CHAPTER 3

Optical Detection

You can't hit what you can't see.

—Walter Johnson (American baseball player)

3.1 INTRODUCTION

Electro-optical systems normally consist of an optical front end and an electronic and software back end, with an optical detector occupying the uncomfortable place of honor in between. The detector subsystem consists not only of the detector itself, but includes any baffles, coolers, windows, or optical filters associated with the detector, as well as amplifiers, electrical filters, and analog signal processing taking place in its vicinity. The optical front end is normally the most expensive part of the system, but is the only place where its photon throughput (and hence the maximum SNR) can be improved; the electronic back end is where the filtering, digitizing, and postprocessing occur, and is what produces the valuable output.

The guiding principle of detector system design is this: a photon, once lost, cannot be recovered. This includes those lost due to poor coatings, inefficient detectors, or poor matching of the optical characteristics of the incoming light to those of the detector, as well as those that are needlessly swamped in technical noise due to a poor choice of amplifier, load impedance, or circuit topology.

Once the measurement principle has been chosen, the choice of detector and the design of the detector subsystem are usually the most critical tasks in engineering a high performance electro-optical instrument; it is easy to get these badly wrong, and even serious errors are not always immediately obvious. Vigilant attention to every decibel there will be repaid with a sensitive, stable, repeatable measurement system. Decibel-chasing should not be limited to sensitivity alone, but should include stability as well; if efficiency is purchased at the price of instability, a measurement that was once merely slow may become impossible. Careful attention must be paid to many second-order sources of artifacts, such as ground loops, spectral response changes with temperature, etalon fringes, parametric effects such as memory in photoconductors, and nonlinear effects such as overload in photomultipliers or Debye shielding and lateral voltage drops in photodiodes. Achieving high stability is a somewhat subtle task and requires the cultivation of a certain healthy paranoia about unconsidered effects. Nevertheless, it is quite possible

Building Electro-Optical Systems, Making It All Work, Second Edition, By Philip C. D. Hobbs

for a measurement system to achieve stabilities of 10^{-5} to 10^{-6} in 1 hour, even at DC, if good design practices are followed.

Linearity is often the most important parameter after efficiency and stability. Many types of photodetector are extremely linear, as well as time invariant; this good performance can only be degraded by the succeeding circuitry. Thus another useful maxim is: if there's one operation in your signal processing strategy that has to be really accurate, do it right at the detector. Examples of this sort of operation are subtracting two signals, where very small phase shifts and amplitude imbalances matter a great deal, or amplifying very small signals in a noisy environment.

Detectors differ also in their resistance to adverse conditions. Photodiodes tend to be vulnerable to ultraviolet damage, photomultipliers to shock, APDs to overvoltage. (PDs are bulletproof by comparison, of course.)

3.2 PHOTODETECTION IN SEMICONDUCTORS

A semiconductor *PN junction* is the interface between regions of n-doping (excess electrons) and p-doping (excess holes). Electrical neutrality (zero E field) would require the extra electrons to stay in the N region and the extra holes in the P region. However, the electrons are in rapid motion, and their thermal diffusion flattens out the density gradient of the free carriers. The mismatch between the bound charge (ions) and free charge (carriers) causes an E field to form in the junction region, even at zero bias. The magnitude of E is just enough to cause a drift current equal and opposite to the diffusion currents. Absorption of light causes the formation of electron–hole pairs, which are pulled apart by the E field, yielding a current at the device terminals. Away from the junction, the E field is shielded out by the free charge. Thus an electron–hole pair generated there will usually recombine before it can be collected, which reduces the quantum efficiency of the device. Applying a reverse bias pulls the free carriers back from the junction, forming a *depletion region* with a large E field throughout. If the doping level in the depletion region is low, applying a reverse bias can cause a very large change in the depletion layer width (Figure 3.1a,b). This reduces device capacitance typically 7× (by effectively separating the capacitor plates), and the extra depletion layer thickness improves quantum efficiency at long wavelengths, where many photons would otherwise be absorbed outside the depletion region. In an avalanche photodiode (Figure 3.1c), the doping profile is changed to make a separate high field region deep inside the device, in which electron multiplication takes place. (See Sze for more.)

3.3 SIGNAL-TO-NOISE RATIOS

There is a significant amount of confusion in the electro-optics field about how to calculate and quote signal-to-noise ratios, and power ratios in general. This arises from the fact that detectors are square-law devices; the electrical power emerging from the detector subsystem is proportional to the square of the optical power incident.

3.3.1 Square-Law Detectors

One way to look at this is that a photon of a given frequency ν carries an energy equal to $h\nu$, so that the optical power is proportional to the number of photons incident on the

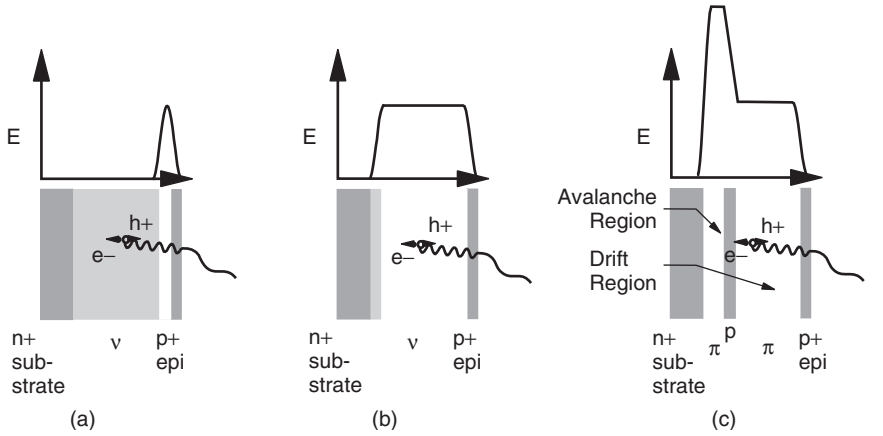

Figure 3.1. Photodetection in semiconductor diode: (a) PIN diode, zero bias; (b) PIN diode, high bias (fully depleted); (c) avalanche photodiode; ν and π are very low-doped n and p regions, respectively.

detector. In a quantum detector, each photon gives rise to an electron–hole pair, so that the electrical current is proportional to the photon flux. Electrical power is proportional to i^2, however, so the electrical power goes as the square of the optical power. Since signal-to-noise ratio (SNR) is defined in terms of power, rather than amplitude, the signal-to-noise ratio on the electronic side is the square of that on the optical side. The same is true of thermal detectors, since $\Delta T \propto P_{\text{opt}}$ and $\Delta V \propto \Delta T$.

An optical signal of 10^6 photons per second has an RMS statistical noise of 10^3 photons in a 1 second DC measurement (0.5 Hz bandwidth); since the optical power is proportional to the photon flux, the signal-to-noise ratio on the optical side is 10^3 in 1 second. On the electrical side, however, the signal power delivered to a load resistor R is $(10^6 e)^2 R$, while the noise power is $(10^3 e)^2 R$, so that the signal-to-noise ratio on this side is 10^6 in the same 1 second measurement. The two ratios behave differently with bandwidth, as well. In a 0.01 second measurement ($B = 50$ Hz), the optical signal-to-noise ratio is $\sqrt{(0.01 \times 10^6)} = 100$; it goes as $B^{-1/2}$. The electrical SNR is $100^2 = 10^4$ in 0.01 second; it goes as B^{-1}.

The author uses the electrical signal-to-noise ratio exclusively, because he finds he makes many fewer blunders that way. It is much easier to convert the signal and shot noise into electrical terms than to convert load resistor Johnson noise, multiplication noise, recombination noise, and so forth into optical terms, because the one conversion is physical and the other is not. Test equipment such as spectrum analyzers, lock-in amplifiers, and A/D boards all work in volts and amps. By using the electrical signal-to-noise ratio, it is unnecessary to mentally extract square roots all the time in order to have good comparisons between different measurements. One drawback to this approach is that if your colleagues are used to quoting SNR in optical terms, they may feel that you're grandstanding by using the electrical SNR. The only defense against this charge is to state clearly and often which ratio is being quoted.

Aside: Where Does the Power Go? Square-law detection is a surprisingly deep subject, leading as it does to coherent detection and large dynamic ranges, but there is a

simpler aspect that people puzzle over too. It is this: If I send an optical signal of power P_O into a photodiode, the electrical power $P_E = R_L(R \cdot P_O)^2$, which is less than the optical power when $R \cdot P_O < h\nu$ and greater than P_O when $R \cdot P_O > h\nu$. What's going on? Am I really wasting almost all my optical power when I'm stuck with small signals?

We need to distinguish between *energy* and *information*. A solar cell, whose job is to turn optical into electrical energy, wastes a lot of the incident energy: even if its quantum efficiency is 1, its power efficiency is $eV_F/(h\nu)$. However, if shot noise dominates, the counting statistics and hence the information content are identical before and after detection, so no information is lost or gained, even though the SNR is squared. The root of the confusion here is that we get so accustomed to identifying electrical SNR with information carrying capacity (correctly, as we'll see in Section 17.11.1) that we tend to forget that it's really the detection statistics that are fundamental for measurements. This is another reason to stick with electrical SNR throughout.

3.3.2 Photons

The photon concept is central to the theory of quantum electrodynamics, the relativistic quantum theory of electromagnetic fields. The interaction of light with matter is quantized; a very weak beam of light, which spreads out over a large area, nevertheless gets absorbed one photon at a time, each one in a very small volume. This is easy to see with an image intensifier. The mathematics of quantum field theory are the most accurate way of predicting the interactions of light with matter that are known at present, and that accuracy is very impressive in many cases. It is clear that this represents one of the towering achievements of 20th century physics.

Now let's come down to Earth and talk about instrument design. Photons are a very important bookkeeping mechanism in calculating photocurrent shot noise, and they are a useful aid in keeping straight the operations of acousto-optic cells. Beyond that, *the photon is the most misleading concept in electro-optics*. The moment people start thinking of photons flying around from place to place, they start finding mares' nests and hens' teeth. For our purposes, it's only the interaction of light and detectors that's quantized—in every other way, light behaves as a wave, and obeys Maxwell's equations to absurdly high accuracy.[†]

3.4 DETECTOR FIGURES OF MERIT

Comparing detectors based on widely differing technologies, manufactured in different sizes, and requiring different circuitry is often difficult. Various figures of merit have been developed to make it easier; unfortunately, these frequently have only oblique connections with the issues facing designers. A healthy scepticism should be maintained about the value of a small difference in a figure of merit, and practical considerations kept firmly in mind.

[†]Nobody knows what a photon *is*, for one thing: see Willis E. Lamb, Jr., "Anti-photon," *Appl. Phys. B* **60**, 77–84 (1995), and the supplementary issue of *Optics & Photonics News* from October 2003, which was devoted to the subject. Each of the eminent contributors had a strong opinion about what a photon was, and no two of them agreed.

3.4.1 Quantum Efficiency

A *quantum detector* is one in which absorbed photons directly create free carriers: a photodiode, photoconductor, or photocathode. The quantum efficiency (QE) η of such a detector is the ratio of the number of photodetection events to the number of photons incident, before any amplification takes place. (We usually get one carrier pair per detection event.) It is the most basic measure of how good a particular quantum detector is—the detector analogue of the transmittance of the optical system. It determines our signal to shot noise ratio, which sets an upper bound on the SNR of the measurement.[†]

Detectors with gain can reduce the effects of circuit noise but are powerless to improve the shot noise (and indeed contribute significant excess noise of their own). Close attention to quantum efficiency is an excellent defense against being led astray in the design of a detector subsystem.

3.4.2 Responsivity

The responsivity of a detector is the ratio of the output current to the optical power input, and is quoted in amps per watt or cute substitutes such as milliamps per milliwatt. It is the most appropriate figure of merit to use when considering the biasing of the detector and the gains and dynamic ranges of subsequent amplifier stages, and also in determining how much optical power will be needed to overcome Johnson noise in the electronics. It is easily seen to be given by

$$\mathcal{R} = \frac{M\eta e}{h\nu}, \tag{3.1}$$

where η is the quantum efficiency and M is the multiplication gain (unity for photodiodes). For a detector of unit quantum efficiency and unit gain, the responsivity is the reciprocal of the photon energy in electron volts; it ranges from 1A/W at 1.24 μm to 0.32 A/W at 400 nm in the violet. The responsivity of real detectors is typically a strong function of wavelength; not only are there fewer photons per joule at short wavelengths, but the detectors themselves are selective.

Responsivity is a reasonable basis for comparison of detectors without gain, such as different types of photodiodes, but should be used with caution otherwise; for example, a PMT with lower photocathode quantum efficiency but higher dynode gain may exhibit a higher responsivity than a high QE device—perhaps yielding a better measurement in very low light, where Johnson noise in the load resistor may dominate, but certainly a worse one in brighter light, in which the amplified shot noise from the photocathode is the major noise source.

