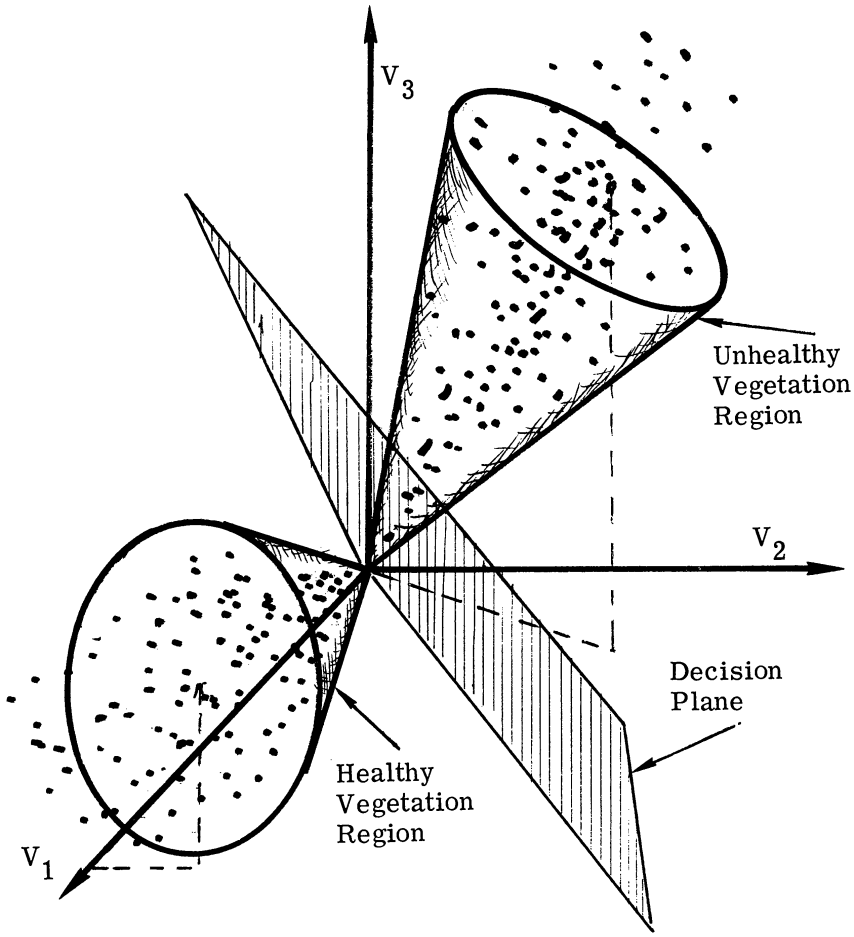


FIGURE 4-6. HYPOTHETICAL PLOT OF DISCRIMINATION OF INFRARED EKTACHROME PLOTTED AS A THREE-ATTRIBUTE DISCRIMINATION SYSTEM FOR HEALTHY AND UNHEALTHY VEGETATION

