Detection of Light

From the Ultraviolet to the Submillimeter

SECOND EDITION

G. H. Rieke
University of Arizona

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 1994
Reprinted 1996
First paperback edition 1996
Second edition 2003

Printed in the United Kingdom at the University Press, Cambridge

Typefaces Times 10.25 / 13.5 pt and Joanna *System* LaTeX 2$_\varepsilon$ [TB]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data

Rieke, G. H. (George Henry)
Detection of light : from the ultraviolet to the submillimeter / G. H. Rieke. – 2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 0 521 81636 X – ISBN 0 521 01710 6
1. Optical detectors. I. Title.
QC373.O59 R54 2002
621.36′2 – dc21 2002023375

ISBN 0 521 81636 X hardback
ISBN 0 521 01710 6 paperback

Contents

1

Introduction

We begin by covering background material in three areas. First, we need to establish the formalism and definitions for the imaginary signals we will be shining on our imaginary detectors. Second, we will describe general detector characteristics so we can judge the merits of the various types as they are discussed. Third, because solid state – and to some extent superconducting – physics will be so pervasive in our discussions, we include a very brief primer on those subjects.

1.1 Radiometry

There are some general aspects of electromagnetic radiation that need to be defined before we discuss how it is detected. Most of the time we will treat light as photons of energy; wave aspects will be important only for heterodyne receivers (involving detection through interference of the signal with a local source of power at nearly the same frequency). A photon has an energy of

$$E_{\mathrm{ph}} = h\nu = hc/\lambda, \tag{1.1}$$

where h ($= 6.626 \times 10^{-34}$ J s) is Planck's constant, ν and λ are, respectively, the frequency (in hertz) and wavelength (in meters) of the electromagnetic wave, and c ($= 2.998 \times 10^{8}$ m s^{-1}) is the speed of light. In the following discussion, we define a number of expressions for the power output of photon sources. Conversion from power to photons per second can be achieved by dividing by the desired form of equation (1.1).

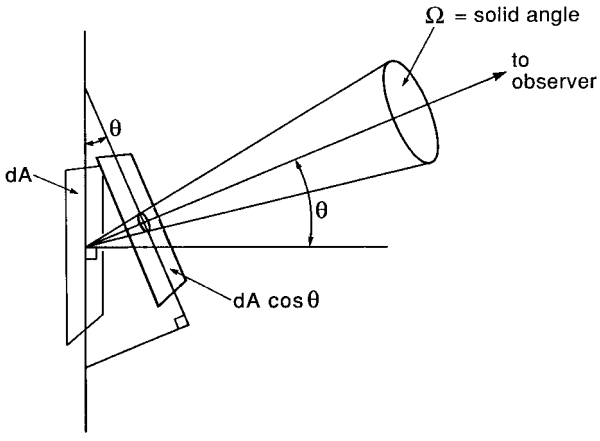

Figure 1.1. Geometry for computing radiance.

To compute the emission of an object, consider a projected area of a surface element *dA* onto a plane perpendicular to the direction of observation. As shown in Figure 1.1, it is $dA\cos\theta$, where θ is the angle between the direction of observation and the outward normal to *dA*. The spectral radiance per frequency interval, $\boldsymbol{L}_\nu$, is the power (in watts) leaving a unit projected area of the surface of the source (in square meters) into a unit solid angle (in steradians) and unit frequency interval (in hertz = 1/seconds). $\boldsymbol{L}_\nu$ has units of $\mathrm{W\,m^{-2}\ Hz^{-1}\ ster^{-1}}$. The spectral radiance per wavelength interval, $\boldsymbol{L}_\lambda$, has units of $\mathrm{W\,m^{-3}\ ster^{-1}}$. The radiance, $\boldsymbol{L}$, is the spectral radiance integrated over all frequencies or wavelengths; it has units of $\mathrm{W\,m^{-2}\ ster^{-1}}$. The radiant exitance, $\boldsymbol{M}$, is the integral of the radiance over the solid angle, Ω, and it is a measure of the total power emitted per unit surface area in units of $\mathrm{W\,m^{-2}}$.

We will deal only with Lambertian sources; the defining characteristic of such a source is that its radiance is constant regardless of the direction from which it is viewed. Blackbodies and "graybodies" are examples. A graybody is defined to emit as a blackbody but at the efficiency of its emissivity, ϵ (ranging from 0 to 1); a blackbody by definition has $\epsilon = 1$. The emission of a Lambertian source goes as the cosine of the angle between the direction of the radiation and the normal to the source surface. From the definition of projected area in the preceding paragraph, it can be seen that this emission pattern exactly compensates for the foreshortening of the surface as it is tilted away from being perpendicular to the line of sight. That is, for the element *dA*, the projected surface area and the emission decrease by the same cosine factor. Thus, if the entire source has the same temperature and emissivity, every unit area of its projected surface in the plane perpendicular to the observer's line of sight appears to be of the same brightness, independent of its actual angle to the line of sight. Keeping in mind this cosine dependence, and the definition of radiant exitance, the radiance and radiant exitance are related as

$$\boldsymbol{M} = \int \boldsymbol{L}\cos\theta\, d\Omega = 2\pi\boldsymbol{L}\int_0^{\pi/2}\sin\theta\cos\theta\, d\theta = \pi\boldsymbol{L}. \tag{1.2}$$

The flux emitted by the source, Φ, is the radiant exitance times the total surface area of the source, that is, the power emitted by the entire source. For example, for a spherical source of radius R,

$$\Phi = 4\pi R^2 \boldsymbol{M} = 4\pi^2 R^2 \boldsymbol{L}. \tag{1.3}$$

Although there are other types of Lambertian sources, we will consider only sources that have spectra resembling those of blackbodies, for which

$$\boldsymbol{L}_\nu = \frac{\epsilon[2h\nu^3/(c/\mathrm{n})^2]}{e^{h\nu/kT} - 1}, \tag{1.4}$$

where ϵ is the emissivity of the source and T its temperature, n is the refractive index of the medium into which the source radiates, and k ($= 1.38 \times 10^{-23}$ J K^{-1}) is the Boltzmann constant. According to Kirchhoff's law, the portion of the energy absorbed, the absorptivity, and the emissivity are equal for any source. In wavelength units, the spectral radiance is

$$\boldsymbol{L}_\lambda = \frac{\epsilon[2h(c/\mathrm{n})^2]}{\lambda^5(e^{hc/\lambda kT} - 1)}. \tag{1.5}$$