The term *responsivity* is reused for a rather different parameter in photoconductors: the change of terminal voltage produced by an incident optical signal, in V/W, with some combination of bias current and bias resistor.

3.4.3 Noise-Equivalent Power (NEP)

The most important single parameter of a detected signal is its signal-to-noise ratio. Because most commonly used optical detectors are very linear, their intrinsic noise is

[†]There are very special situations ("squeezed states") in which this has to be qualified, but they aren't of much practical use.

additive in nature: it remains constant regardless of the incident optical power (PMTs and APDs are exceptions).

Detectors differ in their intrinsic gains and readout mechanisms, so that it is inappropriate to compare the noise performance of different detectors solely on the basis of their output noise currents. It is common to convert the noise into a noise-equivalent (optical) power, NEP. The NEP is defined as the optical signal power required to achieve a SNR of 1 (0 dB) in the given bandwidth. The output noise is broadband, but not in general flat, so that the SNR of your measurement depends on how fast your light is modulated. Accordingly, the NEP is quoted in (optical) watts, at a certain wavelength λ, modulation frequency f (see Section 13.3), and bandwidth B (usually 1 Hz)—NEP(λ, f,B). NEP is somewhat awkward to use, because a good detector has a low NEP; thus its reciprocal, the *detectivity*, is often quoted instead.

In order to be able to compare the performance of detectors of different sizes, the NEP is sometimes normalized to a detector area of 1 cm^2, and then is quoted in units of $W{\cdot}cm^{-1}{\cdot}\ Hz^{-1/2}$ (the somewhat peculiar unit is another example of optical vs. electrical power units—this one is optical). (Thermal detectors do not in general exhibit this area dependence, so their noise should not be normalized this way—see Section 3.10.7.)

Example 3.1: Silicon Photodiode NEP. Designing with visible and NIR photodiodes is easy, at least near DC, because there are really only three noise sources to worry about: Johnson noise from the load resistor, shot noise from leakage, and the shot noise of signal and background. As a concrete example of a noise-equivalent power calculation, consider a 3 mm diameter silicon PIN photodiode, Hamamatsu type S-1722, used with 820 nm diode laser illumination. This device has a quantum efficiency of about 60% at 820 nm, a room-temperature dark current of about 100 pA at 10 V of reverse bias, and a shunt impedance of 10^{10} ohms. Its responsivity is

$$\mathcal{R} = \frac{\eta e}{h\nu}, \tag{3.2}$$

which is 0.39 A/W at 800 nm. The two contributors to the dark NEP (of the diode alone) are the shot noise of the dark current and the Johnson noise of the shunt resistance R_P, so that the total current noise is

$$i_N^2 = \frac{4k_BT}{R_p} + 2ei_{\mathrm{DC}}. \tag{3.3}$$

With these parameters, the second term dominates the first by a factor of nearly 20, and the result is $i_N = 5.7 \times 10^{-15}$ A/Hz$^{1/2}$. The shunt resistance is not in general linear, so it is not safe to assume that $R_P = V_{\mathrm{bias}}/i_{\mathrm{leak}}$, although it'll be the right order of magnitude. To get from here to the NEP, we just multiply by the energy per photon of $hc/\lambda = 1.51$ eV and divide by the quantum efficiency, since a notional signal photon has only 0.6 probability of being detected, so that the optical power required to equal the noise goes up by 1/0.6. The final result for this case is that the NEP is

$$\mathrm{NEP} = \sqrt{2ei_{\mathrm{DC}}}\frac{hc}{\lambda\eta} \approx 1.6 \times 10^{-14}\ \mathrm{W{\cdot}Hz^{-1/2}}. \tag{3.4}$$

With a load resistance of 500 kΩ, this device will be shot noise limited with a current of $2kT/(eR_L)$, or 100 nA, corresponding to an optical power of 260 nW. The shot noise

of the dark current is so small that it would take a 500 MΩ load resistor to make it dominate. This is very typical.

3.4.4 *D**

The most used figure of merit for infrared detectors is D^*, the *specific detectivity*, that is, detectivity normalized to unit area. More specifically,

$$D^* = \frac{\sqrt{A}}{\mathrm{NEP}(f,\ \mathrm{BW})}. \tag{3.5}$$

This choice of normalization takes account of the area dependence of the noise, the wavelength dependence of the responsivity, and the frequency dependence of the noise and the sensitivity. (Consider a photoconductive detector with significant $1/f$ noise below 1 kHz, and a 3 dB cutoff at 1 MHz due to carrier lifetime.) This allows meaningful comparisons of detector types that may not have exactly the same area, and accounts to some degree for the behavior of detectors with gain, such as APDs and PMTs, whose huge responsivities may mask the fact that their noise is multiplied at least as much as their signals.

To choose a detector technology, we first prepare a photon budget, which estimates how much light is available with how much noise, and see if we can meet our electrical SNR target. Lots of common sense and sanity checking are required here, to make sure all relevant noise sources are considered, including signal-dependent terms such as shot and multiplication noise. The maximum NEP in the measurement bandwidth is equal to the optical power divided by the square root of the target electrical SNR (optical vs. electrical again). Once the detector area A has been chosen (normally the minimum size required to collect all the signal photons), the minimum D^* required is given by

$$D^*_{\mathrm{min}} = \frac{\sqrt{A}}{\mathrm{NEP}_{\mathrm{max}}}. \tag{3.6}$$

It must be emphasized that normalizing the response this way is intended to aid comparisons of different technologies and does not necessarily help the designer choose the right detector unit. In a background or Johnson noise limited system, if the signal beam can be focused down onto a 100 μm detector, then choosing a 5 mm one will result in a NEP needlessly multiplied by 50 for a given bandwidth, but D^* won't change. Thermal detectors' NEP is controlled by their thermal mass and thermal conductivity, so they often don't scale with area this way, making D^* useless.

D^* is not very useful in the visible, because it is so high that the detector's intrinsic noise is seldom a limitation. Visible light measurements are typically limited by the background and its noise, the shot noise of the signal, or the Johnson noise of the detector load resistor. In a poorly designed front end, amplifier noise may dominate all other sources, but this need not be. Chapter 18 discusses in detail how to optimize the signal-to-noise ratio of the detector subsystem. In brief, because visible-light photodiodes are such good current sources, the effects of Johnson noise can be reduced by increasing the value of the load resistance R_L. The signal power is proportional to R_L, whereas the Johnson noise power is constant, so the (electrical) SNR goes as R_L (there are circuit tricks to keep the bandwidth from vanishing in the process—see all of Chapter 18).

In the infrared, D^* appropriately includes the shot noise of the leakage current and of the 300 K background, shunt resistance Johnson noise, and lattice G-R noise. D^* is sometimes quoted in such a way as to include the shot noise of the nonthermal background (e.g., room lights), and even nonlinear effects such as multiplication noise in APDs and PMTs. While this is a legitimate way of quoting the noise performance of a measurement, it is unhelpful. Since it lumps all noise contributions into a single number, it totally obscures the questions most important during design: Which is largest, by how much, and what can be done about it? Furthermore, as discussed in the previous section, nonoptical noise sources are important, and translating all noise into optical terms is confusing and unphysical. Accordingly, all the D^* numbers quoted in this book are for detectors operating in the dark.

Another problem with D^* is that an inefficient detector with very low dark current may have a huge D^* but still be a very poor choice for actual use; for example, a 1 cm^2 detector with a quantum efficiency of 10^{-6}, but with a dark current of one electron per week, would have a D^* of about 1.6×10^{15} cm·$Hz^{1/2}$/W in the red, an apparently stunning performance, but actually it's useless. This illustration is somewhat whimsical, but nonetheless illustrates an important point: NEP and D^* are not the whole story.

Example 3.2: Indium Arsenide. Consider the case of a 2 mm diameter InAs photodiode (EG&G Judson type J12-5AP-R02M), operating over the wavelength range of 1.3–2.7 μm, at a temperature of 240 K and zero bias. This temperature was chosen to reduce the otherwise large response nonuniformity and to improve R_{sh}. The device has a long wavelength cutoff of about 3.5 μm. Its shunt resistance R_{sh} is 100 Ω, and the Johnson noise of the shunt resistance dominates the noise. Its quantum efficiency η is about 0.6 across this wavelength range, set mainly by the front surface reflection from the detector, which lacks an AR coating. The mean square noise current in a bandwidth B is given by

$$\langle i_N^2 \rangle = \frac{4kTB}{R_{sh}}, \tag{3.7}$$

where the thermal background has been neglected. The Johnson noise current is 11.5 pA/$Hz^{1/2}$, equal to the shot noise of a photocurrent of 400 μA (about 0.5 mW of optical power, far more than this detector would ever see). D^* is given by

$$D^* = \frac{\eta e}{h\nu}\sqrt{\frac{A}{\langle i_N^2 \rangle}}, \tag{3.8}$$

which at $\lambda = 2$ μm is about 1.5×10^{10} cm·$Hz^{1/2}$/W.

Infrared detection is where the concept of D^* is really crucial, because here the intrinsic noise of the detector is far from negligible. To verify that the detector is Johnson noise limited, we will estimate the shot noise of the background.

We assume that the planar detector is surrounded by a hemispherical black body at a temperature T, and that background photons whose wavelength is shorter than 3.5 μm are detected with $\eta = 0.6$, while all others are lost. Since $h\nu \gg kT$, the thermal photons are uncorrelated, and the thermal photocurrent exhibits full shot noise (see Section 3.10.2).

The photon flux per unit area, between ν and $\nu + d\nu$, from this hemisphere is $M_{q\overline{\nu}}\, d\nu$, where

$$M_{q\overline{\nu}}(T) = \frac{2\pi\nu^2}{c^2(e^{h\nu/kT} - 1)}. \tag{3.9}$$

Since the exponential in the denominator exceeds 10^6 over the wavelength region of interest, the -1 can be neglected, and the result integrated analytically, giving the approximate photocurrent due to the thermal background,

$$I_{BG} = \frac{2\pi\eta e}{c^2}\left[\frac{kT}{h}\right]^3 \int_{\frac{h\nu_0}{kT}}^{\infty} dx' x'^2 e^{-x'}$$

$$= \frac{2\pi\eta e}{c^2}\left[\frac{kT}{h}\right]^3 (x^2 + 2x + 2)e^{-x}\Big|, \tag{3.10}$$

where $x = h\nu_0/kT$ and $\nu_0 = hc/3.5\ \mu$m. After converting to current noise spectral density (see Section 3.10.2), this corresponds to a noise current i_N in 1 Hz of 0.6 pA, well below the Johnson noise. The limiting D^* if only this noise contributed would be 2.7×10^{11} cm·Hz$^{1/2}$/W.

Neglecting the 1 in the denominator causes less than a 1% error in the integrand whenever $h\nu/kT >$ ln(100), which applies for $\lambda < 10.4\ \mu$m at 300 K, or $\lambda < 40.6\ \mu$m at 77 K.

3.4.5 Capacitance

The product of the load resistance and the detector capacitance usually dominates the high frequency performance of the detector system (transit time and carrier lifetime effects may also be important in some instances, especially in photoconductors). Bandwidth is often the limiting factor in how large the load resistor can be, so capacitance partly determines the sensitivity of the detector subsystem. The capacitance per unit area of the detector is thus another important figure of merit.

Reverse Bias. If the detector type allows use of reverse bias, this can reduce the capacitance by as much as 7–10 times in some devices. High speed applications, which require small load resistors to achieve small RC products, may benefit from detectors with intrinsic gain; their gain allows them to overcome the high Johnson noise current of the load. See Sections 18.4.4 and 18.5 for other ways to sidestep the problem.

In CCD and CID detectors, which are inherently integrating, capacitance per pixel should be large, to increase the full well capacity. The maximum signal level depends on how many electrons can be stored, so increasing the capacitance makes the statistical fluctuations and the readout noise a smaller fraction of the maximum signal level.

3.4.6 Spectral Response

A detector with a high peak quantum efficiency isn't much use if it is insensitive to some wavelength you care about. On the other hand, low quantum efficiency can be very helpful sometimes, as in solar blind UV photomultipliers and in GaP or other direct bandgap detectors, whose quantum efficiency drops extremely rapidly at wavelengths longer than cutoff. Spectral flatness means different things in different situations; a bolometer tends to be flat in terms of resistance change per watt, irrespective of the photon energy, whereas a photodiode is flat in terms of electrons per photon. Thermal band broadening gives detectors large values of $\partial\eta/\partial T$ near their long-λ cutoff.

3.4.7 Spatial Uniformity

Detectors are not perfectly uniform in sensitivity. Silicon photodiodes typically have 1–10% variation in η across their active areas, due to coating nonuniformity and the sheet resistance of the very thin top layer. Large-area diodes are usually worse, as one might expect. Fringes due to windows and nonuniform passivation layers can make this even worse. Beyond about 1 μm, you can even get etalon fringes from the back surface of the silicon (especially in CCDs, which have no spatial averaging). Nonuniformity can be a problem in situations such as interferometric detectors, where it can degrade the angular selectivity of the measurement by preventing spatial interference fringes from averaging to zero as they should, in position sensing applications, and in any measurement calling for absolute radiometric accuracy. Some detectors are much better than others; specifications may be available from the manufacturer, but there's no substitute for your own measurements. For the highest accuracy applications, homogenizing devices such as integrating spheres provide a good solution; they are of course better for white light sources than for lasers, due to speckle (see Section 5.7.11). With lasers, it's usually best to put the photodiode at the pupil, because small angular shifts generally cause much less sensitivity variation.