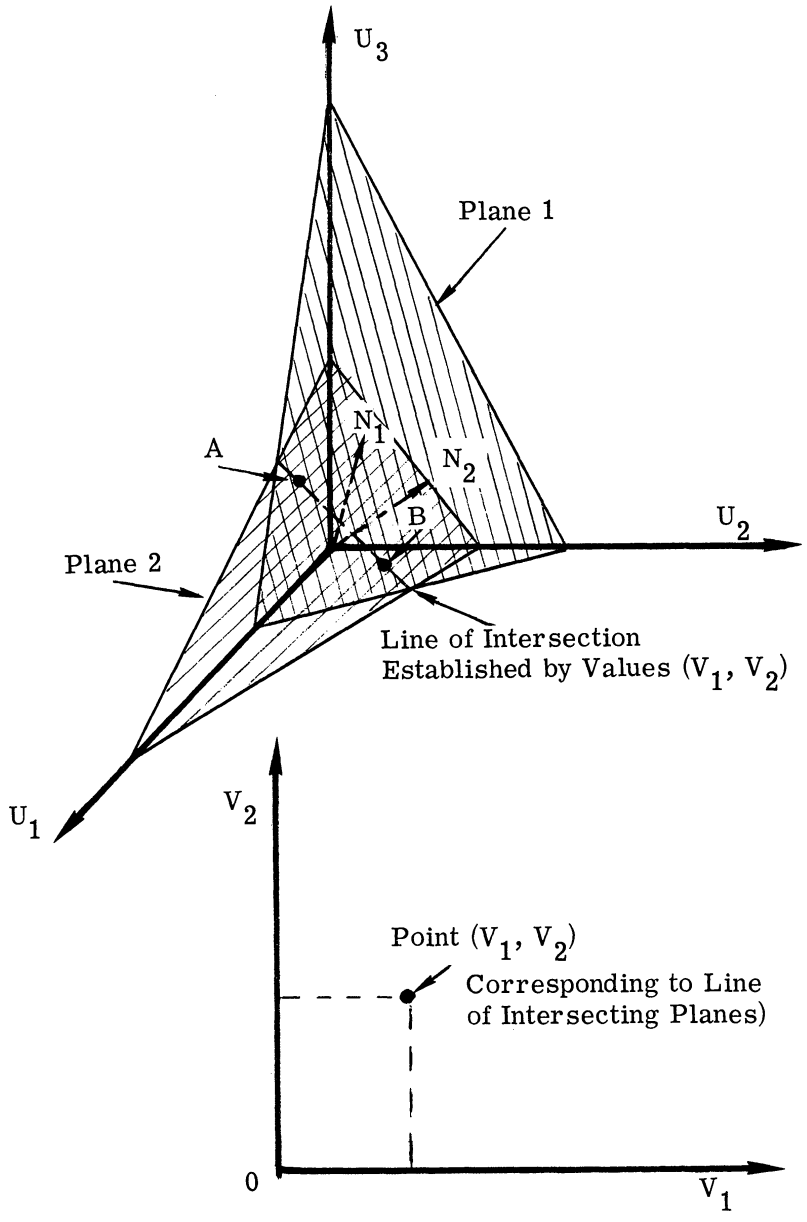


FIGURE 4-7. MAPPING THREE-ATTRIBUTE SPACE INTO TWO-ATTRIBUTE SPACE BY FILTERS

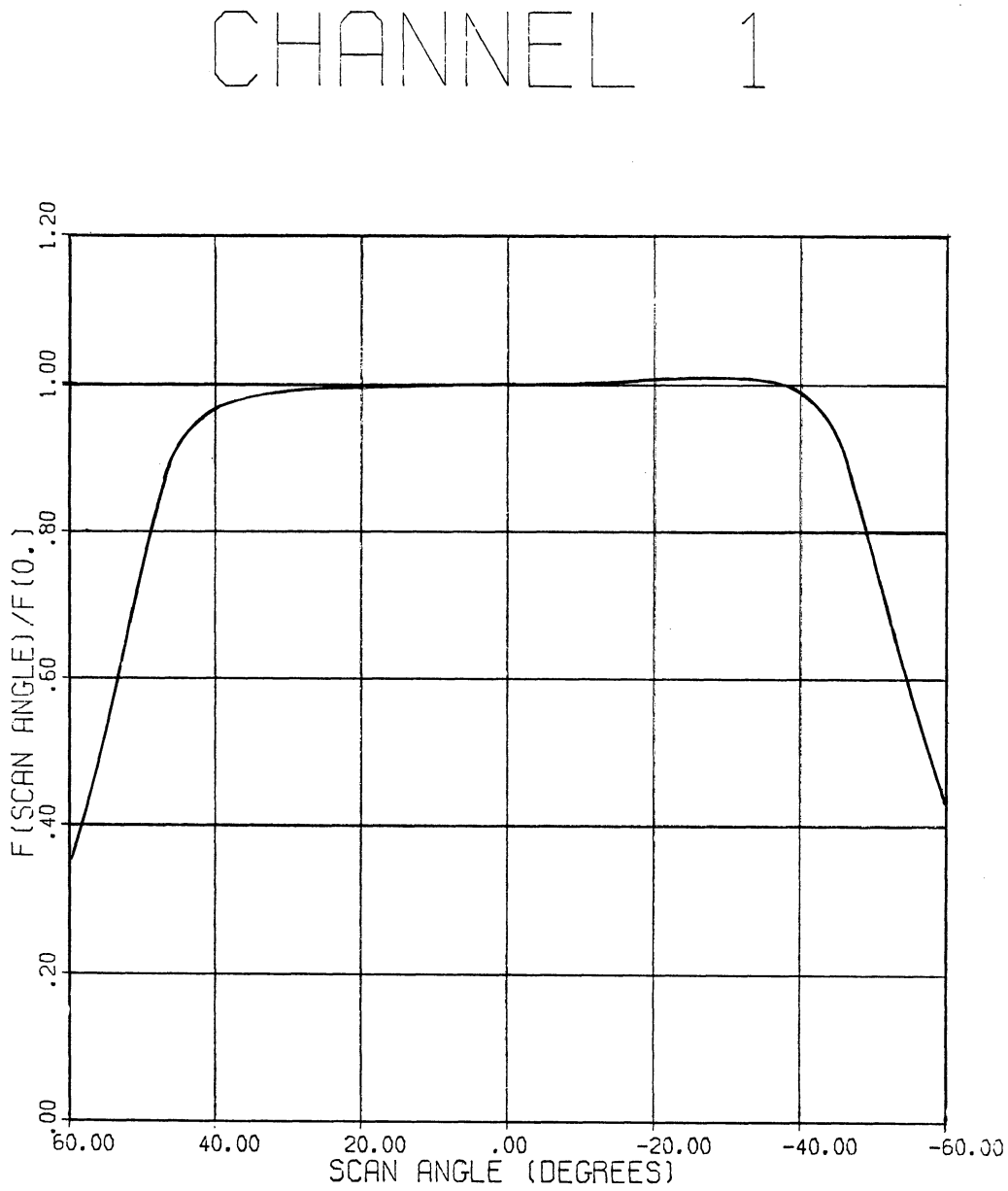

CHANNEL 1
F(SCAN ANGLE)/F(0.)
1.20
1.00
.80
.60
.40
.20
.00
60.00
40.00
20.00
.00
-20.00
-40.00
-60.00
SCAN ANGLE (DEGREES)

FIGURE 4-8.