It can be easily shown from equations (1.4) and (1.5) that the spectral radiances are related as follows:

$$\boldsymbol{L}_\lambda = \left(\frac{c}{\lambda^2}\right)\boldsymbol{L}_\nu = \left(\frac{\nu}{\lambda}\right)\boldsymbol{L}_\nu. \tag{1.6}$$

According to the Stefan–Boltzmann law, the radiant exitance for a blackbody becomes:

$$\boldsymbol{M} = \pi \int_0^\infty \boldsymbol{L}_\nu \, d\nu = \frac{2\pi k^4 T^4}{c^2 h^3} \int_0^\infty \frac{x^3}{e^x - 1}\, dx = \frac{2\pi^5 k^4}{15 c^2 h^3} T^4 = \sigma T^4, \tag{1.7}$$

where σ ($= 5.67 \times 10^{-8}$ W m^{-2} K^{-4}) is the Stefan–Boltzmann constant.

For Lambertian sources, the optical system feeding a detector will receive a portion of the source power that is determined by a number of geometric factors as illustrated in Figure 1.2. The system will accept radiation from only a limited range of directions determined by the geometry of the optical system as a whole and known as the "field of view". The area of the source that is effective in producing a signal is determined by the field of view and the distance between the optical system and the source (or by the size of the source if it all lies within the field of view). This area will emit radiation with some angular dependence. Only the radiation that is emitted in directions where it is intercepted by the optical system can be detected. The range of directions accepted is determined by the solid angle, Ω, that the entrance aperture of the optical system subtends as viewed from the source. Assume that none of the

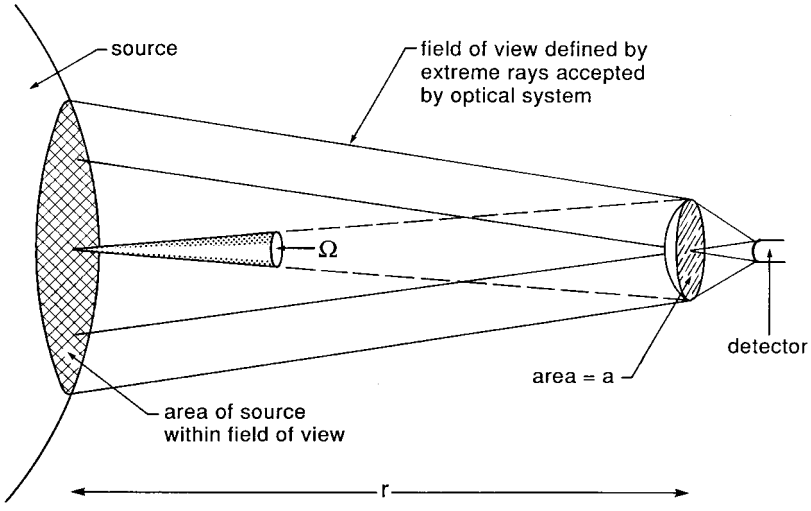

Figure 1.2. Geometry for computing power received by a detector system.

emitted power is absorbed or scattered before it reaches the optical system. The power this system receives is then the radiance in its direction multiplied by (1) the source area within the system field of view times (2) the solid angle subtended by the optical system as viewed from the source.

Although a general treatment must allow for the field of view to include only a portion of the source, in many cases of interest the entire source lies within the field of view, so the full projected area of the source is used. For a spherical source of radius R, this area is πR^2. The solid angle subtended by the detector system is

$$\Omega = \frac{a}{r^2}, \tag{1.8}$$

where a is the area of the entrance aperture of the system (strictly speaking, a is the projected area; we have assumed the system is pointing directly at the source) and r is its distance from the source. For a circular aperture,

$$\Omega = 4\pi \sin^2(\theta/2), \tag{1.9}$$

where θ is the half-angle of the right-circular cone whose base is the detector system entrance aperture, and whose vertex lies on a point on the surface of the source; r is the height of this cone.

It is particularly useful when the angular diameter of the source is small compared with the field of view of the detector system to consider the irradiance, $\boldsymbol{E}$. It is the power in watts per square meter received at a unit surface element at some distance from the source. For the case described in the preceding paragraph, the irradiance is obtained by first multiplying the radiant exitance by the total surface

area of the source, A, to get the flux, $A\pi \boldsymbol{L}$. The flux is then divided by the area of a sphere of radius r centered on the source to give

$$\boldsymbol{E} = \frac{A\boldsymbol{L}}{4r^2}, \tag{1.10}$$

where r is the distance of the source from the irradiated surface element on the sphere. The spectral irradiance, $\boldsymbol{E}_\nu$ or $\boldsymbol{E}_\lambda$, is the irradiance per unit frequency or wavelength interval. It is also sometimes called the flux density, and is a very commonly used description of the power received from a source. It can be obtained from equation (1.10) by substituting $\boldsymbol{L}_\nu$ or $\boldsymbol{L}_\lambda$ for $\boldsymbol{L}$.

The radiometric quantities discussed above are summarized in Table 1.1. Equations are provided for illustration only; in some cases, these examples apply only to specific circumstances. The terminology and symbolism vary substantially from one discipline to another; for example, the last two columns of the table translate some of the commonly used radiometric terms into astronomical nomenclature.

Only a portion of the power received by the optical system is passed on to the detector. The system will have inefficiencies due to both absorption and scattering of energy in its elements, and because of optical aberrations and diffraction. These effects can be combined into a system transmittance term. In addition, the range of frequencies or wavelengths to which the system is sensitive (the spectral bandwidth of the system) is usually restricted by a combination of characteristics of the detector, filters, and other elements of the system as well as by any spectral dependence of the transmittance of the optical path from the source to the entrance aperture. A rigorous accounting of the spectral response requires that the spectral radiance of the source be multiplied by the spectral transmittances of all the spectrally active elements in the optical path to the detector, and by the detector spectral response. The resulting function must be integrated over frequency or wavelength to determine the total power effective in generating a signal.