3.5 QUANTUM DETECTORS

3.5.1 Photodiodes and Their Relatives

Photodiodes are the most popular detectors for optical instruments. This is no accident; they come in a huge variety of types, sizes, and characteristics, their performance is excellent, and their cost is usually low.

A shallow PN junction is formed in a semiconductor. Light entering the chip is absorbed, creating electron–hole pairs. Provided the absorption occurs in or near the depletion region of the junction, the large electric field there (produced by the junction itself and by any externally applied bias) separates the carriers rapidly. Thus they do not recombine appreciably until the charge accumulation is sufficient to overcome the field in the junction, either through forward conduction (as in an open-circuited solar cell) or through Debye shielding in areas of high light intensity.

Ordinary PN junction devices are useful in low speed applications but tend to have very large capacitances, especially at zero bias. PIN diodes have a thick layer of very low-doped (*intrinsic*) semiconductor between the electrodes, which increases the depletion layer thickness, and so reduces the capacitance enormously, to the point where carrier transit time can become the speed limitation. Moderate reverse bias can cause the whole thickness (500 μm or so) of the device to be depleted.

Photodiodes tend to have constant, high quantum efficiency over broad bands; AR-coated silicon photodiodes easily reach 90% η over the visible, and even with no coating, often reach 60%. They roll off at long wavelengths as the semiconductor becomes transparent, so that many of the photons pass completely through the junction region, and at short wavelengths as light is absorbed too shallowly for the carriers to even reach the junction before recombining. So-called blue- and UV-enhanced diodes use very thin top layers to minimize this, at some cost in linearity. Another approach is to use a Schottky barrier instead of a PN junction; a nearly transparent top electrode of thin metal or transparent conductor such as indium tin oxide (ITO) can substitute for the top semiconductor layer.

The quantized nature of the light-to-electricity conversion ensures that photodiodes are extremely linear, which is one of their most important characteristics. Furthermore, photodiodes made of silicon are extremely good current sources when operated at zero or reverse bias; as we saw in Example 3.1, their intrinsic noise arises almost entirely from the shot noise of their leakage current, which is itself low. The low intrinsic noise and leakage are largely traceable to the fact that silicon has a relatively wide bandgap, nearly 50 times the thermal voltage kT/e at room temperature, so that thermally generated carriers and thermal radiation are not significant contributors to the noise in most cases. Leakage can be reduced even further by cooling, although that is rarely necessary in practice.

Nonlinearity in reverse-biased photodiodes comes mainly from excessive current density. High photocurrent densities (>1 mA for a uniformly illuminated 3 mm diameter unit, for example) cause lateral voltage drops and intensity-dependent local increases in the carrier density. Both effects reduce the electric field in the junction, leading to enhanced recombination and slow drift. In other words, if you hit it too hard, it'll go nonlinear and slow down on you. A very small illuminated spot will exhibit this effect at surprisingly low photocurrents—microamps, even CW. See Section 3.5.4 for how much worse this gets with pulsed lasers. *Don't focus the beam down on the diode*.

Photodiodes used at zero bias are considerably less linear than reverse-biased ones, because lateral voltage drops in the thin epitaxial layer cause the diode to be locally forward biased, so that small forward currents flow and partly short out the photocurrent. Radiometrists resort to 20 mm diameter cells for photocurrents of 100 or 200 μA, whereas in reverse bias, the same detector is probably good to 10 mA. It takes a pretty big lateral voltage drop to overcome 5–20 V of reverse bias.

Silicon is the most common material used for photodiodes, both because it is a very good material and because silicon processing is a mature field. In general, silicon is best from 0.2 to 1 μm, and InGaAs from 1 to 1.7 μm (new InGaAs devices reach 2.6 μm). Germanium is also widely used for detectors out to 1.8 μm, but it is being eclipsed by InGaAs, which has been greatly developed for fiber optic communications at 1.3 and 1.55 μm. Beyond there, infrared devices become more exotic and difficult to use. High bandgap semiconductors such as GaP and GaN photodiodes ought to be the best kind in the ultraviolet, but don't seem to be as good as Si in practice. The one exception is silicon carbide, which has the advantage of being nearly totally solar blind; SiC devices have negligible leakage, work from 200 to 400 nm, and can reach $\eta = 0.72$ at 270 nm, which is remarkably good.

3.5.2 Shunt Resistance

Low light applications, where you want to use big diodes and huge feedback resistors (100 MΩ to 10 GΩ) are one place where cooling silicon photodiodes helps, for an interesting reason. To reduce leakage, you run the photodiode at exactly zero bias. Here's the subtlety: the leakage goes to 0, all right, but the shunt resistance deteriorates very badly at zero bias, because of the diode equation. In Section 14.6.1, we'll see that a diode's zero-bias shunt resistance r_0 decreases dramatically at high temperatures and low bandgaps, which is why IR detectors are worse than visible ones, and why cooling them helps. That resistance is a real resistance with real Johnson noise current, and furthermore if it gets too small, it forces the transimpedance amplifier to run at a huge noise gain

(see Section 13.1), which multiplies the amplifier's noise voltage, offset, and drift—bad news for measurement accuracy. Fortunately, we're very rarely in this situation.

Aside: The Zero-Bias Heresy. It is a sad fact that almost all the photodetector circuits published in the history of the world have been designed by circuits people who weren't also optics people. The author has no statistics to back this up, but he feels that it must be so, because almost all show the photodiode being operated at zero bias, often with great care being exerted to make the bias exactly zero. This will reduce the dark current through the photodiode, all right, but that isn't the problem we need to solve (see Section 18.2.1). Photodiode dark current is almost never the limiting factor in a visible or near-IR measurement. Fixing this nonproblem costs you a factor of 5–7× in bandwidth (or the same factor in high frequency SNR), as well as destroying the large-signal linearity, which makes it an expensive blunder. *Don't do it*.

3.5.3 Speed

For the highest speed applications, such as 40 Gb/s optical Ethernet, the speed of photodiodes becomes a serious issue. There are two effects that limit speed: transit time and *RC* delays. The transit time is how long it takes a carrier to reach the terminals. It can be reduced by making the diode thin, and maximizing the field in the junction by using as high a bias as possible and making the diffusion-dominated *p* and *n* layers extremely thin. This of course tends to increase the capacitance, so fast photodiodes tend to be very small. Another problem is that the QE suffers, because at the 850 nm wavelength used in fiber LANs, the absorption depth in the silicon is several microns; thus narrow-gap III–V materials such as InP are commonly used in fast NIR applications, which costs extra. One approach to fixing this is to use transverse devices, where a very small detector is coupled to a guided wave—the light travels perpendicular to the current flow, so the light path can be long and the current path short.

Aside: Plastic Packages. It appears that a few types of plastic-packaged photodiodes are susceptible to QE changes due to triboelectric charging of their plastic packages. The effect usually goes away with a bit of surface leakage (e.g., by breathing on the package) or slight heating. If your accuracy requirements are high, you may want to test for this, or stick with metal-can devices.

3.5.4 Photodiodes and Pulses

Now that femtosecond lasers are so popular, it's worth giving a bit of thought to the plight of the poor photodiode used to detect the pulses. Of course, the photodiode doesn't have the glamour job—or the easy one either. Consider a 100 fs laser with a 100 kHz pulse repetition rate, and an average detected power of 2.5 mW. A 100 fs pulse of 250 nJ (assuming it's a Ti:sapphire around 800 nm) produces a peak photocurrent of 160,000 A, and even though the diode can't possibly respond that rapidly, one wouldn't expect it to be terribly linear under such treatment, as indeed it isn't. One very good solution is to use a small (25 mm) integrating sphere on the diode (see Section 5.7.7). If it has a reflectivity of 97%, then incoming photons will typically bounce around over about a meter's distance before being absorbed, which will broaden the pulse to 2.5 nanoseconds or so. This is a better match to the capabilities of photodiodes. If your detector's diameter

is larger than the input aperture's, and larger than about 1/8 of the sphere's diameter, you'll collect most of the photons. Make sure the first bounce of the beam in the sphere isn't in the field of view of the detector, or you may still have nonlinearity problems. (Don't ablate the paint.)

3.5.5 Phototransistors

A phototransistor looks like a silicon bipolar transistor with a photodiode connected between base and collector. They have gain and are widely available at low cost, which about exhausts their virtues. Although their gain makes smallish photocurrents conveniently large, they are slow, leaky, nonlinear, and very noisy. They come only in very small sizes. Photodarlingtons have an additional gain stage and are even slower. These devices are classical examples of why amplification is not always useful; except for the very lowest performance applications, avoid them at all costs.

3.5.6 Prepackaged Combinations of Photodiodes with Amplifiers and Digitizers

Several detector and op amp manufacturers build packaged combinations of photodiodes with transimpedance amplifiers, current to frequency converters, "current amplifiers" (Hamamatsu), or even $\Delta-\Sigma$ type A/D converters. These devices are intended for people who are not comfortable designing detector circuits, and you pay in cost and performance for the convenience of using them despite frequent claims to the contrary by their manufacturers. (Their data sheets very seldom mention noise performance, which is a very bad sign.) The performance problems mostly arise from their use of fixed value internal feedback resistors (often 1 MΩ), and failure to put in a cascode transistor to reduce junction capacitance effects or a tee network to decrease the second-stage noise contribution (see Section 18.4.12 for more details). They can be useful where there is lots of light available and speed is not a limitation, or where skilled engineering time is very limited; people commit much worse blunders than these all the time, and using these devices at least limits the damage. These devices may improve in quality and cost effectiveness in the future; at present, however, designers who are willing to do a bit more work can usually achieve much better performance at a considerably lower cost. The exception to this rule is some packaged APD/GaAs FET amplifier combinations, which provide a level of speed and noise performance that is not trivial to reproduce.

3.5.7 Split Detectors

It is often handy to be able to measure the position of a beam on a detector. If image sensing is not absolutely required, the best way to do this is by using split detectors (bi-cells and quadrant cells) or lateral effect devices. A split detector is just that: two or more detectors spaced very closely, often arranged as sectors of a circular disc. Each detector has at least one lead of its own (often all the anodes or all the cathodes are common, to save leads and ease fabrication). By combining the separate photocurrents in various ways, such as subtracting them or computing their ratio, small changes in the beam position can be measured very accurately (often to 0.1 nm or better near null, at least at AC). Because their geometry is defined lithographically, they are

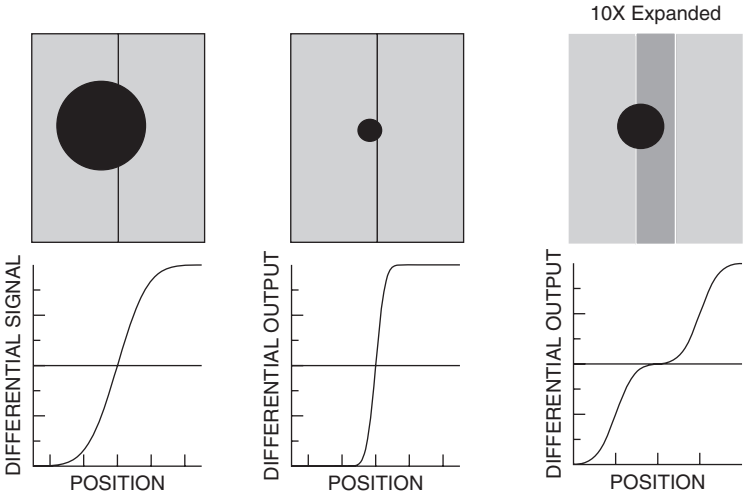

Figure 3.2. Split photodiode, showing the dependence of output signal on beam diameter and position.

very stable, and because each segment is operated as an independent photodiode, split detectors have the same virtues of linearity, low noise, and high efficiency we expect from single photodiodes. Split detectors are currently available in Si, Ge, InSb, and HgCdTe.

The main problems with split detectors are that some of the light is lost by falling into the kerf between the cells, and that the position sensitivity depends reciprocally on the beam diameter (See Figure 3.2). This is physically obvious, since a full-scale change results as the beam moves from being entirely on one side to entirely on the other. If the position information is obtained by subtraction, the sensitivity will depend on the optical power as well. Using analog division instead of subtraction helps.

Another approach is to use the cells in the open-circuit photovoltaic mode, where the photocurrent is allowed to forward bias the cell (as in a solar cell), and subtract the open-circuit voltages. These voltages depend logarithmically on the photocurrent, so when they are subtracted, a ratiometric measurement results. The circuit responds slowly, but this trick is good for alignment sensors (see Section 12.9.11).