CHANNEL 3

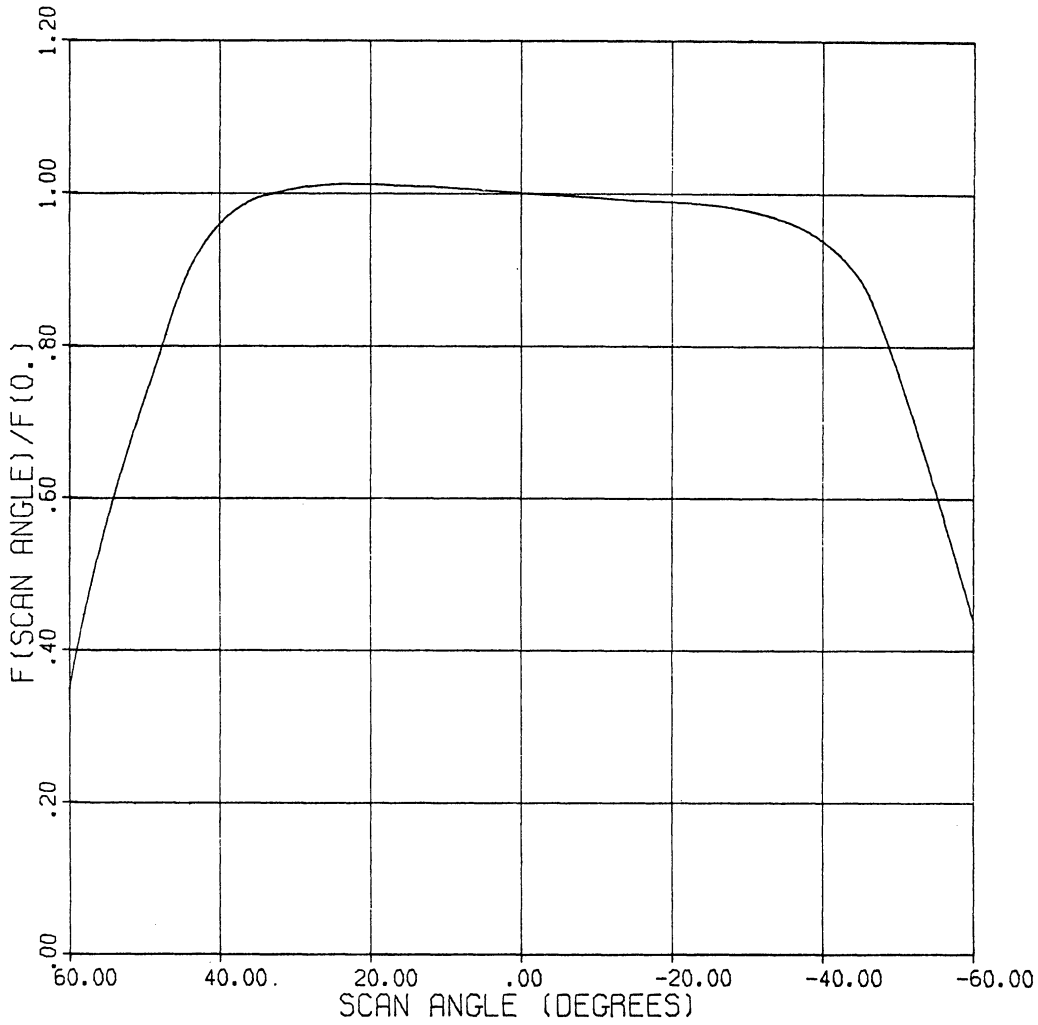


FIGURE 4-9.

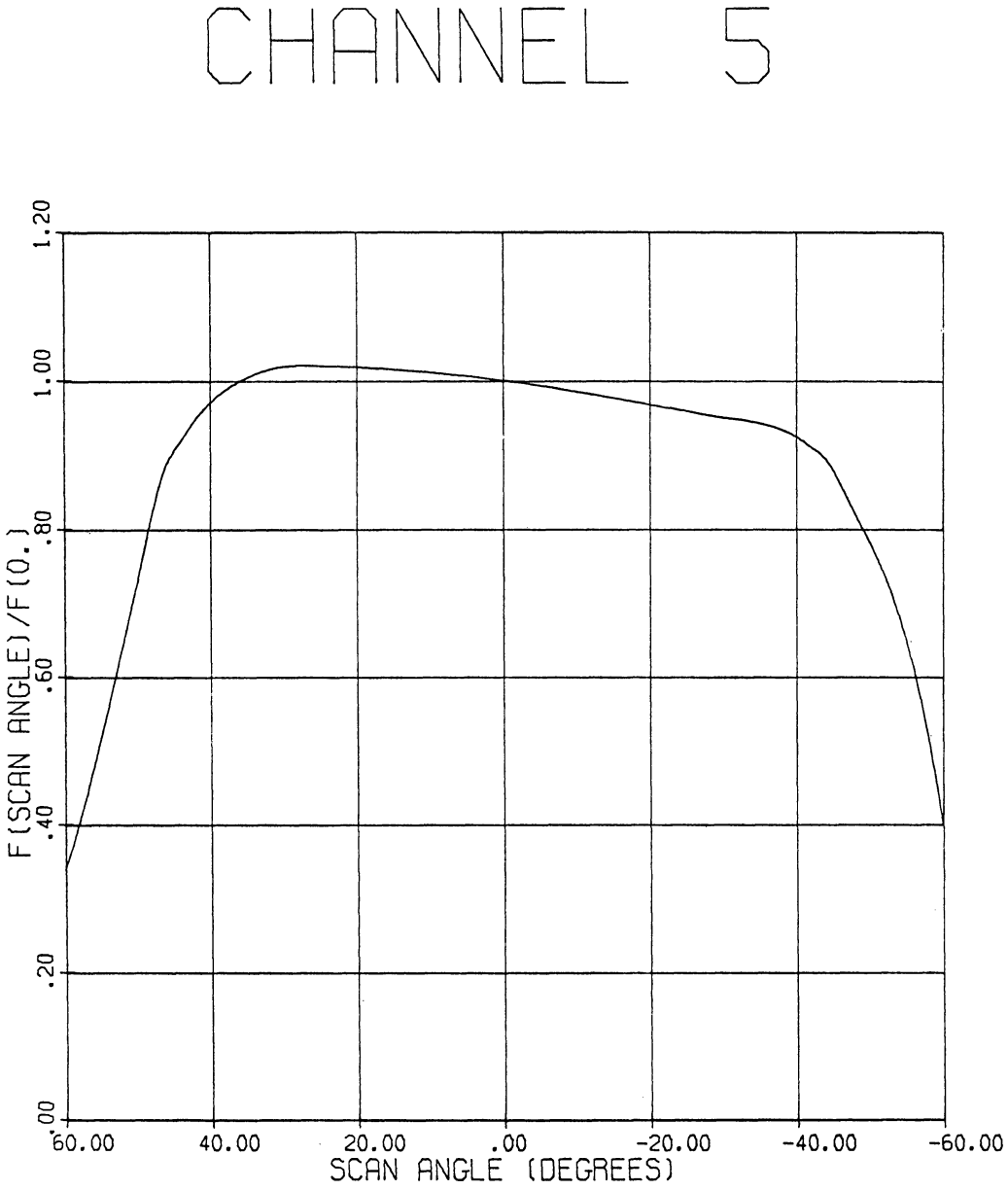

CHANNEL 5
F(SCAN ANGLE)/F(0.)
1.20
1.00
.80
.60
.40
.20
.00
60.00
40.00
20.00
.00
-20.00
-40.00
-60.00
SCAN ANGLE (DEGREES)

FIGURE 4-10.