In many cases, the spectral response is intentionally restricted to a narrow range of wavelengths by placing a bandpass optical filter in the beam. It is then useful to define the effective wavelength of the system as

$$\lambda_0 = \frac{\displaystyle\int_0^\infty \lambda\, \mathcal{T}(\lambda)\, d\lambda}{\displaystyle\int_0^\infty \mathcal{T}(\lambda)\, d\lambda}, \tag{1.11}$$

where $\mathcal{T}(\lambda)$ is the spectral transmittance of the system, that is, the fraction of incident light transmitted as a function of wavelength. Often the spectral variations of the other transmittance terms can be ignored over the restricted spectral range of the filter. The

Table 1.1 *Definitions of radiometric quantities*

Symbol	Name	Definition	Units	Equation	Alternate name	Alternate symbol
$\boldsymbol{L}_\nu$	Spectral radiance (frequency units)	Power leaving unit projected surface area into unit solid angle and unit frequency interval	$\mathrm{W\,m^{-2}\,Hz^{-1}\,ster^{-1}}$	(1.4)	Specific intensity (frequency units)	$\boldsymbol{I}_\nu$
$\boldsymbol{L}_\lambda$	Spectral radiance (wavelength units)	Power leaving unit projected surface area into unit solid angle and unit wavelength interval	$\mathrm{W\,m^{-3}\,ster^{-1}}$	(1.5)	Specific intensity (wavelength units)	$\boldsymbol{I}_\lambda$
$\boldsymbol{L}$	Radiance	Spectral radiance integrated over frequency or wavelength	$\mathrm{W\,m^{-2}\,ster^{-1}}$	$\boldsymbol{L} = \int \boldsymbol{L}_\nu d\nu$	Intensity or specific intensity	$\boldsymbol{I}$
$\boldsymbol{M}$	Radiant exitance	Power emitted per unit surface area	$\mathrm{W\,m^{-2}}$	$\boldsymbol{M} = \int \boldsymbol{L}(\theta) d\Omega$		
Φ	Flux	Total power emitted by source of area A	W	$\Phi = \int \boldsymbol{M} dA$	Luminosity	$\boldsymbol{L}$
$\boldsymbol{E}$	Irradiance	Power received at unit surface element; equation applies well removed from the source at distance r	$\mathrm{W\,m^{-2}}$	$\boldsymbol{E} = \frac{\int \boldsymbol{M} dA}{(4\pi r^2)}$		
$\boldsymbol{E}_\nu$, $\boldsymbol{E}_\lambda$	Spectral irradiance	Power received at unit surface element per unit frequency or wavelength interval	$\mathrm{W\,m^{-2}\,Hz^{-1}}$, $\mathrm{W\,m^{-3}}$		Flux density	$\boldsymbol{S}_\nu$, $\boldsymbol{S}_\lambda$

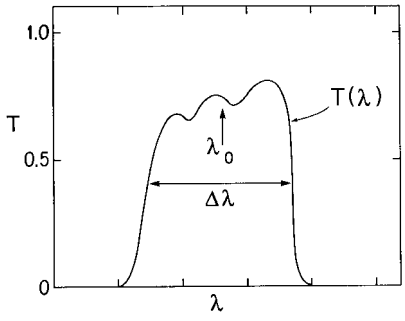

Figure 1.3. Transmittance function $T(\lambda)$ of a filter. The FWHM $\Delta\lambda$ and the effective wavelength λ_0 are indicated.

bandpass of the filter, $\Delta\lambda$, can be taken to be the full width at half maximum (FWHM) of its transmittance function (see Figure 1.3). If the filter cuts on and off sharply, its transmittance can be approximated as the average value over the range $\Delta\lambda$:

$$\mathcal{T}_F = \frac{\int_{\Delta\lambda} \mathcal{T}(\lambda)\, d\lambda}{\Delta\lambda}. \tag{1.12}$$

If $\Delta\lambda/\lambda_0 \leq 0.2$ and the filter cuts on and off sharply, the power effective in generating a signal can usually be estimated in a simplified manner. The behavior of the bandpass filter can be approximated by taking the spectral radiance at λ_0 (in wavelength units) and multiplying it by $\Delta\lambda$ and the average filter transmittance over the range $\Delta\lambda$. The result is multiplied by the various geometric and transmittance terms already discussed for the remainder of the system. However, if λ_0 is substantially shorter than the peak wavelength of the blackbody curve (that is, one is operating in the Wien region of the blackbody) or if there is sharp spectral structure within the passband, then this approximation can lead to significant errors, particularly if $\Delta\lambda/\lambda_0$ is relatively large.

Continuing with the approximation just discussed, we can derive a useful expression for estimating the power falling on the detector:

$$P_d \approx \frac{A_{\text{proj}}\, a\, \mathcal{T}_P(\lambda_0)\mathcal{T}_O(\lambda_0)\mathcal{T}_F\, \boldsymbol{L}_\lambda(\lambda_0)\Delta\lambda}{r^2}. \tag{1.13}$$

Here A_{proj} is the area of the source projected onto the plane perpendicular to the line of sight from the source to the optical receiver. $\mathcal{T}_P$, $\mathcal{T}_O$, and $\mathcal{T}_F$ are the transmittances, respectively, of the optical path from the source to the receiver, of the receiver optics (excluding the bandpass filter), and of the bandpass filter. The area of the receiver entrance aperture is a, and the distance of the receiver from the source is r. An analogous expression holds in frequency units. The major underlying assumptions for equation (1.13) are that: (a) the field of view of the receiver includes the entire source; (b) the source is a Lambertian emitter; and (c) the spectral response of the

detector is limited by a filter with a narrow or moderate bandpass that is sharply defined.