3.5.8 Lateral Effect Cells

Another position sensing device, which avoids the problems of split detectors, is the lateral effect cell. Both 1 D and 2 D devices are available, in Si, Ge, InSb, and InAs. They are single large photodiodes that use a thin, highly resistive layer for the top electrode of the cell, each end of which has its own lead. The output leads are connected to low impedance points (such as op amp summing junctions). The light beam appears as a current source located somewhere on the surface, so that the photocurrent divides itself between the output pins in proportion to the conductance of each path. Because the conductance depends on the distance from the light beam to the output pin, the ratio of the currents in each pin gives the location of the light source. Because the cell surface is uniform, and the current division linear (i.e., two or more light sources shining on the

same cell produce the sum of the outputs each would produce by itself), the position signal is much less sensitive to beam diameter, until the edges of the beam approach the edges of the cell or the current density or lateral voltage drops get so high that response nonlinearity sets in. Two-dimensional lateral effect cells come in two varieties: *pincushion*, where the readouts are in the corners, and *linear*, where the anode is split in one axis and the cathode in the other, so the anode sheet gives x and the cathode y. The linear ones give position information that is accurate to 1% or so, with the best ones achieving 0.1% over 80% of their aperture. Lateral effect cells tend to be much slower than split detectors, since they have a fairly large series resistance (1–200 kΩ), and also more difficult to use. The difficulty is that, because of the series resistance, even moderate photocurrents can cause lateral voltage drops large enough to locally forward bias the junction, leading to serious nonlinearity. The simplest way to avoid this problem is to reverse bias the junction. There are also some inferior devices where the top electrode has four long contacts, along the sides of a square, and the bottom has only one. These superficially resemble linear lateral effect devices but differ in one crucial respect: the two axes compete for current. If the illuminated spot is near the $-x$ electrode, that electrode will suck up almost all the photocurrent, leaving little to drive the y axis, which therefore is rendered nearly indeterminate by noise and drifts.

The low-light performance of lateral effect cells is relatively poor, because the resistance of the silicon layer appears in parallel with the outputs. It thus contributes a large amount of Johnson noise current, with a noise source connected between the two outputs of each axis. This is an obnoxious problem, because although these noise currents of course sum to zero, they must be subtracted or divided instead in order to make a position measurement; this causes them to augment instead of canceling. Unless the photocurrent is sufficient to drop $2kT/e$ (50 mV at room temperature) across the cell resistance, the measurement will be Johnson noise limited. It does not help much to increase the load resistance, since the differential signal will always be loaded by the cell resistance, so that the effective load for differential measurements will be the parallel combination of the cell resistance and the external load. In general the lateral effect cell is a fairly noisy, kilohertz-range device for sensing beam positions, but it is relatively immune to beam-size effects that can plague split-cell measurements.

3.5.9 Position Sensing Detector Pathologies

Split detectors are more sensitive to etalon fringes than single element ones. A smooth beam profile on a single element detector does a much better job of preserving the orthogonality of misaligned beams, whose fringe patterns average to zero over the surface. In a split cell, there's a great big cliff in the middle of the fringe pattern, so a small phase shift with temperature can cause an overwhelming amount of drift—all the more so since the desired signal is a small difference between two large currents. Position-sensitive detectors are also vulnerable to nonorthogonalities of the measurement directions caused by sensitivity variations across the detector.

3.5.10 Other Position Sensing Detectors

For position sensing applications, lower cost solutions such as a few discrete diodes plus a shadow mask should not be overlooked (Figure 3.3). For many purposes, they can be very useful and are dramatically cheaper than the combination of a lens plus a position

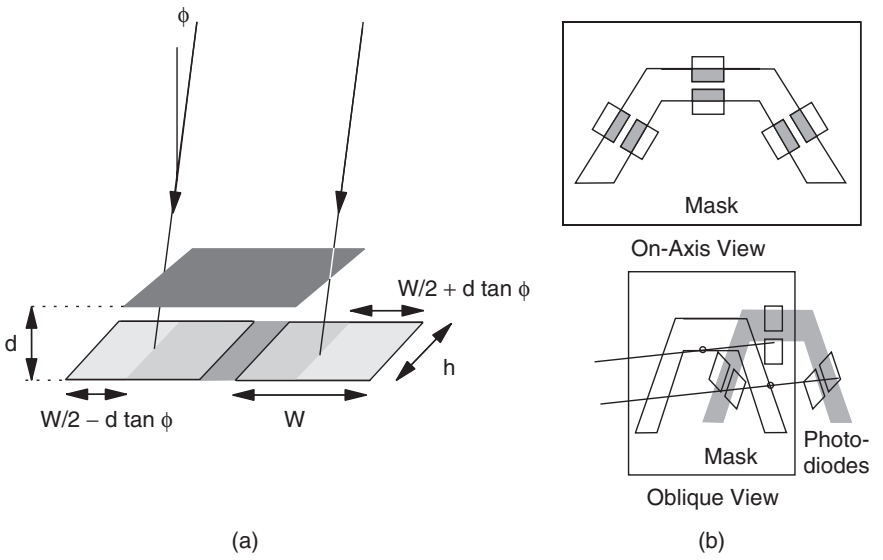

Figure 3.3. Three-dimensional shadow mask detector: (a) schematic and (b) drawing of actual device.

sensing diode. When using mid- and far-infrared detectors, where split detectors and lateral effect cells are harder to get, this may be the only choice feasible.

Example 3.3: Two Solar Cells and a Mask Versus a Bi-cell. As an example of a cheap position sensing detector, consider two rectangular solar cells of width W and height h, plus a thin mask at a height d, covering the right half of the left cell and the left half of the right cell, as shown in Figure 3.3. Uniform illumination of intensity I comes in at an incidence angle ϕ, illuminating $W/2 + d \tan \phi$ of cell 1, and $W/2 - d \tan \phi$ of cell 2. Obviously, if the angle is too large, one diode will be completely covered, and the other one completely illuminated. For angles smaller than this, if the two photocurrents are subtracted, the result is

$$i_- = 2I\mathcal{R}\, dh \sin \phi, \tag{3.11}$$

where I is the power per unit area, measured in a plane normal to the incoming beam. If the difference is normalized by dividing by the sum of the two currents, the result is

$$i_{\text{norm}} = \frac{2d \tan \phi}{Wh}, \tag{3.12}$$

which is plotted in Figure 3.4, together with the angular uncertainty due to shot noise alone. You can also divide instead of subtracting, which gives

$$\frac{i_2}{i_1} = \frac{W + 2d \tan \phi}{W - 2d \tan \phi}. \tag{3.13}$$

It is apparent that the sensitivity of the shadow mask/multiple element detector can be adjusted over a wide range by changing d, a desirable property. If the light comes

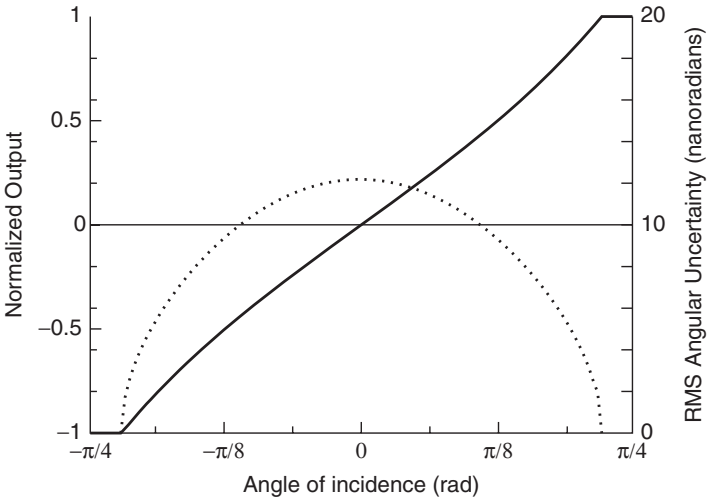

Figure 3.4. Output and RMS angular error of a shadow mask sensor, with 5 mm square detectors and a 3 mm mask spacing. $P_{\text{opt}} = 10$ mW/cm^2.

from a wide spectrum of angles, it is sensible to integrate Eq. (3.11) over angle. If the differential characteristic is not required, a single photodiode with a shadow mask can be used instead, at an even greater cost saving.

If two axes are required, two such single-ended detectors can be used, with the edges of their shadow masks mutually perpendicular. If the differential character is important, then three or four detectors can be combined. If four are used, the 1D expressions are approximately correct, except for the additional effect of obliquity in the other axis. For the three-detector case, as shown in Figure 3.4, the X and Y directional signals are given by

$$X = \frac{i_3 - i_1}{i_1 + i_2 + i_3} \tag{3.14}$$

and

$$Y = \frac{i_3 + i_1 - 2i_2}{i_1 + i_2 + i_3}, \tag{3.15}$$

respectively. This saves one photodiode, while preserving the differential property: light coming in exactly on axis gives $X = Y = 0$; the third dimension comes from intensity or (for a nearby point source) parallax, in which case the simple expressions given have to be modified.The noise performance of shadow mask detectors, like that of quadrant cells, is excellent. In a 1 Hz bandwidth, the shot noise limited RMS angular uncertainty is

$$\left\langle \frac{\delta\theta}{\Delta\theta_{\text{pp}}} \right\rangle \approx \sqrt{\frac{2e}{i_0}}, \tag{3.16}$$

where i_0 is the response of one unobscured detector to the incident optical intensity. A HeNe laser beam of 1 mW per detector will give $i_0 \approx 300\ \mu$A. Assuming a mask spacing

of half the detector width, which results in a full scale range of $\pm\pi/4$ radians, the 1 Hz RMS angular uncertainty is around 50 nanoradians, or 0.01 arc second.

The accuracy attainable by this technique is limited mainly by nonideal beam characteristics, such as amplitude nonuniformity, multiple scattering between mask and detectors, and diffraction of the beam at the edge of the mask. †

3.5.11 Infrared Photodiodes

Photoelectric detectors for the mid- and far-infrared region also exist, based on compound semiconductors such as indium antimonide (InSb), indium arsenide (InAs), platinum silicide (PtSi), and mercury cadmium telluride (HgCdTe, sometimes written MCT, and pronounced "mercadtell"). Their characteristics are not as good as those of silicon and InGaAs photodiodes, and they are much more expensive. Compound semiconductor materials are more difficult to process, because small errors of stoichiometry appear as huge dopant densities, and because their small markets dictate that their processing is much less developed than that for silicon (GaAs and InGaAs are the exceptions). In the mid-IR, the standard detector at present is HgCdTe.

Of the near-IR devices, Ge was king for a long time. It has the advantage of being an element, so the stoichiometry is not a problem, but its poor leakage performance means that it requires cooling in situations where compound semiconductor devices such as InGaAs do not. In the past, InGaAs detectors were too expensive for widespread use, but now that demand for detectors in fiber optics applications has funded their further development, costs have fallen to a few dollars for a usable InGaAs detector at 1.5 μm.

Mid- and long-wavelength detectors, such as HgCdTe, PbS, PbSe, PtSi, and InSb, frequently require cryogenic cooling, which is inconvenient and expensive ($2500 for a single cooled detector element). InAs, with response out to 3.5 μm, is often useful with only thermoelectric (-20 to $-60\,^{\circ}$C) cooling.

Unlike silicon detectors, the shunt resistance of an infrared diode can be very low, as in Example 3.2; the Johnson noise of this low resistance is the dominant additive noise source in IR devices that are not cryogenically cooled. Because of the low shunt resistance, carriers generated far from the electrodes are sometimes lost, leading to very large spatial nonuniformities of response (Figure 3.5 shows a $4\times$ changes in η with position).

3.5.12 Quantum Well Infrared Photodiodes

A promising new mid- and far-IR detector technology is based on quantum wells. These are basically the famous square box potential well, whose energy levels can be tailored by adjusting the depth and width of the well. Accordingly, the bandgap of the detector can be as narrow as desired, without having to deal with the poor properties of narrow-gap semiconductors. Thermionic emission is still a problem with these devices, so they must be cooled, but their D^* can be as high as 10^{12} at 9 μm, a remarkable performance. They can also be made in arrays. We'll see more of these in the future; the technology is driven by space-borne and military sensors.

†Do the diffraction ripples at the leading and trailing edges of the shadow cause ripples in the *XYZ* outputs versus angular position? Why or why not?