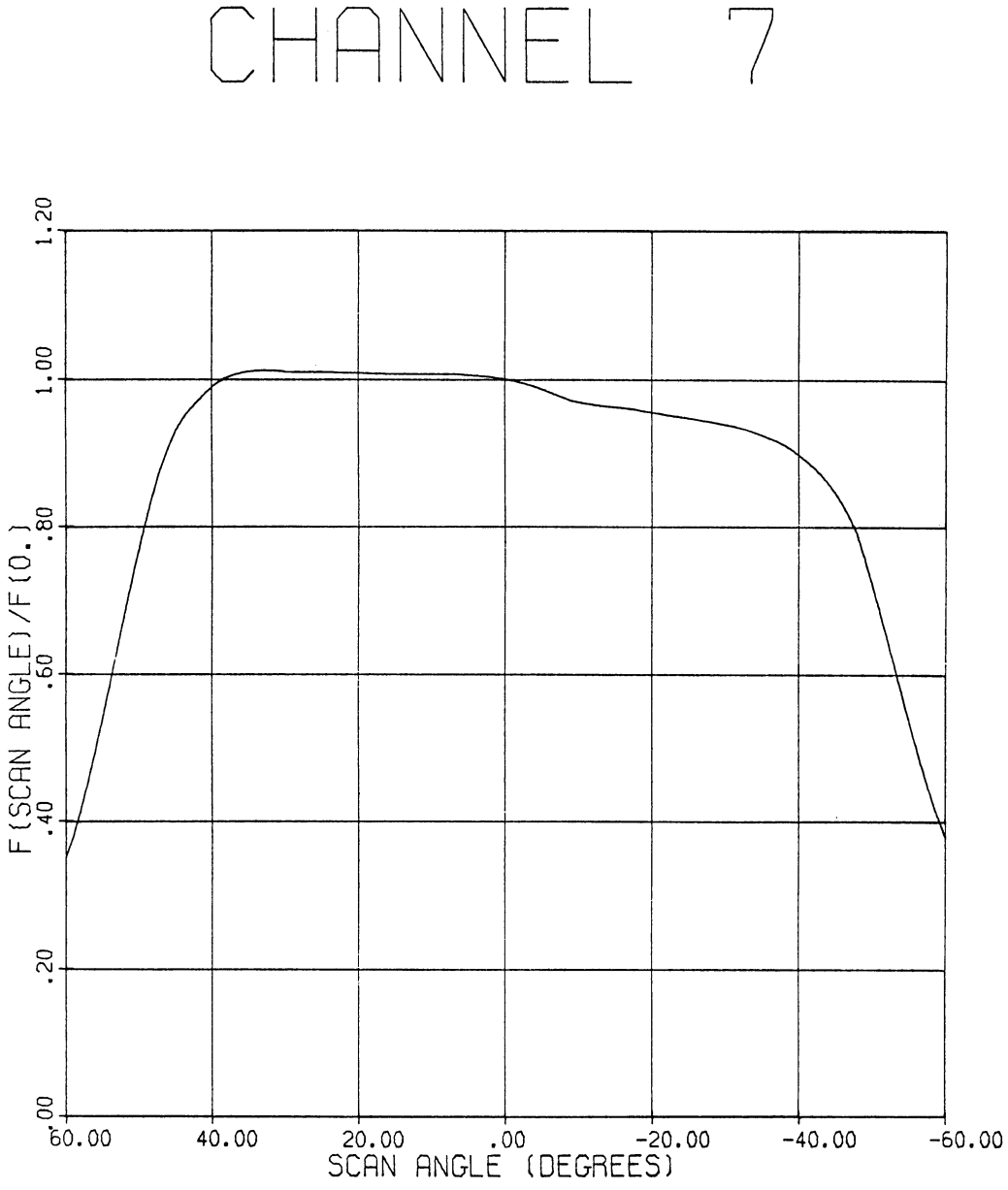

CHANNEL 7
F(SCAN ANGLE)/F(0.)
1.20
1.00
.80
.60
.40
.20
.00
60.00
40.00
20.00
.00
-20.00
-40.00
-60.00
SCAN ANGLE (DEGREES)

FIGURE 4-11.