1.2 Detector types

Nearly all detectors act as transducers that receive photons and produce an electrical response that can be amplified and converted into a form intelligible to suitably conditioned human beings. There are three basic ways that detectors carry out this function:

(a) *Photon detectors* respond directly to individual photons. An absorbed photon releases one or more bound charge carriers in the detector that may (1) modulate the electric current in the material; (2) move directly to an output amplifier; or (3) lead to a chemical change. Photon detectors are used throughout the X-ray, ultraviolet, visible, and infrared spectral regions. Examples that we will discuss are photoconductors (Chapters 2 and 3), photodiodes (Chapter 4), photoemissive detectors (Chapter 7), and photographic plates (Chapter 8).

(b) *Thermal detectors* absorb photons and thermalize their energy. In most cases, this energy changes the electrical properties of the detector material, resulting in a modulation of the electrical current passing through it. Thermal detectors have a very broad and nonspecific spectral response, but they are particularly important at infrared and submillimeter wavelengths, and as X-ray detectors. Bolometers and other thermal detectors will be discussed in Chapter 9.

(c) *Coherent receivers* respond to the electric field strength of the signal and can preserve phase information about the incoming photons. They operate by interference of the electric field of the incident photon with the electric field from a coherent local oscillator. These devices are primarily used in the radio and submillimeter regions and are sometimes useful in the infrared. Coherent receivers for the infrared are discussed in Chapter 10, and those for the submillimeter are discussed in Chapter 11.

1.3 Performance characteristics

Good detectors preserve a large proportion of the information contained in the incoming stream of photons. A variety of parameters are relevant to this goal:

(a) *Spectral response* – the total wavelength or frequency range over which photons can be detected with reasonable efficiency.

(b) *Spectral bandwidth* – the wavelength or frequency range over which photons are detected at any one time; some detectors can operate in one or more bands placed within a broader range of spectral response.

(c) *Linearity* – the degree to which the output signal is proportional to the number of incoming photons that was received to produce the signal.

(d) *Dynamic range* – the maximum variation in signal over which the detector output represents the photon flux without losing significant amounts of information.

(e) *Quantum efficiency* – the fraction of the incoming photon stream that is converted into signal.

(f) *Noise* – the uncertainty in the output signal. Ideally, the noise consists only of statistical fluctuations due to the finite number of photons producing the signal.

(g) *Imaging properties* – the number of detectors ("pixels") in an array determines in principle how many picture elements the detector can record simultaneously. Because signal may blend from one detector to adjacent ones, the resolution that can be realized may be less, however, than indicated just by the pixel count.

(h) *Time response* – the minimum interval of time over which the detector can distinguish changes in the photon arrival rate.

The first two items in this listing should be clear from our discussion of radiometry, and the next two are more or less self-explanatory. However, the remaining entries include subtleties that call for more discussion.

1.3.1 Quantum efficiency

To be detected, photons must be absorbed. The absorption coefficient in the detector material is indicated as $a(\lambda)$ and conventionally has units of cm^{-1}. The absorption length is just the inverse of the absorption coefficient. The absorption of a flux of photons, φ, passing through a differential thickness element dl is expressed by

$$\frac{d\varphi}{dl} = -a(\lambda)\,\varphi, \tag{1.14}$$

with the solution for the remaining flux at depth l being

$$\varphi = \varphi_0\, e^{-a(\lambda)l}, \tag{1.15}$$

where φ_0 is the flux entering the detector. The quantum efficiency, η, is the flux absorbed in the detector divided by the total flux incident on its surface. There are two components: (1) the portion of photons that enter the detector that are absorbed within it; and (2) the portion of photons incident on the detector that actually enter it. The portion of the flux absorbed within the detector divided by the flux that enters it is

$$\eta_{\mathrm{ab}} = \frac{\varphi_0 - \varphi_0\, e^{-a(\lambda)d_1}}{\varphi_0} = 1 - e^{-a(\lambda)d_1}, \tag{1.16}$$

where d_1 is the thickness of the detector. The quantity η_{ab} is known as the absorption factor. Photons are lost by reflection from the surface before they enter the detector volume, leading to a reduction in quantum efficiency below η_{ab}. Minimal reflection occurs for photons striking at normal incidence:

$$\mathcal{R} = \frac{(\mathrm{n}-1)^2 + (a(\lambda)\lambda/4\pi)^2}{(\mathrm{n}+1)^2 + (a(\lambda)\lambda/4\pi)^2}, \tag{1.17}$$

where the reflectivity, $\mathcal{R}$, is the fraction of the incident flux of photons that is reflected, n is the refractive index of the material (= c/(the speed of light in the material)), $a(\lambda)$ is the absorption coefficient at wavelength λ, and we have assumed that the photon is incident from air or vacuum, which have a refractive index of $\mathrm{n} \approx 1$. In most circumstances of interest for detectors, the absorption coefficients are small enough that the terms involving them can be ignored. Reflection from the back of the detector can result in absorption of photons that would otherwise escape. If we ignore this potential gain, the net quantum efficiency is

$$\eta = (1 - \mathcal{R})\,\eta_{ab}. \tag{1.18}$$

For example, for a detector operating at a wavelength of 0.83 μm that is 20 μm thick and made of material with $\mathrm{n} = 3.5$ and $a(0.83\,\mu\mathrm{m}) = 1000\ \mathrm{cm}^{-1}$, $\eta = (1 - 0.31) \times (1 - 0.13) = 0.60$.