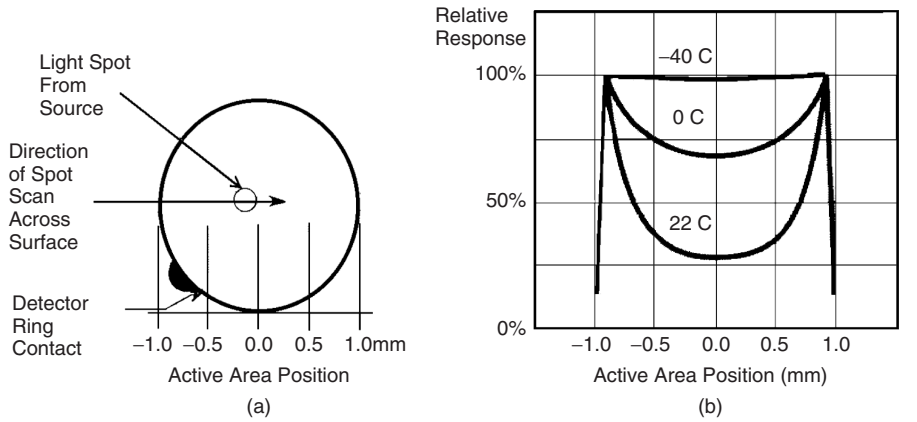

Figure 3.5. Response nonuniformity and shunt resistance of a commercial InAs photodiode (EG&G Judson J12-5AP-R02M) versus temperature. High lateral and low shunt resistance leads to poor uniformity at 300 K. (Courtesy of EG&G Judson Inc.)

3.6 QUANTUM DETECTORS WITH GAIN

3.6.1 Photomultipliers

At wavelengths shorter than 2 μm, thermally generated photons are very rare, so in principle a detector should be limited only by signal photon statistics (shot noise). However, Johnson noise in load resistors, amplifier input noise, and other circuit noise typically dominates shot noise at low photocurrents. That's where photomultiplier tubes (PMTs) come in. PMTs use *electron multiplication*, which we will see more of later, to amplify the generated photoelectrons before they become mixed with the noise currents of the electronics. Electron multiplication occurs when an electron hitting a surface causes it to emit more than one *secondary electron*. The way this works in a PMT is as follows: a photocathode is exposed to incoming light. By the photoelectric effect, some fraction of the incoming photons cause the photocathode to emit photoelectrons from its surface. These photoelectrons are electrostatically accelerated and focused onto another electrode, the *first dynode*, which is coated with a material selected for high secondary electron yield. The secondaries from the first dynode are accelerated and focused onto the second dynode, and so on for 5 to 14 stages, before finally being collected by the anode. By the end, each photoelectron has become a pulse 300 ps to several nanoseconds long, containing perhaps 10^5 to 10^7 electrons, which is easily detectable above the Johnson noise.

Aside: Electron Affinity. When a photoelectron is generated inside a photocathode or dynode, it must escape into the vacuum before it's any use to us, and to do so it has to overcome a potential barrier. The height of this barrier is a material property called the work function W. The photoelectron is continually losing energy through collisions, and the requirement that it arrive at the surface with at least W limits the photoelectron yield. The work function is made up of two parts, the classical image potential, which is the work required to separate the electron from its image charge in the metal surface,

and the electron affinity.[†] The image potential is always positive, but negative electron affinity (NEA) materials exist, and their lower work function leads to improved electron yield, at the expense of slower response.

3.6.2 PMT Circuit Considerations

This elegant scheme requires a certain amount of circuit support: a power supply of from −500 to −2000 volts, with taps for the photocathode and all the dynodes. This is usually provided by a powerful (1–2 W) high voltage supply and a multitap voltage divider string made of high value (100–300 kΩ) resistors. The high supply power is almost all dissipated in the resistors, which sets a practical lower limit on their values. The exact bias voltages are not terribly critical (unlike APDs, for example).

Because of the high electron gain, the last few dynodes need a fair amount of bias current, which is supplied only poorly by the resistor string. Bypass capacitors or Zener diodes on the last couple of stages are a help, but nevertheless the nonlinearity of most photomultiplier systems at high light intensity is dominated by voltage drops in the dynode bias string. It is often convenient to change the supply voltage to vary the gain, and having zeners on in the string makes this hard, because the voltage distribution on the dynodes will change as the supply voltage is altered. Since the linearity depends significantly on this distribution, zeners reduce the flexibility of the system.

The Cockroft–Walton (C-W) generator is a many-section voltage multiplier based on a diode–capacitor ladder structure. An N-stage C-W generator produces N nearly equally spaced voltage taps, making it a natural fit for biasing PMT dynodes. Some photomultiplier modules now include C-W multipliers rather than voltage dividers.[‡] C-Ws have the useful property that the lower taps have much lower impedance than the high voltage end, a good match for the needs of PMTs. Besides compactness and low power, this leads to enormously improved linearity in Cockroft–Walton devices, so that the resistor scheme is obsolescent for most purposes. Watch out for the modules with preamps built in—they all cut off at 20 kHz to avoid the power supply ripple.

The most appropriate uses for resistor biasing nowadays are in applications where linearity is not vital but power supply ripple is extremely objectionable, or in devices where the photocathode cannot be run at a large negative bias, and DC coupling is not needed. Running the photocathode near ground means that the anode and last dynodes must be run at a high positive voltage, which significantly reduces the advantages of the Cockroft–Walton technique, since the high current electrodes are at the high impedance end of the supply.

Applications requiring a grounded photocathode include scintillation detectors, where the scintillator is an ionic crystal such as NaI, which must be operated at ground for safety reasons. If the photocathode is not grounded, the large potential gradient across the glass envelope of the PMT in an end-on tube can lead to electrophoretic motion of ions toward the photocathode, which destroys the tube by photocathode corrosion. Side-looking PMTs with opaque photocathodes (electrons are emitted from the same side the light hits) are immune to this since the photocathode doesn't touch the envelope.

[†]For insulators the electron affinity may be less than this, because an added electron goes into the conduction band, whereas a photoelectron comes from the valence band.

[‡]This idea took awhile to catch on—it was first published in 1960 (R. P. Rufer, Battery powered converter runs multiplier phototube. *Electronics* **33**(28), 51 (1960)).

The noise of PMTs is dominated by thermionic emission from the photocathode, leading to dark current spikes, and by variations in the dynode gain (especially at the first dynode), which leads to multiplicative noise. The average number of secondary electrons is only 5 or so, although the NEA material GaP(Cs) can achieve 20–50. As we'd expect from a low yield random process, the size of the pulse from a single photon event varies within a 1σ range of $\pm\sqrt{5}/5 \approx \pm 45\%$, although with larger photocurrents these error bounds are greatly reduced through averaging. PMTs require very high insulation resistances between their internal elements, but use volatile metals such as cesium and antimony, which are prone to migrate at high temperatures; thus PMTs cannot be baked out as thoroughly as most vacuum systems. There is always some residual gas (perhaps 10^{-6} torr, mainly water), which leads to artifacts known as *ion events*. An ion event takes place when a positive ion is generated in the residual gas inside the PMT near the photocathode. Because of its positive charge, it accelerates and hits the photocathode hard enough to knock loose many electrons. This large pulse is amplified through the dynode chain, producing a very large current pulse at the anode. Ion events are even more common in old or poor quality tubes, or those operated near radioactive sources.

Related to ion events are afterpulses, which are secondary pulses that sometimes occur, usually 20–100 ns after a photon is detected. These arise from photoemission inside the PMT due to electron impact on a surface, or a generated ion that hits a dynode instead of the photocathode. Afterpulses are a problem with fast photon counting setups; putting in 100 ns of dead time after each photocount will more or less cure the problem. This dead time reduces the integration period at higher light levels, so it has to be corrected for (see Section 7.3.1).

PMTs get old and wear out, at a rate largely controlled by the anode current. They exhibit strong (several percent) effects due to warmup, intensity and voltage hysteresis, and other historical causes. Under good conditions, over its life a PMT can produce an integrated anode charge of a few hundred coulombs per square centimeter of photocathode. In long-term dark storage, or in low current applications, the lifetime approaches a constant value of a few years, limited by helium diffusion through the tube envelope and by surface changes inside. Tubes with soda lime glass envelopes are the most vulnerable; high partial pressures of helium can kill one of those in an hour.

PMTs can be quite fast; ordinary ones have rise and fall times of a few tens of nanoseconds and the fastest are around 250 ps, with timing uncertainties of 40 ps or thereabouts. Fall times are rather slower than rise times. In a high gain PMT, a single photon can produce 10^7 output electrons, which in an 8 ns wide pulse amounts to 200 μA. Such a large current can produce 10 mV signals even across a 50 Ω load, making it feasible to count individual photons at high speed. The pulses are repeatable enough in height that simple thresholding can easily distinguish a pulse corresponding to a single detected photon from noise, and from multiphoton or ion events. Such *photon counting* is an important application of PMTs, and photon rates of up to 30 MHz can be accommodated with good accuracy with commercial gear (200 MHz with special hardware). The combination of high gain and low dark count rate makes PMTs uniquely suited to photon counting.

In photon counting, the photoelectrons arrive more or less one at a time. PMTs can also be used in analog mode, where the average anode current is measured in much the same way as with a photodiode. Of the two, photon counting is better behaved. This is mainly because the gain of a PMT depends on everything you can think of. In the analog

mode, this directly affects the signal level, whereas in photon counting it merely changes the pulse height, without making a pulse significantly more or less likely to be detected.

A photon counting PMT has a sensitivity similar to that of a cooled CCD. It has no spatial resolution, but on the other hand you don't have to wait for the end of the integration time to see the data.

The amplification mechanism of PMTs is very clever and effective, and their optical performance has recently improved by a factor of nearly 2. Conventional bialkali PMTs tend to be narrowband, peaking strongly in the violet. PMTs are available in large sizes, up to about 300 mm in stock devices, and up to 600 mm in custom devices. The combination of huge area and low dark count rates is unique to PMTs.

Choosing a Photocathode Material. The choice of photocathode material depends on the application; they are typically made of a mixture of antimony with alkali metals or of a direct bandgap compound semiconductor with traces of cesium. Infrared units are available too. The classic Ag-O-Cu S-1 material reaches 1.1 μm, but has extremely low quantum efficiency (0.01–1%) and has a lot of dark current, often requiring cryogenic cooling.

NEA photocathodes work further into the IR than alkali metal ones, but have a much slower response (1 ns rather than <50 ps), making them less suitable for exotic applications such as streak cameras. The best InGaAs NEA photocathodes reach 1.7 μm, with quantum efficiencies of around 20%, and the best enhanced bialkali and GaAsP ones achieve about 45–50% peak quantum efficiency in the blue and near-UV.

How to Kill a PMT. Photomultipliers can be destroyed by exposure to daylight when powered, and their dark current can be dramatically increased by such exposure even when unpowered; several days of powered operation may be required to bring it back to normal.

Making Accurate Measurements. The photon counting mode is pretty trouble-free, but it is not trivial to make accurate analog measurements of absolute light intensity. For instance, the total gain typically varies $\pm 10\%$ with position on the photocathode. Accuracies of around 1% result from ordinary care, but by taking sufficient pains to control or calibrate small effects, accuracies and repeatabilities of 0.1% can be achieved.

The gain of a PMT is affected by many internal and external effects, including shock, age, dynode fatigue due to high current operation, stray fields, and even changes in geometry caused by gravity or acceleration. The efficiency and speed with which the first dynode collects the primary photoelectrons depends on position, leading to variations of $\sim$10–20% in sensitivity and a few percent in delay. Due to Fresnel reflection at the photocathode surface, the quantum efficiency also depends on incidence angle and (off-normal incidence) on polarization. Besides these parabolic-looking variations, the sensitivity of some PC types shows ripples with position, which get much worse at longer wavelengths, so it's often helpful to put diffusers in front of PMTs.

The gain of some tubes can be reduced 40% by a 1 gauss magnetic field (the Earth's field is about 0.5 gauss), although others can work in 1 kilogauss. Mu-metal shields improve this greatly, but mu-metal's shielding properties are easily destroyed by shock or bending. The lower energy electrons between the photocathode and first dynode are most susceptible to magnetic steering, so mount the tube so the shield sticks out about one diameter in front of the photocathode, to allow the fringing fields space to die off.

Static charges on the envelope or nearby grounded objects can cause electrons to strike the tube envelope, generating spurious light. Photoemission from the gas and from dynode surfaces, as well as Čerenkov light from cosmic rays and radioactive decay, cause spurious counts. Light from these sources can be guided by the glass tube envelope directly to the photocathode, where it will cause spurious counts. Graphite paint (DAG) applied to the envelope and kept at cathode potential (via a 10 MΩ high voltage resistor for safety) eliminates the light guiding and provides an electrostatic shield. High electric fields inside the envelope can cause scintillation as well, so use DAG and really good insulation (e.g., 4 mm of high quality silicone rubber, not 10 layers of PVC tape). Taking care here can reduce the dark counts by a factor of 10. High humidity is a disaster for PMT performance due to external leakage currents.

Being vacuum tubes, PMTs are easily destroyed by shock and may be microphonic in high vibration environments; also, of course, they are expensive. PMT manufacturers provide 200-odd page manuals on how to apply PMTs in measurements, and these are full of lore. If you need PMTs, get a few of these application manuals.