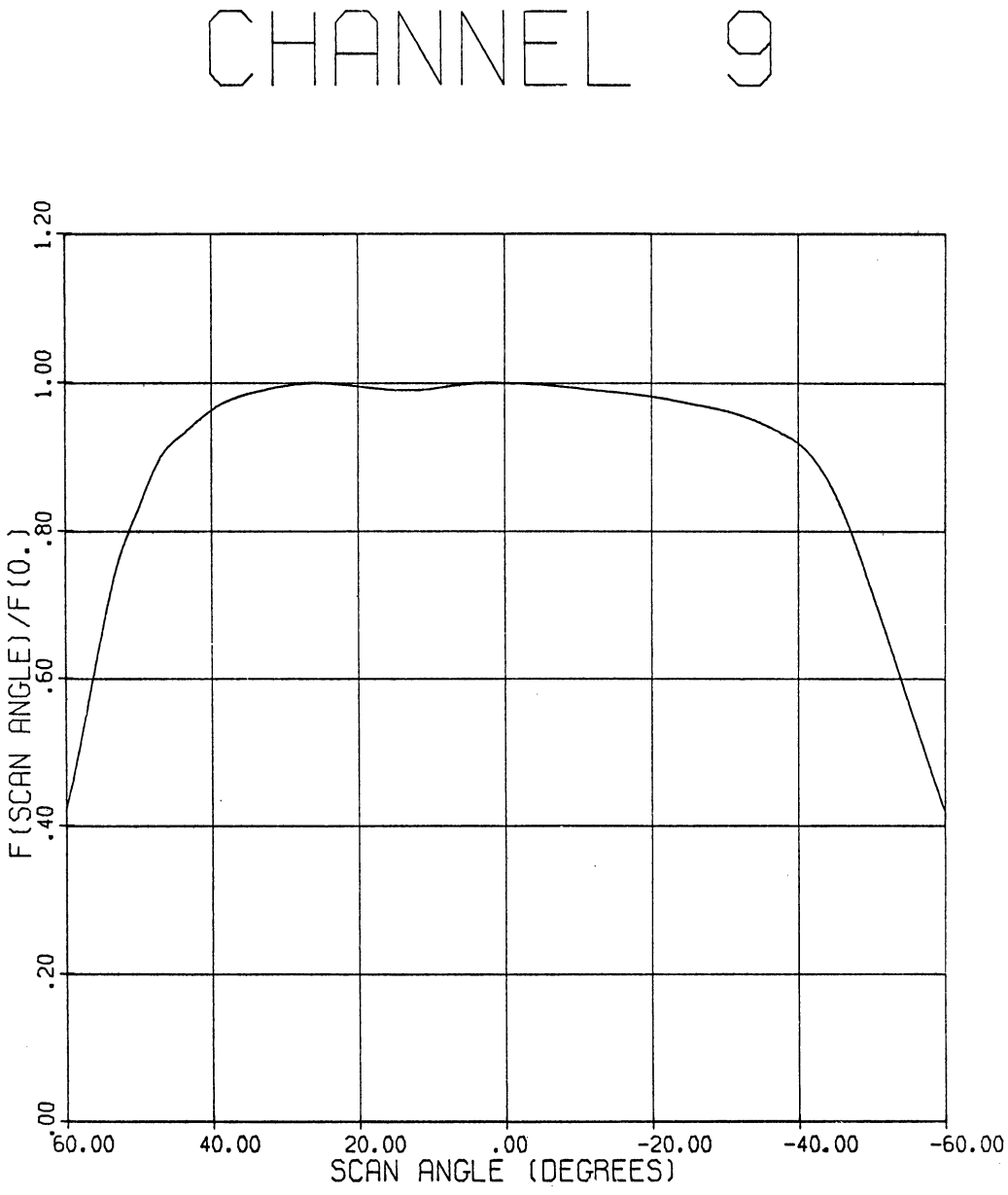

CHANNEL 9
F(SCAN ANGLE)/F(0.)
1.20
1.00
.80
.60
.40
.20
.00
60.00
40.00
20.00
.00
-20.00
-40.00
-60.00
SCAN ANGLE (DEGREES)

FIGURE 4-12.

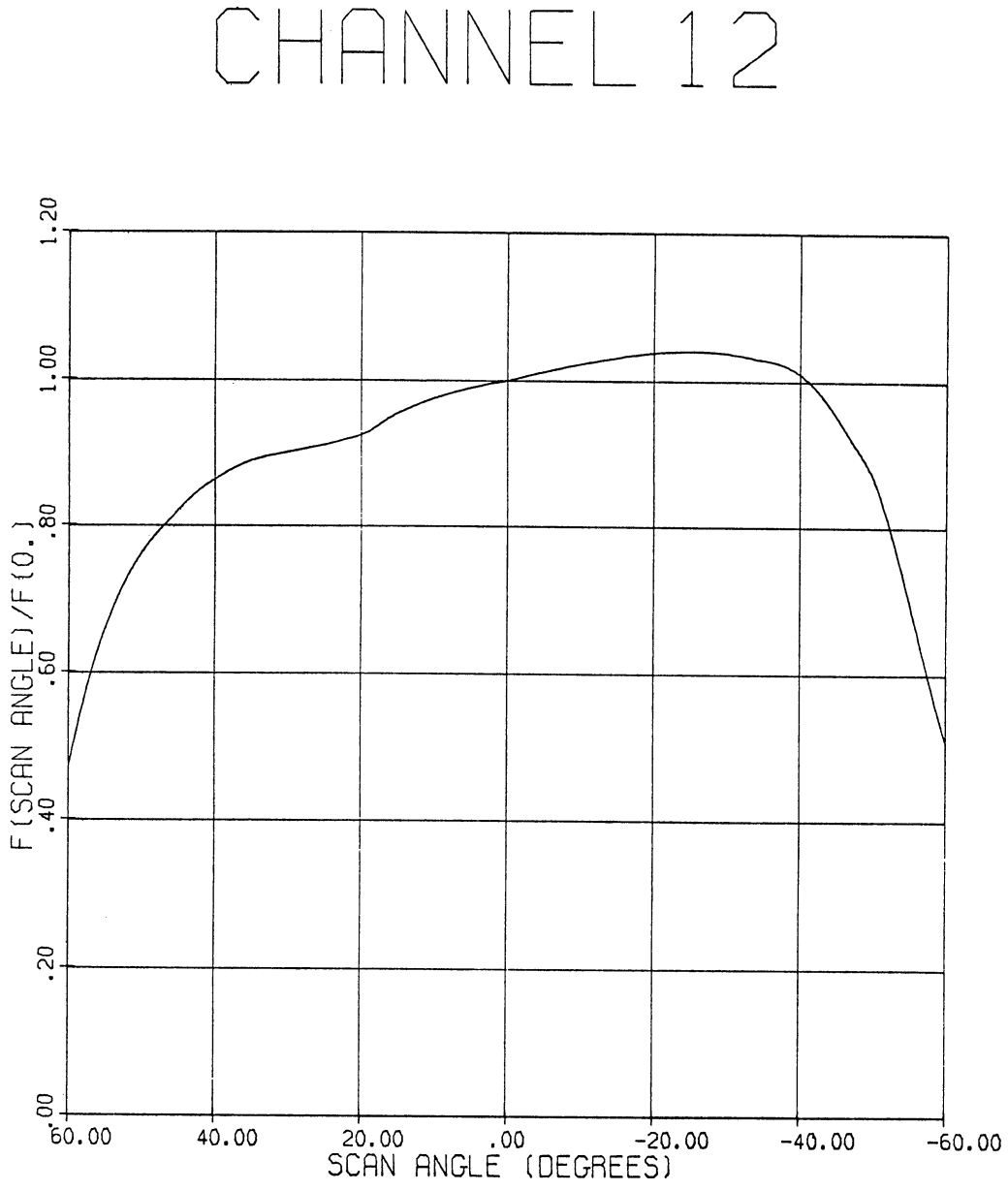

CHANNEL 12
F(SCAN ANGLE)/F(0.)
1.20
1.00
.80
.60
.40
.20
.00
60.00
40.00
20.00
.00
-20.00
-40.00
-60.00
SCAN ANGLE (DEGREES)

FIGURE 4-13.