1.3.2 Noise and signal to noise

The following discussion derives the inherent ratio of signal, S, to noise, N, in the incoming photon stream and then compares it with what can be achieved in the detector as a function of the quantum efficiency. Ignoring minor corrections having to do with the quantum nature of photons, it can be assumed that the input photon flux follows Poisson statistics,

$$P(m) = \frac{e^{-n} n^m}{m!}, \tag{1.19}$$

where $P(m)$ is the probability of detecting m photons in a given time interval, and n is the average number of photons detected in this time interval if a large number of detection experiments is conducted. The root-mean-square noise N_{rms} in a number of independent events each with expected noise N is the square root of the mean, n,

$$N_{rms} = \langle N^2 \rangle^{1/2} = n^{1/2}. \tag{1.20}$$

The errors in the detected number of photons in two experiments can usually be taken to be independent, and hence they add quadratically. That is, the noise in two measurements, n_1 and n_2, is

$$N_{rms} = \langle N^2 \rangle^{1/2} = \left[\left(n_1^{1/2}\right)^2 + \left(n_2^{1/2}\right)^2 \right]^{1/2} = (n_1 + n_2)^{1/2}. \tag{1.21}$$

From the above discussion, the signal-to-noise ratio for Poisson-distributed events is $n/n^{1/2}$, or

$$S/N = n^{1/2}. \tag{1.22}$$

This result can be taken to be a measure of the information content of the incoming photon stream as well as a measure of the confidence that a real signal has been detected.[1] †

From the standpoint of the detector, photons that are not absorbed cannot contribute to either signal or noise; they might as well not exist. Consequently, for n photons incident on the detector, equation (1.22) shows that the signal-to-noise ratio goes as $\eta n/(\eta n)^{1/2}$, or

$$\left(\frac{S}{N}\right)_{\mathrm{d}} = (\eta n)^{1/2} \tag{1.23}$$

in the ideal case where both signal and noise are determined only by the photon statistics.

The quantum efficiency defined in equation (1.18) refers only to the fraction of incoming photons converted into a signal in the first stage of detector action. Ideally, the signal-to-noise ratio attained in a measurement is controlled entirely by the number of photons absorbed in the first stage. However, additional steps in the detection process can degrade the information present in the photon stream absorbed by the detector, either by losing signal or by adding noise. The detective quantum efficiency (*DQE*) describes this degradation succinctly. We take n_{equiv} to be the number of photons that would be required with a perfect detector (100% quantum efficiency, no further degradation) to produce an output equivalent in signal to noise to that produced with the real detector from n_{in} received photons. We define

$$DQE = \frac{n_{\mathrm{equiv}}}{n_{\mathrm{in}}} = \frac{(S/N)^2_{\mathrm{out}}}{(S/N)^2_{\mathrm{in}}}. \tag{1.24}$$

Converting to signal to noise, $(S/N)_{\mathrm{out}}$ is the observed signal-to-noise ratio, while $(S/N)_{\mathrm{in}}$ is the potential signal-to-noise ratio of the incoming photon stream, as given by equation (1.22). By substituting equations (1.22) and (1.23) into equation (1.24), it is easily shown that the *DQE* is just the quantum efficiency defined in equation (1.18) if there is no subsequent degradation of the signal to noise.

1.3.3 Imaging properties

The resolution of an array of detectors can be most simply measured by exposing it to a pattern of alternating white and black lines and determining the minimum spacing of line pairs that can be distinguished. The eye can identify such a pattern if the light–dark variation is 4% or greater. The resolution of the detector array is

† Superscript numbers refer to Notes at end of chapter

expressed in line pairs per millimeter corresponding to the highest density of lines that produces a pattern at this threshold.

Although it is relatively easy to measure resolution in this way for the detector array alone, a resolution in line pairs per millimeter is difficult to combine with resolution estimates for other components in an optical system used with it. For example, how would one derive the net resolution for a camera with a lens and photographic film whose resolutions are both given in line pairs per millimeter? A second shortcoming is that the performance in different situations can be poorly represented by the line pairs per millimeter specification. For example, one might have two lenses, one of which puts 20% of the light into a sharply defined image core and spreads the remaining 80% widely, whereas the second puts all the light into a slightly less well-defined core. These systems might achieve identical resolutions in line pairs per millimeter (which requires only 4% modulation), yet they would perform quite differently in other situations.

A more general concept is the modulation transfer function, or *MTF*. Imagine that the detector array is exposed to a sinusoidal input signal of period P and amplitude $F(x)$,

$$F(x) = a_0 + a_1 \sin(2\pi f x), \tag{1.25}$$

where $f = 1/P$ is the spatial frequency, x is the distance along one axis of the array, a_0 is the mean height (above zero) of the pattern, and a_1 is its amplitude. These terms are indicated in Figure 1.4(a). The modulation of this signal is defined as

$$M_{\text{in}} = \frac{F_{\max} - F_{\min}}{F_{\max} + F_{\min}} = \frac{a_1}{a_0}, \tag{1.26}$$

where $F_{\max}$ and $F_{\min}$ are the maximum and minimum values of $F(x)$. Assuming that the resulting image output from the detector is also sinusoidal (which may be only approximately true due to nonlinearities), it can be represented by

$$G(x) = b_0 + b_1(f) \sin(2\pi f x), \tag{1.27}$$

where x and f are the same as in equation (1.25), and b_0 and $b_1(f)$ are analogous to a_0 and a_1. Because of the limited response of the array to high spatial frequencies, the signal amplitude, b_1, is a function of f. The modulation in the image will be

$$M_{\text{out}} = \frac{b_1(f)}{b_0} \leq M_{\text{in}}. \tag{1.28}$$

The modulation transfer factor is

$$MT = \frac{M_{\text{out}}}{M_{\text{in}}}. \tag{1.29}$$

A separate value of the *MT* will apply at each spatial frequency; Figure 1.4(a) illustrates an input signal that contains a range of spatial frequencies, and Figure 1.4(b) shows a corresponding output in which the modulation decreases with increasing