3.6.3 Avalanche Photodiodes (APDs)

When a high electric field is applied to a semiconductor, free carriers can acquire enough energy to excite other carriers through *impact ionization*. These newly generated carriers can themselves create others, so that a chain reaction or *avalanche* results. At higher fields still, carriers can be generated spontaneously by the field, and breakdown occurs. When this mechanism is used to multiply the photocurrent in a photodiode, the result is an APD. Reasonably stable multiplication gains (M) of up to 100 or so are possible by controlling the bias voltage carefully at a value of about 90% of the breakdown voltage, and compensating its strong temperature dependence.

Holes and electrons have different values of the ionization coefficient α (a normalized cross section). One might think that the performance would be best when both contribute equally, but in fact that's the worst case. All the carriers have the same speed, so in a pure electron avalanche, all the secondary electrons arrive at the same time as the primary photoelectron, and the holes are spread out by the transit time τ. In a bipolar avalanche, the holes cause ionizations, so that the avalanche spreads out in both directions and bounces back and forth until it dies away due to statistical fluctuations in the rates. This makes it become very slow and very noisy as the gain increases. The figure of merit for this is $k = \alpha_h/\alpha_e$, the ratio of the ionization cross sections of the less ionizing species (holes) to the more ionizing (electrons). The situation to avoid is $k \approx 1$. In silicon, k is very small, because holes cause almost no impact ionization, but in low bandgap materials like InGaAs, k is around 0.3 at low voltage, rising to nearly 1 at high voltage. Heterostructure APDs exist, in which the detection and multiplication are done in different semiconductors; this gives the best of both worlds at the price of complexity and expense. The noise is always worse at high gain; the noise power tends to increase as M^{2+m}, where the noise exponent m is 0.3–1.0, so that the SNR goes down as $M^{-0.3}$ to M^{-1}. The exact exponent is device dependent, and manufacturer's specifications should be consulted—if you're lucky enough to find a data sheet that has that level of detail. (Optical detector manufacturers should be ashamed of themselves for the uniformly poor quality of their data sheets.) The excess noise performance of APDs has been improving, due to innovations such as separating the photodetector region from the multiplication region, so this may become less of a problem in future.

The other major excess noise contributions come from uncertainties in the position of the first ionization event of the avalanche. Simple designs where the multiplication occurs in the absorption region have a 6 dB noise penalty, because some photons are absorbed deep into the multiplication region, so that the available gain is much less (they also have horrible variations of gain with wavelength, for the same reason). Newer designs that separate the absorption and multiplication regions via control of the doping density are much quieter and flatter with λ.

APDs exhibit a gain–bandwidth trade-off almost like an op amp's. Due to the finite value of k, the avalanche takes more time to build up and die away as M increases, so the bandwidth tends to go as $1/M$. Even at low gain, the time it takes carriers to transit the (thick) multiplication zone limits the ultimate bandwidth. That's why the quickest photoreceivers (40 Gb/s or $\sim$30 GHz) use Schottky photodiodes and accept the reduced sensitivity.

The inherent SNR of the photocurrent generated by an APD is monotonically decreasing with gain; the signal quality gets worse and worse—so why on earth use them? The great virtue of APDs shows itself in wide bandwidth systems. For speed, these systems have to use low load impedances, and so are limited by the large additive Johnson noise. When the additive noise dominates all other noise sources, the (electrical) signal-to-noise ratio improves by M^2 until the excess noise from the APD becomes comparable to the Johnson noise, at which point the SNR peaks and begins to deteriorate. Thus the operating M should be chosen so that the excess noise roughly equals the Johnson noise of the amplifier (actually a bit higher since the falloff in SNR is very slow in the direction of increasing M, so the peak is not exactly where the two noises are equal). Alternatively, compared to a PIN diode, you can reduce the load resistance by a factor M^2, which can help the bandwidth a lot. The price you pay is slightly lower SNR due to multiplication noise, narrower optical bandwidth, extra cost, and uncertainty in the exact operating gain.

The gain of an APD is a very strongly increasing function of bias voltage near breakdown, as shown in Figure 3.6 for a Hamamatsu S5343; a $\pm 20\,^{\circ}$C change will make a nominal gain of 70 vary between 30 and 200. While this can be calibrated, and the bias

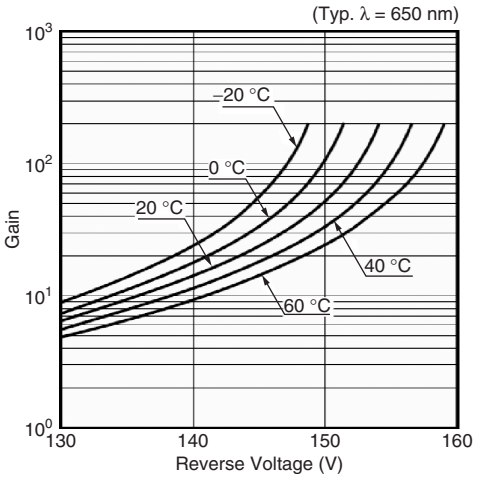

Figure 3.6. Gain of a Hamamatsu S5343 Si avalanche photodiode versus bias voltage, for various temperatures.

voltage or temperature controlled to keep it within closer bounds, we're never going to get even 1% accuracy over temperature with wild swings like that. This makes APDs poorly suited to accurately calibrated jobs in analog mode.

If you have matched APDs, either monolithically or by sorting, it is possible to stabilize the multiplied dark current (and hence the gain) of an illuminated APD by putting a constant current into the dark diode and applying the resulting voltage to the illuminated device. (You have to use a lowpass filter and a buffer, plus some thoughtfully chosen safety limits to avoid blowing up expensive APDs.) This is twice as expensive, somewhat noisier, and vulnerable to gradients, but saves the trouble of running temperature calibrations, which tend to be slow and hence expensive.

3.6.4 Photon Counting with APDs

APDs operated in the breakdown region (*Geiger mode*) are often used in medium-performance photon counting applications. The bias must be reduced after each event to stop the avalanche, typically by putting a huge resistor ($\approx$100 kΩ) between the device and the bias supply, with the output taken across the diode. The resulting *RC* time constant makes the recovery slow (1 μs), but this does not affect their rise time, which can easily be below 1 ns—20 ps has been reported.

Since all light pulses produce an output of the same size, pulse height discrimination is impossible. Compared with compact photomultipliers, these devices are more rugged, have higher quantum efficiency, and are available in larger diameters, but because of their slowness they have no compelling advantage in signal detection performance. Circuit improvements[†] can get them down to around 50 ns, which is somewhat better.

APDs emit a small amount of light when they break down, so if you have a multi-APD setup, you can get optical crosstalk. It's a good idea to reset all the APDs whenever any of them fires. You can also get segmented APDs intended for counting bursts of multiple photons in Geiger mode. In these devices, the active region is pixellated, but all the segments (as many as 14,000) are wired in parallel, via integrated quench resistors. These have higher dynamic range (since many segments can break down at once) but no better timing characteristics. Pulses from different segments are reasonably uniform in size, so that bursts of a few dozen photons can be counted with good pulse height discrimination. Besides the ordinary dead time correction (see Section 7.3.1), there is a small nonlinearity due to segments with multiple events. However, the main drawback of APDs for photon counting is their very high dark count rate—something like 50 MHz/cm^2 for an APD (Hamamatsu S10362-33-100C 3 $\times$ 3 mm segmented APD) versus 30 Hz/cm^2 (Hamamatsu H9319-01/11 25 mm PMT module), more than six orders of magnitude worse. This restricts photon counting APDs to very small areas.

APDs should be avoided if possible, since a PIN diode operating anywhere near the shot noise limit will always be superior, as well as cheaper and much easier to use. Unfortunately, we can't always have all the light we'd like, and when we can't use coherent detection, APDs can be a big help in the nasty 10 pA to 1μA region.

[†]A. Spinelli, L. M. Davis, and H. Dautet, Actively quenched single photon avalanche diode for high repetition rate time gated single photon counting. *Rev. Sci. Instrum*. **67**, 1 (January 1996); A. Lacaita et al., Performance optimization of active quenching circuits for picosecond timing with single photon avalanche diodes. *Rev. Sci. Instrum*. **66**, 4289–4295 (1995).

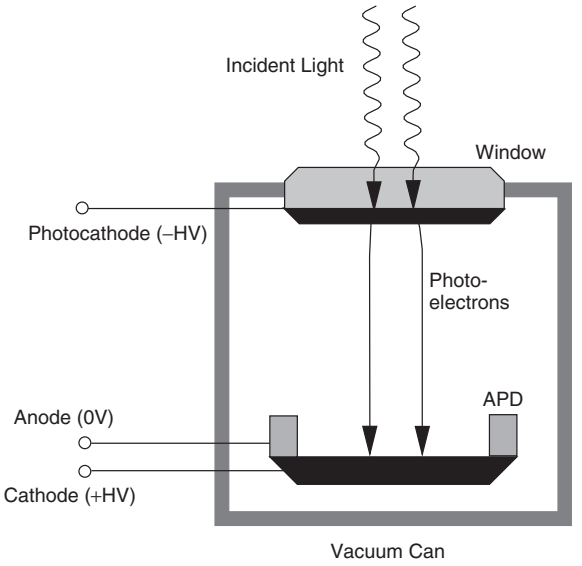

Figure 3.7. The vacuum APD (or hybrid PMT), an imaginative cross between APD and PMT; the high energy of the photoelectron hitting the APD produces a very well-defined pulse.

3.6.5 Vacuum APDs

An imaginative hybrid between the PMT and APD has been developed, the vacuum APD or hybrid photomultiplier (Figure 3.7). It consists of a photocathode and a silicon APD sealed into opposite ends of a vacuum tube, with a high potential (5–10 kV) applied between them. A photoelectron emitted by the photocathode slams into the APD surface with enough energy to generate as many as 2000 carrier pairs. Combined with avalanche multiplication, the overall gain is 10^5–10^6, easily enough for photon counting. Because of the huge electron gain due to impact (equivalent to the first dynode gain in a PMT), the pulse height is much better controlled than in either a PMT or a regular APD, and is much larger than APD dark pulses, so the two are easily distinguished by thresholding.

You can also get these with regular PIN diodes as the anode, and those have the usual PIN advantages of stability and low noise, at the expense of gain. Either type gives a tight enough pulse height distribution that you can accurately compute the number of photons in a burst, at least up to about 10 photons or so. This is important, because it extends the usefulness of photon counting measurements to higher flux levels. The disadvantages of VAPDs include high cost, photocathode inefficiency, and the requirement of two very high voltage (+2 kV and −10 kV) power supplies. Still, these devices appear to be the state of the art for nonimaging detectors with gain.

3.6.6 Photoconductors

True photoconductive detectors must be distinguished from photodiodes operated at reverse bias, in the misnamed "photoconductive mode." The two bear a superficial resemblance, in that the same circuit is used for both, but their operating principles and performance are quite distinct. A true photoconductor is a resistive chip or film made of a substance whose conductivity changes when it is illuminated, due to the generation of

electron–hole pairs when photons are absorbed. If a bias current is applied, the change in conductivity gives rise to a voltage change between the terminals.

Except in special circumstances, such as far-IR detection, where no good photodiodes exist, photoconductors are a poor choice of detector.

An ideal photoconductor is 3 dB noisier than an ideal photodiode; in a photoconductor, recombination becomes a Poisson process with the same variance as generation, so that the noise power doubles. More than that, though, photoconductors exhibit a fundamental trade-off between sensitivity and speed that photodiodes are free of. Once a carrier pair is generated in a photoconductor, it contributes to the conductivity until it recombines. Recombination can occur at the surface or in the bulk; surface recombination is usually faster, so that at short wavelengths, where the carriers are generated near the surface, the responsivity may drop.

A photoconductor exhibits gain equal to the ratio of the carrier lifetime τ to the transit time τ_{tr}, since it basically gets to reuse the carriers M times, where

$$M = \frac{\tau}{\tau_{tr}}. \tag{3.17}$$

This can be increased by increasing the lifetime or shortening the transit time. This relation appears very favorable, since it allows the possibility of useful gain with noise lower than that of PMTs or APDs (since the multiplication contributes no additional noise except for recombination). Unfortunately, the carrier lifetime limits the speed of the device, and so it must be made short for practical devices.

The only common photoconductors used in the visible are cadmium sulfide (CdS), cadmium selenide (CdSe), and their offspring, CdSSe. Besides being very slow, these devices have a sharply peaked spectral response that depends strongly on processing, a memory effect that can lead to factor-of-5 changes in cell resistance due to previous history, and a response time that depends on the illumination level; on the other hand, they exhibit a truly gigantic photoconductive response, which is helpful in applications such as night lights, where the amount of circuitry must be minimized. They are not terribly useful in high performance applications.