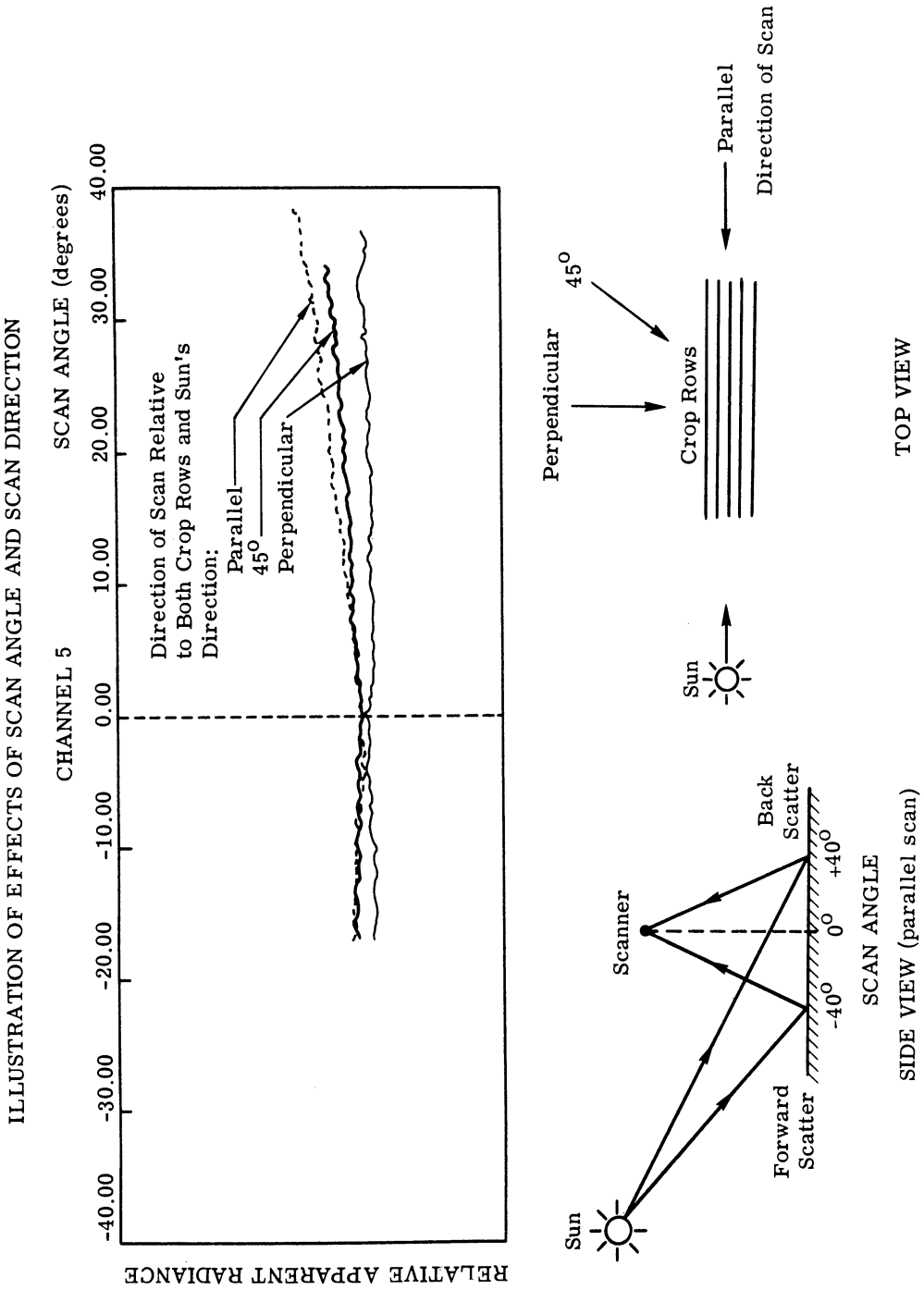


ILLUSTRATION OF EFFECTS OF SCAN ANGLE AND SCAN DIRECTION
CHANNEL 5
SCAN ANGLE (degrees)
-40.00
-30.00
-20.00
-10.00
0.00
10.00
20.00
30.00
40.00
RELATIVE APPARENT RADIANCE
Direction of Scan Relative to Both Crop Rows and Sun's Direction:
Parallel
45°
Perpendicular
Perpendicular
45°
Crop Rows
Parallel
Direction of Scan
Sun
TOP VIEW
Scanner
Back Scatter
Forward Scatter
-40°
0°
+40°
SCAN ANGLE
Sun
SIDE VIEW (parallel scan)

FIGURE 4-14.