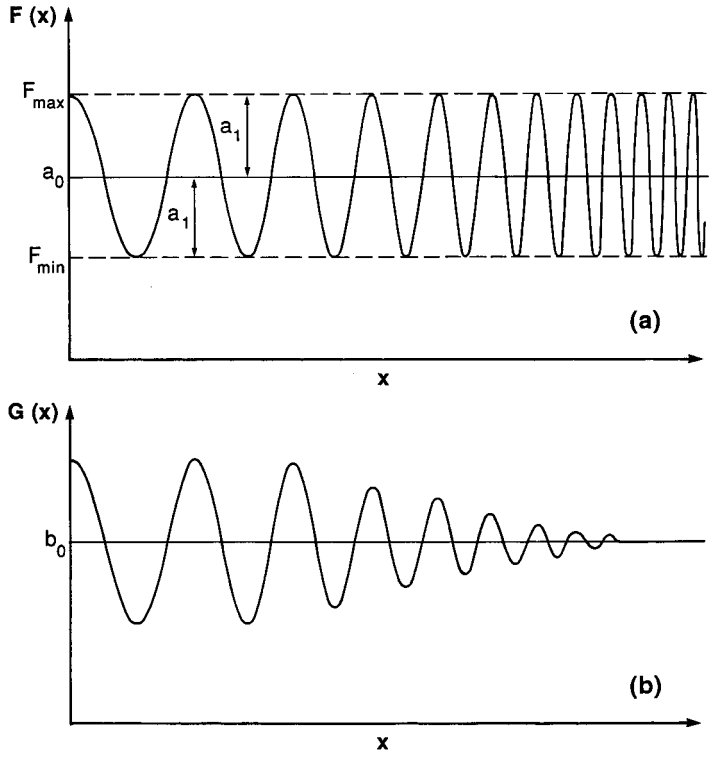

Figure 1.4. Illustration of variation of modulation with spatial frequency.
(a) Sinusoidal input signal of constant amplitude but varying spatial frequency.
(b) How an imaging detector system might respond to this signal.

spatial frequency. This frequency dependence of the *MT* is expressed in the modulation transfer function (*MTF*). Figure 1.5 shows the *MTF* corresponding to the response of Figure 1.4(b).

In principle, the *MTF* provides a virtually complete specification of the imaging properties of a detector array. However, one must be aware that the *MTF* may vary over the face of the array and may have color dependence. It also cannot represent nonlinear effects such as saturation on bright objects. In addition, the *MTF* omits time-dependent imaging properties, such as latent images that may persist after the image of a bright source has been put on the array and removed.

Computationally, the *MTF* can be determined by taking the absolute value of the Fourier transform, $\mathbf{F}(u)$, of the image of a perfect point source. This image is called the point spread function. Fourier transformation is the general mathematical technique used to determine the frequency components of a function $f(x)$ (see, for example, Press *et al.*, 1986; Bracewell, 2000). $\mathbf{F}(u)$ is defined as

$$\mathbf{F}(u) = \int_{-\infty}^{\infty} f(x) e^{j2\pi ux} dx, \tag{1.30}$$

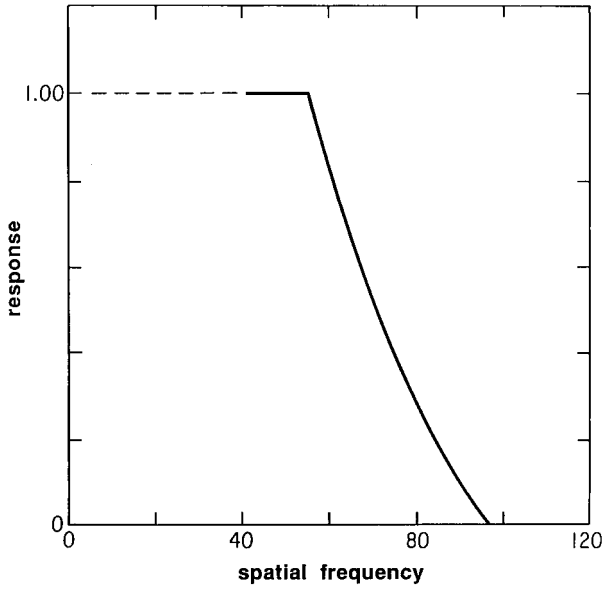

Figure 1.5. The modulation transfer function (*MTF*) for the response illustrated in Figure 1.4(b).

with inverse

$$f(x) = \int_{-\infty}^{\infty} \mathbf{F}(u) e^{-j2\pi xu} du, \tag{1.31}$$

where j is the (imaginary) square root of -1. The Fourier transform can be generalized in a straightforward way to two dimensions, but for the sake of simplicity we will not do so here. The absolute value of the transform is

$$|\mathbf{F}(u)| = [\mathbf{F}(u)\mathbf{F}^*(u)]^{1/2}, \tag{1.32}$$

where $\mathbf{F}^*(u)$ is the complex conjugate of $\mathbf{F}(u)$; it is obtained by reversing the sign of all imaginary terms in $\mathbf{F}(u)$.

If $f(x)$ represents the point spread function, $|\mathbf{F}(u)|/|\mathbf{F}(0)|$ is the *MTF* with u the spatial frequency. This formulation holds because a sharp impulse, represented mathematically by a δ function, contains all frequencies equally (that is, its Fourier transform $C\int \delta(x) e^{(-j,2\pi ux)}\, dx = C$, a constant). Hence the Fourier transform of the image formed from an input sharp impulse (the image is the point spread function) gives the spatial frequency response of the detector.

The *MTF* is normalized to unity at spatial frequency 0 by this definition. As emphasized in Figure 1.5, the response at zero frequency cannot be measured directly but must be extrapolated from higher frequencies.

Table 1.2 *Fourier Transforms*

$f(x)$	$F(u)$
$F(x)$	$f(-u)$
$aF(x)$	$aF(u)$
$f(ax)$	$(1/\|a\|)F(u/a)$
$f(x)+g(x)$	$F(u)+G(u)$
1	$\delta(u)$[c]
$e^{-\pi x^2}$	$e^{-\pi u^2}$
$e^{-\|x\|}$	$2/(1+(2\pi u)^2)$
e^{-x}, $x>0$	$(1-j\,2\pi u)/(1+(2\pi u)^2)$
$\operatorname{sech}(\pi x)$	$\operatorname{sech}(\pi u)$
$\|x\|^{-1/2}$	$\|u\|^{-1/2}$
$\operatorname{sgn}(x)$[a]	$-j/(\pi u)$
$e^{-\|x\|}\operatorname{sgn}(x)$	$-j\,4\pi u/(1+(2\pi u)^2)$
$\Pi(x)$[b]	$\sin(\pi u)/\pi u$

[a] $\operatorname{sgn}(x)=-1$ for $x<0$ and $=1$ for $x\geq 0$.
[b] $\Pi(x)=1$ for $|x|<1/2$ and $=0$ otherwise.
[c] $\delta(u)=0$ for $u\neq 0$, $\int\delta(u)\,\mathrm{d}u=1$; that is, $\delta(u)$ is a spike at $u=0$.