Infrared photoconductors are a bit better, but their main attraction is that they exist in a field with limited alternatives. They include HgCdTe, whose cutoff wavelength can be tuned from 5 to 22 μm by adjusting the Hg/Cd ratio; InSb, useful to 5.5 μm; and lead salts PbS and PbSe, which are very slow and should be avoided if possible. Most of these require cryogenic cooling for good performance, which tends to cost thousands of dollars per unit. Room temperature photoconductors are available out to 15 μm, but are so insensitive that they are good only for detecting laser beams.

Far-infrared (>20 μm) photoconductors are also available; these are made from doped silicon and germanium. Due to dopant solubility limitations, they are nearly transparent to the radiation they are designed to detect, which seriously limits their quantum efficiencies; they also require cooling to liquid helium temperatures, and cost $10,000 or more.

Designers who find themselves thinking that there must be an easier way should consider pyroelectric and thermal detectors. Photoconductor performance varies widely; consult the manufacturers' data sheets.

3.7 THERMAL DETECTORS

Not all photodetectors are based on quantum effects; the other major class is thermal devices and includes bolometers, thermocouples and thermopiles, pyroelectric detectors,

and mixed technology devices such as Golay cells. These devices are characterized by room temperature operation, low sensitivity, broad wavelength response, and low speed (except for some pyroelectrics).

A bolometer is a device whose resistance changes with temperature, such as a carbon thermistor or platinum film, and which has a very black surface. It is limited by the Johnson noise of its resistance and (if care is not taken) by the noise of its bias current and by thermal gradients. Increasing the bias current improves the sensitivity, but self-heating effects set a limit. Bolometers are used in a bridge configuration with a second, similar element shielded from incident radiation, to provide compensation for ambient temperature variations.

Thermocouples and thermopiles were the first IR detectors, dating from the 1830s. These devices do not require bias currents, as they generate their own output voltage from the temperature difference between two junctions of dissimilar metals or semiconductors. A thermopile is a series array of thermocouples, which gives a higher output voltage.

Bolometers and thermocouples are frequently used in spectrometers, where their room temperature operation and wide wavelength range are important advantages. Superconducting bolometers exploit the very large temperature coefficient of a superconductor near its transition temperature; they are extremely sensitive but obviously cumbersome.

Infrared bolometer arrays are becoming popular, in spite of poor sensitivity, because room temperature operation makes them much easier to use than cooled quantum detectors; they are somewhat cheaper as well.

Pyroelectric detectors are basically capacitors made of lithium tantalate ($LiTaO_3$), molecular crystals such as triglycine sulfate (TGS) and its derivatives,[†] or ferroelectric plastics such as polyvinylidene difluoride (PVDF), which is basically fluorinated Saran Wrap. By *poling* the material (causing a net dielectric polarization to be "frozen in"), the dependence of the polarization on temperature can be converted to a surface voltage change. High impedance AC amplifiers are used to detect these changes. Pyroelectric (PE) detectors do not work at low frequency, so PE detectors require their inputs to be modulated (e.g., with a chopper). Pyroelectrics work at room temperature but are not as sensitive as cooled semiconductor detectors. Their response increases as their Curie temperature (where they depole spontaneously) is approached. These devices have a very wide range of speeds and sensitivities, from submicrosecond devices used with pulsed lasers to pyroelectric vidicons that can see small (0.1–0.5 °C) variations around room temperature. Circuit hacks can get very competitive sensitivity (~0.13 K NEΔT) in a low resolution image sensor very cheaply (see Section 3.11.16 and Example 17.1).

3.8 IMAGE INTENSIFIERS

3.8.1 Image Tubes

These are two kinds of spatially resolved image amplifiers. Image intensifier tubes use a photocathode and a scintillator in a vacuum tube with an electrostatic lens in between, to make a single photoelectron emitted from the photocathode produce many photons from the scintillator (up to 2×10^3 in one stage). These can be detected directly or can be run into another image intensifier tube. Stacks of three tubes have been popular, requiring

[†]Philips used to use *deuterated triglycine fluoroberylate* in their PE vidicons. To the author's knowledge this stuff has the most jaw cracking name of any material in practical use.

power supplies of 15–50 kV at low current, and yielding photon/photon gains of 5×10^4 or so.

Image tubes are somewhat inconvenient to use, since electrostatic lenses have curved focal surfaces, so that the tube faces must be convex. Controlling field curvature requires negative lenses. Laplace's equation forbids this, because the field would have to be a maximum away from the boundary. This is a particularly inconvenient shape to match to the focal surface of a lens, which also tends to be convex (in the other direction of course). Even when a fiber optic face plate is used to flatten the focal surface, image tubes have relatively poor resolution. The image intensifier tube is not used as widely as it was 20 years ago, because of the advantages of microchannel plate (MCP) image intensifiers.

3.8.2 Microchannel Plates

A microchannel plate is a slab of glass containing a close-packed hexagonal array of small holes or channels, 5–15 μm in diameter, running from one face right through to the other. It is typically made by fusing together a bundle of specially made optical fibers, whose cores are of an easily etched glass. After fusing and slicing, the cores are etched away, leaving channels. The channels are lined with a thin layer of an electron multiplication material like that used in photomultiplier dynodes. A small current flows along this layer, allowing the layer to replace both the dynodes and the bias string of a PMT. A few kilovolts' bias is applied across the thickness of the plate, so that due to the voltage drop across a strong potential gradient exists along the length of the channel. An electron hitting the wall of the channel gives rise to a few secondary electrons, which cascade down the length of the channel, hitting the walls and resulting in a large multiplication factor from face to face. The length of the channels is typically 50 times their diameter.

A MCP image intensifier has a photocathode near one face and a scintillator near the other; due to the electron multiplication, a much higher photon–photon gain ($>10^4$) can be realized in a single stage with MCP, so multiple stages are usually unnecessary. The spatial resolution of a MCP is much better than that of an image tube, since it is defined by the geometry of the channels, rather than that of a rather fuzzy electrostatic lens. Microchannel plates are vulnerable to ion events; a straight MCP channel is an effective ion accelerator. Modern MCPs have their channels at an angle to the slab faces or arranged in a chevron or curved ("J") arrangement. This increases the gain and reduces its spread, since it tends to force all electrons to hit the wall near the top of the channel. In addition, tortuous channels help force any ions to hit the sides instead of being launched into the photocathode.

MCPs are extremely fast; typical rise times for MCP electron multipliers are around 100 ps, and fall times 100–500 ps. The transit time spread is usually a few times less than the rise time. They are also very linear for channel currents up to perhaps 10% of the bias current per channel, at which point signal-dependent voltage drops along the channel begin to modulate the gain. Their multiplication noise is rather worse than a conventional dynode-chain multiplier's, because the secondary electron yield depends more on the trajectory of the electrons.

You can get MCPs with up to 64-segment anodes for time- and space-resolved photon counting applications. Their time resolution is excellent ($\sim$40 ps RMS) but they take a lot of circuitry.

Due to their huge internal surface area and constricted channels, MCPs cannot be baked out very well. Their interiors are thus much dirtier than those of PMTs, which limits their lifetime to about 0.2–0.5 coulomb of anode charge per square centimeter, *about 1000 times shorter than that of a regular PMT*.[†]

3.8.3 Streak Tubes

The high gain and fast shuttering of MCPs has one major disadvantage; all the photons from all times are superimposed on one another, so that there is no way of discovering the time evolution of the signal afterwards. One can use several MCPs, or make the signal periodic and use stroboscopic sampling, but these are not always possible; fortunately, there is a better way: the streak tube.

A streak tube is nothing more than an image tube with deflector plates in it, so that the image can be steered about on the output phosphor. A one-dimensional (1D) line image can be scanned across the output phosphor to produce a two-dimensional (2D) grey scale picture of the time evolution of every point in the line image—much like a 2D optical oscilloscope, but with a time resolution down to 1 ps. The moderate photon-to-photon gain of a streak tube, 10–500, is useful, because a signal has to be pretty intense to produce many photons in a picosecond. Fast-responding photocathodes such as S-20 are *de rigueur*—an NEA photocathode will smear the time response out to 1 ns or even slower. Electron storage materials can be used to make IR streak cameras. Streak cameras tend to cost $100k or so, but they're 30 times faster than the fastest digitizing scopes.

3.9 SILICON ARRAY SENSORS

3.9.1 Charge-Coupled Devices

Silicon photodiodes are very good detectors in the visible and are made on the same substrates as ICs; thus they are natural candidates for making imaging array detectors. In operation, each array element operates as an isolated photosensor, accumulating charge in a potential well for some integration time, after which it is read out. The classical silicon array detector is the charge-coupled device (CCD). In CCDs, the readout mechanism is destructive but quiet; the charge from each element is shifted out by a clever analog shift register technique analogous to a worm screw feed, until it reaches a device lead or an on-chip sense amplifier. CCDs usually have on-chip amplifiers, which have enough gain that the subsequent stages are easy to design (one to a few microvolts per electron is typical).

The limits on CCD performance are set by photon efficiency, dark current, readout noise, and charge transfer losses. Provided the CCD is not shifted too rapidly, transfer efficiency will be 99.99% per stage at least. (This sounds great, but remember that this is raised to the 512th or even 4096th power—in a good device, you can get 99.9999% by slowing down a bit, which is dramatically better.) Nearly all the charge lost in transfer winds up in the following pixel, so that a bit of postprocessing can help.

All but the most expensive CCDs have some bad pixels, which have anomalously low sensitivity (*dead pixels*) or high dark current (*hot pixels*). Really cheap CCDs have whole dead columns. Your processing strategy has to take these into account.

[†]Philips Photonics, *Photomultiplier Tubes: Principles & Applications*, 1994, p.1–21.

3.9.2 Types of CCD

The two major classes of area-array CCDs are interline transfer (ILT) and frame transfer (FT). In an ILT CCD, the columns of sensor pixels are interdigitated with readout columns. When the integration time is finished, the entire image is transferred into the readout columns and is transferred out to be read. The areal efficiency of ILT devices is poor, 25–30%, since the readout columns take up space in the focal plane. On the other hand, the very fast lateral transfer makes these devices good shutters. The sensitive area in ILT CCD pixels is often weirdly shaped, which make its Fourier-domain properties less desirable, and the small size and odd shapes lead to serious moiré effects when imaging strong geometric patterns. Newer ILT CCDs having microlenses deposited on the top can have fill factors nearly the same as FT devices, with the attendant FOV reduction.

A frame transfer CCD has no interdigitated columns. When it is read out, the columns are shifted down into a shielded area array and read out from there. This keeps the fill factor high but requires hundreds of times longer to get the last of the pixels shifted into the dark array, so the shuttering performance is correspondingly worse. On the other hand, the fill factor can be 100%, and the square, full-pitch pixels reduce the moiré effects.

3.9.3 Efficiency and Spectral Response

Some CCDs are cheap, being produced in huge volumes for video cameras and other applications (usually with many dead pixels). These have poor long wavelength response, some cutting off at 600 nm or even shorter, although Sony makes some that work at some level out to 1000 nm. Front-illuminated CCDs are also insensitive in the blue and UV due to absorption of the films on the top of the silicon. Antireflection coating helps somewhat.

Illuminating the CCD from the back helps a lot, because of the greater absorption depth available and the absence of top surface obstacles. This requires thinning the die drastically, which is extremely expensive and hurts yield; nevertheless if you have the money, you can get CCDs whose fill factor is 100% and whose QE exceeds 85% after AR coating.

Typical commodity black-and-white CCDs have $\eta \approx 60\%$ and fill factors of 25%, so that their overall efficiency is only 15% or thereabouts. These specifications must be taken into account when designing a CCD detector subsystem. Some devices are becoming available with microlenses on the chip, to gather more of the incoming light into the active area, which helps a lot but causes spatial pattern problems, especially etalon fringes with lasers (see Section 3.9.6).

3.9.4 Noise and Dark Current

The readout noise of a CCD is additive and can be very low: as low as 2 electrons per pixel in a cooled device with a full well capacity of 10^6 electrons, read out at 20 kHz or so. Floating-gate readout devices have demonstrated sub-electron noise levels but are not commonly available. Over the past 20 years or so, commercial CCD readout noise has been stuck at the few-electron level.

There is a sharp trade-off between speed and readout noise; going faster requires more bandwidth, and the noise power scales linearly with speed. It's possible to achieve sub-electron noise by sampling slowly enough, but then it takes forever to read out the

camera. More commonly, the readout noise of a decent camera is around 30 electrons RMS. Achieving this noise level requires eliminating the *kTC* noise (see Example 13.3) contributed by the reset operation, which is done by sampling the output level before and after each charge readout, and subtracting the two voltages, a procedure called *correlated double sampling*. (Interestingly, since no charge is dissipated in the shifting operations, no *kTC* noise is introduced by shifting the charge packet into the readout cell.)

Increasing the integration time increases the signal linearly, so the electrical SNR of low light images goes as t^2, until the dark current noise $(i_{\text{dark}}t)^{1/2}$ equals the readout noise, after which the SNR is shot noise limited and so increases only linearly with time. The attainable (electrical) SNR is limited by photon statistics to a value equal to the full well capacity in electrons, around 47 dB for the camcorder CCD to perhaps 60 dB for a scientific device.