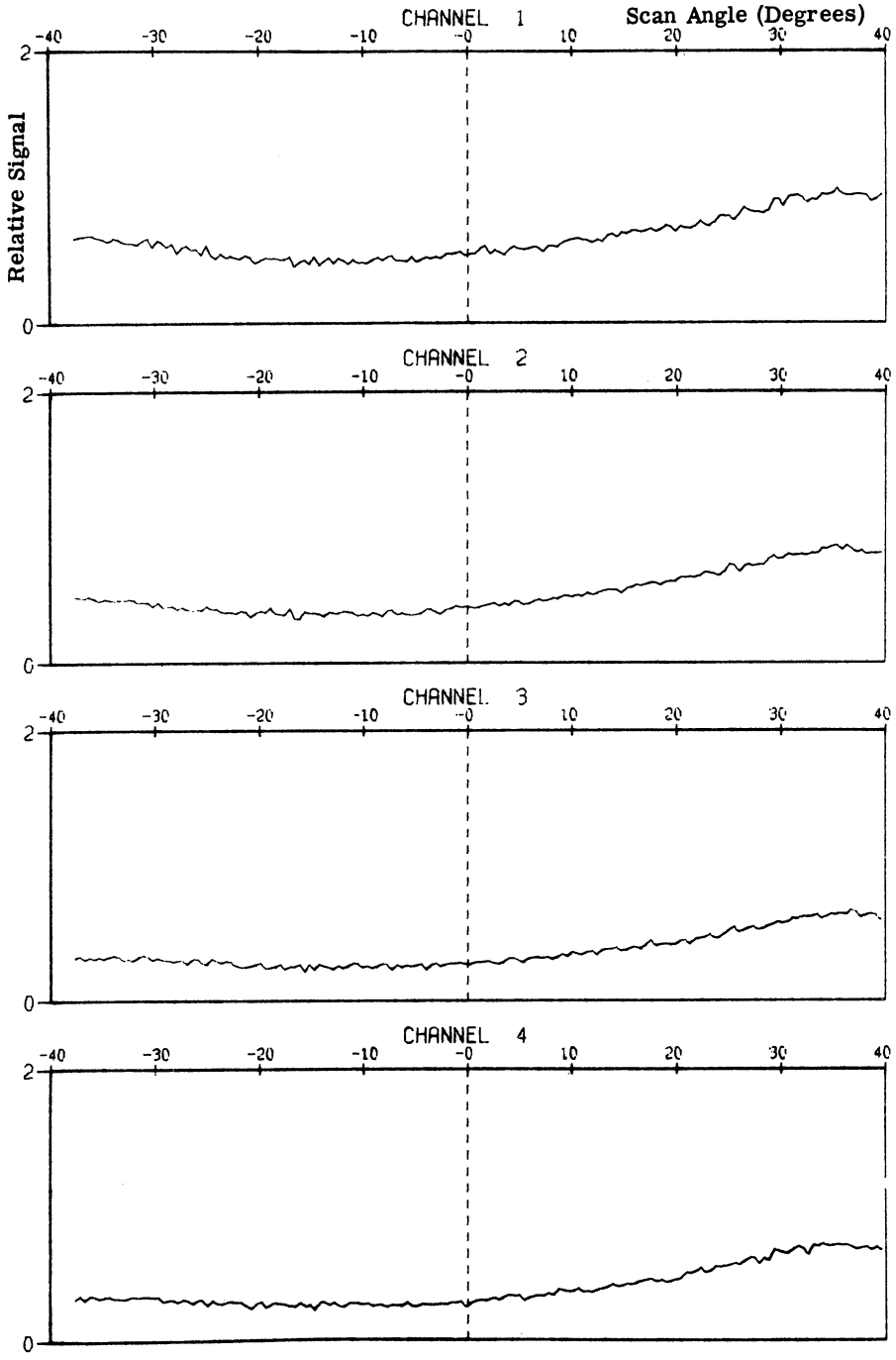

CHANNEL 1
Scan Angle (Degrees)
-40 -30 -20 -10 -0 10 20 30 40
2
Relative Signal
0
CHANNEL 2
-40 -30 -20 -10 -0 10 20 30 40
2
0
CHANNEL 3
-40 -30 -20 -10 -0 10 20 30 40
2
0
CHANNEL 4
-40 -30 -20 -10 -0 10 20 30 40
2
0

FIGURE 4-15.

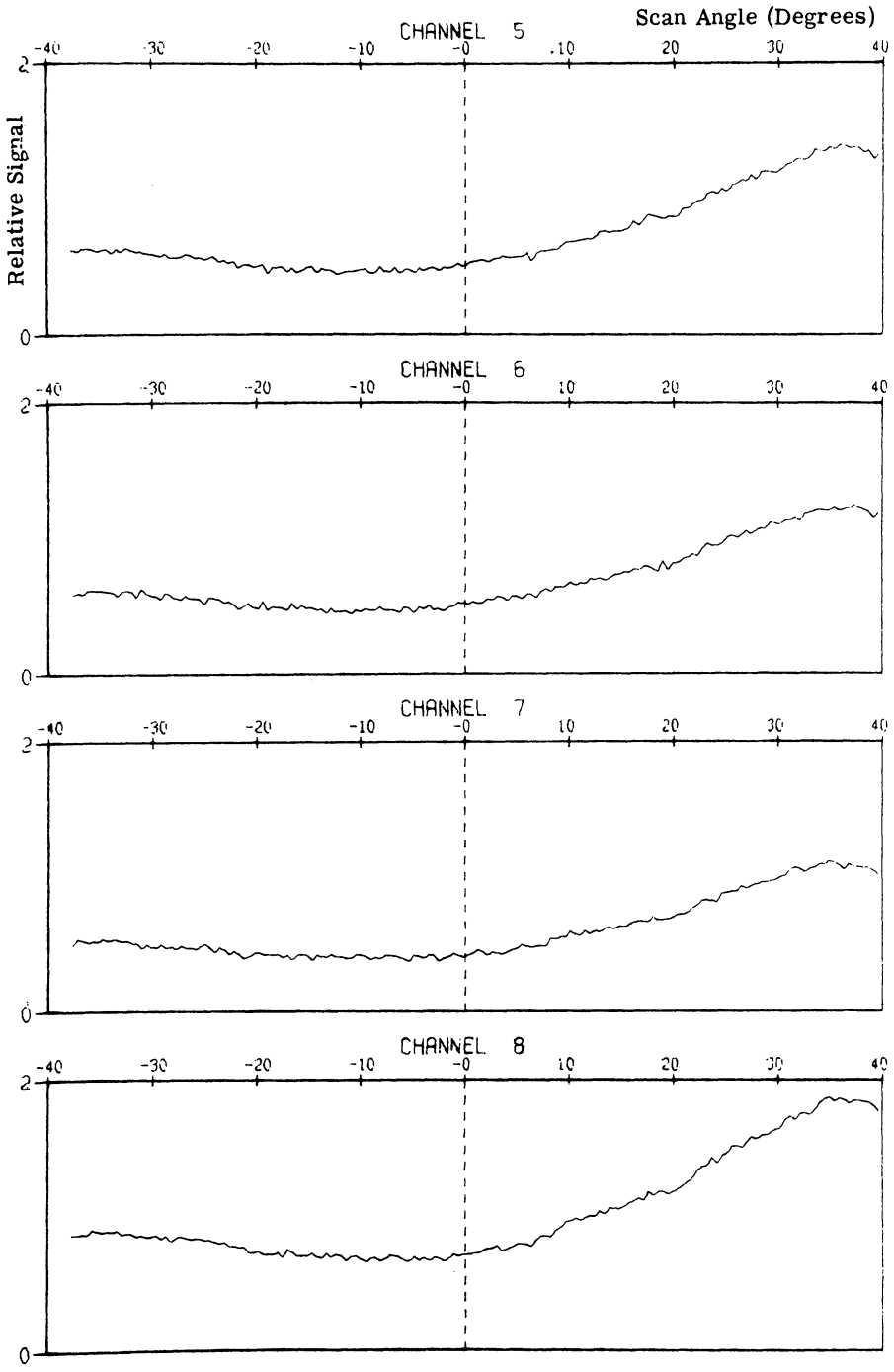

Scan Angle (Degrees)
CHANNEL 5
CHANNEL 6
CHANNEL 7
CHANNEL 8
Relative Signal
-40
-30
-20
-10
-0
10
20
30
40
2
0

FIGURE 4-16.