Only a relatively small number of functions have Fourier transforms that are easy to manipulate. Table 1.2 contains a short compilation of some of these cases. With the use of computers, however, Fourier transformation is a powerful and very general technique.

The image of an entire linear optical system is the convolution of the images from each element. By the "convolution theorem", its *MTF* can be determined by multiplying together the *MTF*s of its constituent elements, and the resulting image is determined by inverse transforming the *MTF*. The multiplication occurs on a frequency by frequency basis, that is, if the first system has $MTF_1(f)$ and the second $MTF_2(f)$, the combined system has $MTF(f) = MTF_1(f)\,MTF_2(f)$. The overall resolution capability of complex optical systems can be more easily determined in this way than by brute force image convolution.

1.3.5 Frequency response

The response speed of a detector can be described very generally by specifying the dependence of its output on the frequency of an imaginary photon signal that varies sinusoidally in time. This concept is analogous to the modulation transfer function described just above with regard to imaging; in this case it is called the electrical frequency response of the detector.

A variety of factors limit the frequency response. Many of them, however, can be described by an exponential time response, such as that of a resistor/capacitor electrical circuit. To be specific in the following, we will assume that the response is given by the *RC* time constant of such a circuit, although we will find other uses for the identical formalism later. If the capacitor is in parallel with the resistance, charge deposited on the capacitance bleeds off through the resistance with an exponential time constant

$$\tau_{RC} = RC. \tag{1.33}$$

Sometimes a "rise time" or "fall time" is specified rather than the exponential time constant. The rise or fall time is the interval required for the output to change from 10% to 90% of its final value or vice versa (measured relative to the initial value). For an exponential response, this time is $2.20\tau_{RC}$.

Let a voltage impulse be deposited on the capacitor,

$$v_{\text{in}}(t) = v_0\delta(t), \tag{1.34}$$

where v_0 is a constant and $\delta(t)$ is the delta function (defined in the footnote to Table 1.2). We can observe this event in two ways. First, we might observe the voltage across the resistance and capacitance directly, for example with an oscilloscope. It will have the form

$$v_{\text{out}}(t) = \begin{cases} 0, & t < 0 \\ \dfrac{v_0}{\tau_{RC}} e^{-t/\tau_{RC}}, & t \geq 0. \end{cases} \tag{1.35}$$

The same event can be analyzed in terms of the effect of the circuit on the input frequencies rather than on the time dependence of the voltage. To do so, we convert the input and output voltages to frequency spectra by taking their Fourier transforms. The delta function contains all frequencies at equal strength, that is, from Table 1.2,

$$V_{\text{in}}(f) = v_0 \int_{-\infty}^{\infty} \delta(t) e^{-j2\pi f t} dt = v_0. \tag{1.36}$$

Since the frequency spectrum of the input is flat ($V_{\text{in}}(f)$ = constant), any deviations from a flat spectrum in the output must arise from the action of the circuit. That is, the output spectrum gives the frequency response of the circuit directly. Again from Table 1.2, it is

$$V_{\text{out}}(f) = \int_{-\infty}^{\infty} v_{\text{out}}(t) e^{-j2\pi f t} dt = v_0 \left[\frac{1 - j2\pi f \tau_{RC}}{1 + (2\pi f \tau_{RC})^2} \right]. \tag{1.37}$$

The imaginary part of $V_{\text{out}}(f)$ represents phase shifts that can occur in the circuit. For a simple discussion, we can ignore the phase and describe the strength of the signal only in terms of the frequency dependence of its amplitude. The amplitude

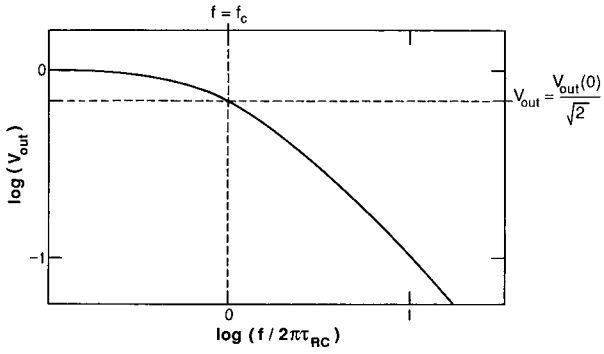

Figure 1.6. Frequency response of an *RC* circuit. The cutoff frequency is also indicated.

can be determined by taking the absolute value of $V_{\text{out}}(f)$:

$$|V_{\text{out}}(f)| = (V_{\text{out}} V_{\text{out}}^{*})^{1/2} = \frac{v_0}{[1 + (2\pi f \tau_{RC})^2]^{1/2}}, \tag{1.38}$$

where V_{out}^{*} is the complex conjugate of V_{out}. This function is plotted in Figure 1.6. As with the *MTF*, the effects of different circuit elements on the overall frequency response can be determined by multiplying their individual response functions together.

The frequency response is often characterized by a cutoff frequency

$$f_{\text{c}} = \frac{1}{2\pi \tau_{RC}}, \tag{1.39}$$

at which the amplitude drops to $1/\sqrt{2}$ of its value at $f = 0$, or

$$|V_{\text{out}}(f_{\text{c}})| = \frac{1}{\sqrt{2}} |V_{\text{out}}(0)|. \tag{1.40}$$

1.4 Solid state physics

The electrical properties of a semiconductor are altered dramatically by photoexcitation due to the absorption of an ultraviolet, visible, or infrared photon, making this class of material well adapted to a variety of photon detection strategies. Metals, on the other hand, have high electrical conductivity that is only insignificantly modified by the absorption of photons, and insulators require more energy to excite electrical changes than is available from individual visible or infrared photons.

In addition, adding small amounts of impurities to semiconductors can strongly modify their electrical properties at and below room temperature. Consequently, semiconductors are the basis for most electronic devices, including those used for amplification of photoexcited currents as well as those used to detect photons with too little energy to be detected through photoexcitation.