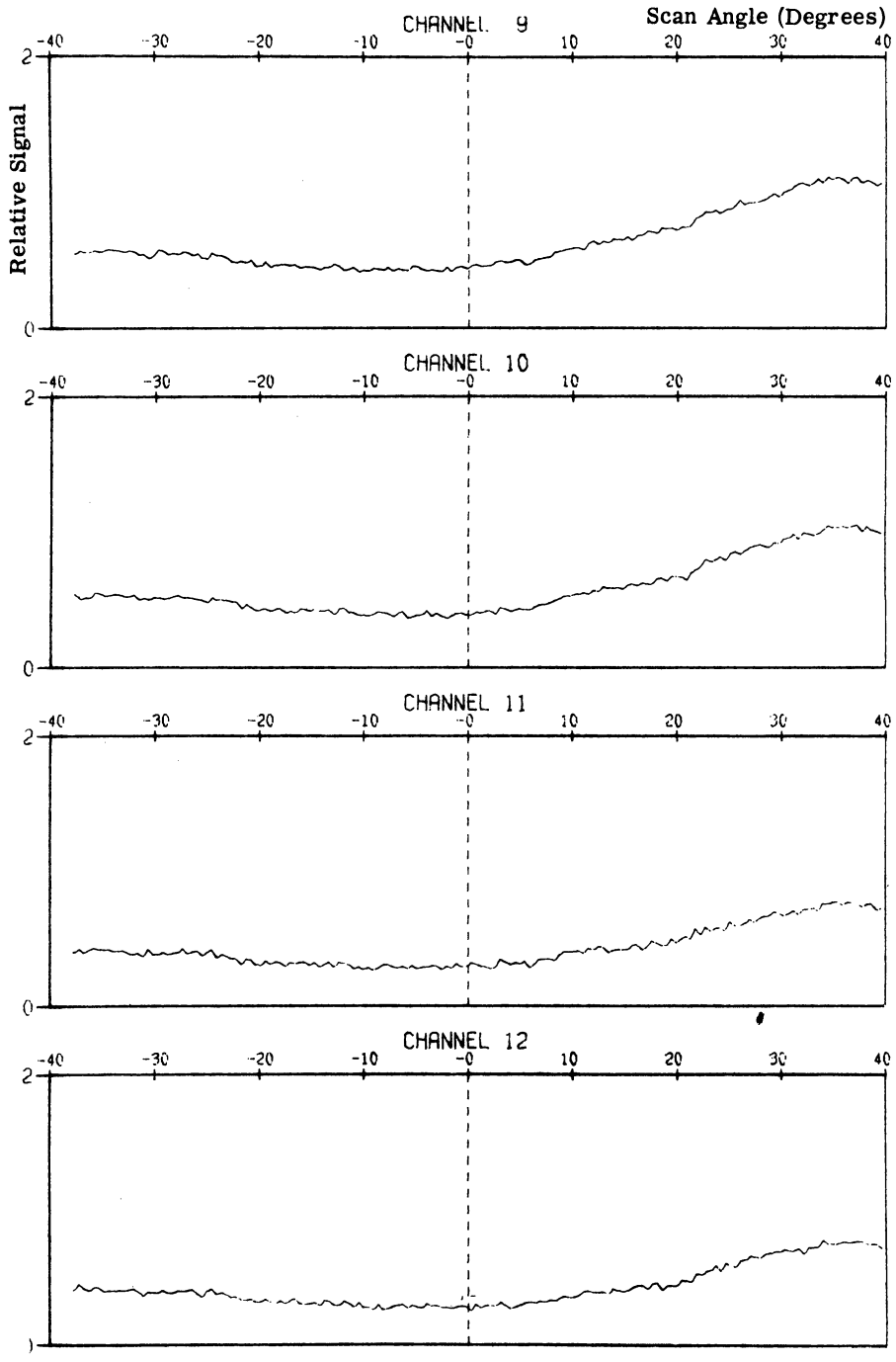

CHANNEL 9
Scan Angle (Degrees)
-40 -30 -20 -10 -0 10 20 30 40
2
Relative Signal
0
CHANNEL 10
-40 -30 -20 -10 -0 10 20 30 40
2
0
CHANNEL 11
-40 -30 -20 -10 -0 10 20 30 40
2
0
CHANNEL 12
-40 -30 -20 -10 -0 10 20 30 40
2
1

FIGURE 4-17.

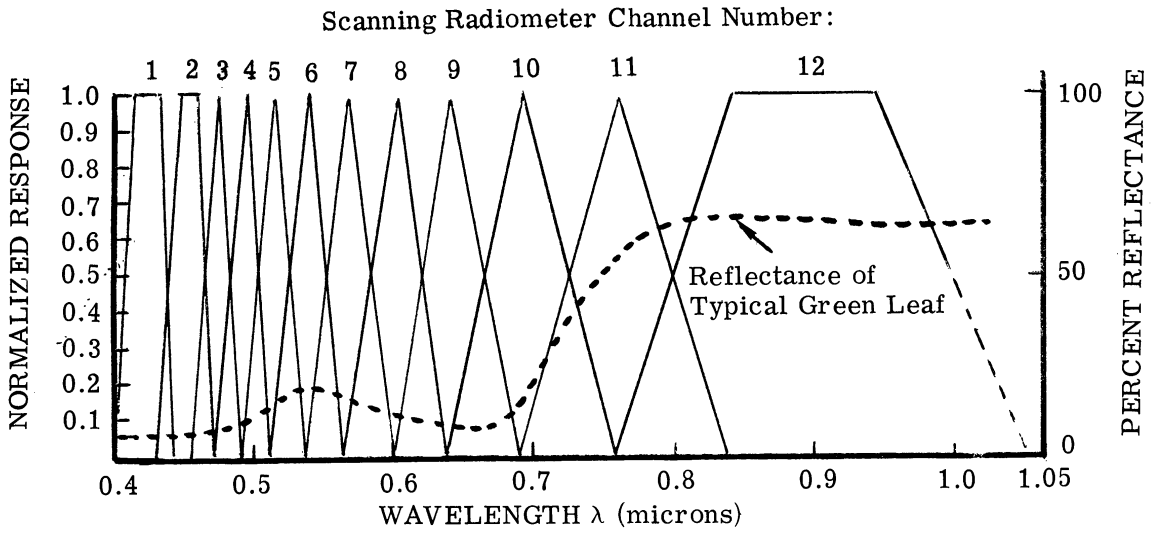


FIGURE 4-18.