Because of these properties of semiconductors, virtually every detector we shall discuss depends on these materials for its operation. To facilitate our discussion, we will first review some of the properties of semiconductors. The concepts introduced below are used throughout the remaining chapters.

The elemental semiconductors are silicon and germanium; they are found in column IVa of the periodic table (Table 1.3). Their outermost electron shells, or valence states, contain four electrons, half of the total number allowed for these shells. They form crystals with a diamond lattice structure (note that carbon is also in column IVa). In this structure, each atom bonds to its four nearest neighbors; it can therefore share one valence electron with each neighbor, and vice versa. Electrons are fermions and must obey the Pauli exclusion principle, which states that no two particles with half-integral quantum mechanical spin can occupy identical quantum states.[2] Because of the exclusion principle, the electrons shared between neighboring nuclei must have opposite spin (if they had the same spin, they would be identical quantum mechanically), which accounts for the fact that they occur in pairs. By sharing electrons, each atom comes closer to having a filled valence shell, and a quantum mechanical binding force known as a covalent bond is created.

The binding of electrons to an atomic nucleus can be described in terms of a potential energy "well" around the nucleus. Electrons may be in the ground state or at various higher energy levels called excited states. There is a specific energy difference between these states which can be measured by detecting an absorption or emission line when an electron shifts between energy levels. The sharply defined energy levels of an isolated atom occur because of constructive interference of electron wave functions within the potential well; there is destructive interference at all other energies. When atoms are brought close enough together to allow the electron wave functions to begin overlapping, the energy levels of the individual atoms split due to the coupling between the potential wells. The splitting occurs because the electrons must distribute themselves so that no two of them are in an identical quantum state, according to the exclusion principle. In a compact structure such as a crystal, the energy levels split multiply into broad energy zones called bands. The "valence states" and "conduction states" in a material are analogous to the ground state and excited states, respectively, in an isolated atom. Band diagrams such as those in Figure 1.7 can represent this situation.

For any material at a temperature of absolute zero, all available states in the band would be filled up to some maximum level. The electrical conductivity would be zero because there would be no accessible states into which electrons could move. Conduction becomes possible when electrons are lifted into higher and incompletely filled energy levels, either by thermal excitation or by other means. There are two distinct possibilities. In a metal, the electrons only partially fill a band so that a very small amount of energy (say, a temperature just above absolute zero) is required to gain access to unfilled energy levels and hence to excite conductivity. Metal atoms

Table 1.3 *Periodic table of the elements*

Ia	IIa	III b	IV b	Vb	VIb	VII b	VIII			Ib	IIb	III a	IV a	V a	VI a	VII a	0
1 H													↓				2 He
3 Li	4 Be											5 B	6 **C**	7 N	8 O	9 F	10 Ne
11 Na	12 Mg											13 **Al**	14 **Si**	15 **P**	16 **S**	17 **Cl**	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 **Zn**	31 **Ga**	32 **Ge**	33 **As**	34 **Se**	35 **Br**	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 **Ag**	48 **Cd**	49 **In**	50 **Sn**	51 **Sb**	52 **Te**	53 **I**	54 Xe
55 Cs	56 Ba	57 La	72 Hf	73 Ta	74 W	75 Re	76 Os	77 ir	78 Pt	79 Au	80 **Hg**	81 Tl	82 **Pb**	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	89 Ac															

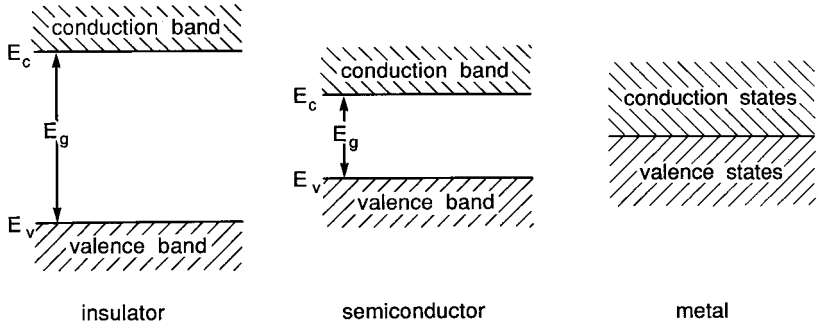

Figure 1.7. Energy band diagrams for insulators, semiconductors, and metals.

have a small number of loosely bound, outer-shell electrons that are easily given up to form ions. In a bulk metal, these electrons are contributed to the crystal as a whole, creating a structure of positive ions immersed in a sea of free electrons. This situation produces metallic bonding.

On the other hand, in a semiconductor or an insulator, the electrons would completely fill a band at absolute zero. To gain access to unfilled levels, an electron must be lifted into a level in the next higher band, resulting in a threshold excitation energy required to initiate electrical conductivity. In this latter case, the filled band is called the valence band and the unfilled one the conduction band. The bandgap energy, E_g, is the energy between the highest energy level in the valence band, E_v, and the lowest energy level in the conduction band, E_c. It is the minimum energy that must be supplied to excite conductivity in the material. Semiconductors have $0 < E_g < 3.5$ eV.

The band diagrams for insulators and semiconductors are similar to each other, but the insulators have larger values of E_g because the conduction electrons are more tightly bound to the atoms than they are in semiconductors. It therefore takes more energy to break these bonds in insulators so the electrons can move through the material. A common kind of insulator is a compound containing atoms from opposite ends of the periodic table (one example is NaCl). In this case, the valence electron is taken from the metal atom and added to the outer valence band of the halide atom; both atoms then have filled outer electron shells. The electrostatic attraction of the positive metal and negative halide ions forms the crystal bond. This bonding is called ionic.[3]

Despite their differing electrical behavior, the band diagrams for semiconductors and insulators are qualitatively similar. Semiconductors are partially conducting under typically encountered conditions because the thermal excitation at room temperature is adequate to lift some electrons across their modest energy bandgaps. However, their conductivity is a strong function of temperature (going roughly as $e^{-E_g/2kT}$; $kT \approx 0.025$ eV at room temperature), and near absolute zero they behave as insulators. In such a situation, the charge carriers are said to be “frozen out”.