A garden-variety camcorder CCD operated at room temperature has a dark current of about 100 electrons per pixel in a $\frac{1}{30}$ second integration time, and a well capacity on the order of 5×10^4 electrons. A cooled CCD, as used in astronomy, can achieve dark currents far below 1 electron/s per pixel. These very low dark currents are achieved by multiphase pinning (MPP), which eliminates the effects of surface states (MPP is also called *inverted mode*). Since the dark current is thermally activated, it drops by a factor of 2 every 10 °C. Cooling cannot be taken too far, however, due to the "freezing out" of trap sites, which leads to serious charge transfer losses at low temperatures (−30 °C for some devices, to −100 °C for others). Freeze-out is not subtle; you get severe streaks along the transfer direction.

3.9.5 Bloom, Bleed, and Fringing

Prominent CCD pathologies are bloom, bleed, and fringing. Bloom is the picturesque name given to the artifacts that result when some elements become full and spill over into the readout circuitry or the substrate, causing spurious lines and a general leakage of charge into adjoining elements. A badly bloomed image is full of bright lines and blobs. Bloom is controlled by additional electrodes that extract the charge before it can migrate far; antiblooming compromises linearity, fill factor, and full well capacity, so use it only if you really need it. Note that bloom is not necessarily restricted to illuminated areas; by dumping charge into the substrate, saturated pixels can send charge into storage areas as well, a particular problem in scientific applications, when the readout time is sometimes long compared to the integration time.

Fringing and bleed cause loss of resolution in CCDs operated at wavelengths near the 1.1 μm absorption edge of silicon. The silicon becomes increasingly transparent, and light can bounce back and forth between the front and back surfaces of the wafer, causing nasty irregular fringes and scatter. This is not a trivial effect: at short wavelengths the fringes tend to be 1% or below, but in the IR they can get up to 5% or even more, which is very objectionable.

If we try to make the silicon thicker, so that more absorption occurs, lateral diffusion of carriers in the field-free region (away from the depletion zone of the junction) causes bleed, where carriers from one pixel diffuse into neighboring ones. The best solution to this is to use a thick piece of high resistivity silicon, which can be depleted throughout the volume (like a back-biased PIN diode). High resistivity CCDs have high quantum efficiency even at 1000 nm. If you haven't got one of these special devices, there's not a lot you can do about it.

3.9.6 Spatial Pattern

CCDs also have stable, accurate spatial patterns; subpixel position measurements of images a few pixels in size can be done by interpolating. Carefully designed, back-illuminated CCDs have high fill factors, which makes their Fourier domain properties simple, but front-surface devices are much iffier.

Fancier Fourier processing techniques should not assume that the spatial sensitivity pattern of a PMT pixel looks like a nice rectangular box, even if the pixel is rectangular: most of the time, it has sloping sides and a bit of dishing in the middle, so that it resembles a fedora more closely than it does a top hat. Some types have serious asymmetry between two halves of the pixel (e.g., when the sensitive area is L-shaped). These effects make the optical transfer function of your CCD do things you might not expect; for example, the zero in the optical transfer function of the CCD is not at 2π/pixel width but is somewhat further out; irregular pixel shapes make it even worse because of their high spatial harmonic content. The only defense against this in serious Fourier processing is to oversample by a big factor (say, $4\times$ to $8\times$ the Nyquist limit), so that the big irregularities in the OTF happen out where you don't care about them, and you can correct for the small ones that remain. Devices that are not efficiently AR coated will do unintuitive things, because of reflectance changes with incidence angle. For an air–silicon interface, the normal incidence reflectance is 0.3, whereas at 30° it's 0.36 for s and 0.26 for p. Measuring pixel sensitivities directly with a high NA flying spot will thus produce some funny results due to this asymmetric pupil apodization.

At long wavelengths, bleed spreads incident light into adjacent pixels and causes the OTF to be seriously wavelength dependent; how much this happens depends on the absorption of the silicon.

CMOS image sensors and interline transfer CCDs with microlenses exhibit horrible etalon fringes when used with temporally coherent sources.

3.9.7 Linearity

Unlike imaging tubes such as vidicons, CCDs are normally extremely linear, although the antiblooming provisions of some devices can cause serious nonlinearity above about half-scale. Infrared focal plane arrays are much less linear and usually require multiple calibrations at different light intensities.

At low temperatures and long wavelengths, front-illuminated CCDs exhibit QE hysteresis due to trapping of charge at the interface between the bulk and epitaxial layers (this is polished away in back-illuminated CCDs). The CCDs used on the Hubble, Galileo, SXT, and Cassini space missions had QE variations as much as 10% depending on signal level and operating temperature.[†]

Aside: CCD Data Sheets. From an instrument designer's viewpoint, the worst thing about CCDs is their data sheets. Op amps, microprocessors, diode lasers—all these have reasonably standard spec sheets, but not CCDs. Designing a CCD system that will be replicated more than a few times depends on the designer having a deep knowledge of the details of each kind of CCD considered. The data sheets are also nearly all hopelessly out of date. Consult the manufacturer of your devices for the latest specs—and don't be

[†]J. Janesick, posted to the CCD-world mailing list http://www.cfht.hawaii.edu/~tmca/CCD-world/, March 18, 1999.

too surprised if they pay more attention to their camcorder-building customers than they do to you.

3.9.8 Driving CCDs

CCDs take a lot of circuit support. This is somewhat specialized, so if you're rolling your own CCD drivers, you probably want a copy of Janesick. In general, CCDs are forgiving of mildly ugly clock signals as long as they're highly repeatable. People often use transmission gates connected to well-filtered analog voltages to produce the funny clock levels required.

3.9.9 Time Delay Integration (TDI) CCDs

Linear CCD sensors are used in line-scan cameras and in pushbroom-style remote sensing satellites, where the spacecraft motion supplies the frame scan. The SNR can be improved by *time-delay integration* (TDI), in which the linear array is replaced by a narrow area array (perhaps 4096×64 pixels) and clocked in the narrow direction at the same rate that the image crosses the detector, so that the same object point illuminates the same bucket of electrons throughout. This requires accurate alignment and places severe constraints on the geometric distortion of the imaging system, but a 64-deep TDI sensor can get you a 36 dB signal increase.

Another TDI advantage is uniformity. Because each object point is imaged by (say) 64 pixels in turn, the fixed pattern noise tends to average out. There's a bit more in Section 10.5.3.

3.9.10 Charge-Multiplying CCDs

Because the charge transfer efficiency of a CCD is so high, and its dark current can be made very low, its major noise source is its output amplifier. The noise can be reduced to subelectron levels by slowing the readout clock and filtering or averaging the output. The noise/speed trade-off prevents the use of CCDs for real-time imaging at low light levels, as we've seen. Hynecek[†] and more recently Mackay et al.[‡] have made a clever electron-multiplying CCD that overcomes this trade-off almost completely, the *low-light-level CCD* or LLLCCD (or L^3CCD). This device can be used either as a normal CCD of quantum efficiency η, or as the equivalent of a noiselessly intensified CCD with a QE of $\eta/2$ and a readout speed of 20 MHz. The way it works is to take an ordinary single-output CCD and add a few hundred extra transfer stages before the readout amplifier. In the extension, one of the three readout phases is run at a much higher voltage, enough that there is a small amount ($\approx 1-2\%$) of electron multiplication in each stage. Because there are many stages, reasonably well-controlled gains from 1 to several thousand are easily realized by changing the voltage in the one phase, so the same device can go from sunlight to photon counting by changing one voltage. Since the multiplication occurs inside the silicon, there is no dirty vacuum to limit its lifetime. The complexity increase is much smaller than that of an MCP. Furthermore, blooming

[†]Jaroslav Hynecek, CCM—a new low-noise charge carrier multiplier suitable for detection of charge in small pixel CCD image sensors. *IEEE Trans. Electron Devices* **30**, 694–699 (1992).

[‡]Craig D. Mackay, Robert N. Tubbs, Ray Bell, David Burt, Paul Jerram, and Ian Moody, Sub-electron read noise at MHz pixel rates. *SPIE Proc*. January 2001.

of the multiplication stages automatically protects against damage from bright lights that would reduce an MCP to lava.

The LLLCCD needs cooling to get its dark current low enough to count photons. In addition, its dynamic range is limited by the full well capacity of the multiplication stages, and multiplication causes the shot noise to go up by 3 dB, which is actually amazingly good. In a PMT, secondary emission is a Poisson process, so 512 dynode stages, each with a secondary electron yield of 1.01, would produce a pulse height histogram that looked like a pancake (its variance would be about 100 times the mean; see the chapter problems in the Supplementary Material). Electron multiplication in the CCD involves a Poisson process too, with one key difference: with 512 stages of 1.01 gain, the 1 is deterministic and hence noiseless—only the 0.01 is Poissonian, so the variance is only twice the mean, not 100 times. (Why?)

There is a slight linearity trade-off involved at very high gains—bright pixels see a somewhat *higher* gain than dim ones, the opposite of the usual situation. The error can be a factor of 2 in really bad situations, but as it's monotonic, with care it can be calibrated out. A significant advantage of L^3CCDs is the ability to achieve high frame rates, because the usual trade-off of readout noise versus frame rate has been greatly improved by the amplification ahead of the readout amplifier. Because of their sensitivity, flexibility, potentially low cost, and long life, electron-multiplying CCDs will probably replace image intensifiers in applications where cooling is feasible and fast shuttering is not required.

For applications such as astronomical spectroscopy, where the 3 dB SNR loss is too great, L^3CCDs can be operated in photon counting mode, provided the frame rates are high enough that you don't lose too many counts due to two photons hitting a given pixel in the frame time. Because the device operational parameters don't need to change between analog multiplication and photon counting modes, you could in principle change the voltage on the multiplying phase during readout, which would give different pixels different gains, increase the dynamic range.

3.9.11 Charge Injection Devices (CIDs)

CIDs are like CCDs only backwards: the well starts out full, and light removes charge instead of adding it. This adds noise at low light levels but makes CIDs intrinsically resistant to bloom, and so suitable for high contrast applications. Their quantum efficiencies are typically fairly low, around 30%. CIDs use multiplexers instead of shift registers and can be read nondestructively, since the charge need not be removed from the element during readout. CMOS imagers also use big multiplexers (see below). These multiplexer-based technologies offer random access, so we can use different integration times for different pixels on the array. This means that an image with extremely high contrast can be read adaptively, yielding the best of all worlds: rapid acquisition of bright objects and long integration times to enhance detection of faint ones. Since CMOS imagers don't exhibit bloom either, CIDs are no longer so unique in that way.

3.9.12 Photodiode Arrays

Photodiode arrays (commonly known as "Reticons" for historical reasons) look like CCDs but are actually read out via a big multiplexer instead of a bucket brigade, which makes them much easier to control. They are competitive only in 1D arrays, as in OMA

spectrometers, but they work very well for that. Compared with CCDs, they are not limited by charge transfer inefficiency, which suits them well to high contrast applications, where the data will come under close scrutiny, as in spectroscopy. They are a bit less sensitive than CCDs of the same pixel size, however, because of multiplexer noise and charge injection. More insidiously, they are dramatically less linear than CCDs.

Photodiode arrays, unlike CCDs, generally have no bias applied to the diodes during integration; thus they are somewhat nonlinear at large signals, because of the forward conduction of the photodiodes. The forward voltage drops and the dark current increases with temperature, so the linearity, full well capacity, and available integration time all degrade more quickly with temperature than in a CCD. A 10 °C increase doubles the dark current and reduces the charge capacity by about 25%.

3.9.13 CMOS Imagers

The CMOS imager is similar in performance to an interline transfer CCD but is manufactured on an ordinary CMOS process, as used in logic and memory ICs. It consists of an array of pixels, each with its own buffer amplifier. The amplifiers are connected to a gigantic crossbar multiplexer, as in a dynamic memory IC. Often there is a single stage of CCD-style charge transfer between the pixel and the amplifier, to provide the fast shuttering capability of an ILT CCD. The attraction of a CMOS imager is that lots of the ancillary circuitry required for a CCD, such as clock generation, amplification, A/D conversion, panning, zooming, and even image processing (e.g., halftoning), can be done on the imager chip. The optical characteristics are not generally as good as CCDs; the dark current is generally higher, and the linearity and accuracy less. The large number of amplifiers all have slightly different offset voltages, so that there is a lot of fixed-pattern noise in CMOS imagers that is not present in CCDs, which have many fewer amplifiers (often only one). This makes their dim-light performance poor, but work is underway to bring these devices to the capabilities of scientific CCDs. Being able to use imaging sensors without dragging along six tons of ancillary hardware is a powerful and useful capability, but on the other hand the flexibility of custom controllers is sometimes invaluable.

CMOS imagers often exhibit severe fringing due to reflections between the fairly dense metal wiring layers and the silicon surface, and microlenses make it worse. Effective optical path lengths are in the 10 μm range. This isn't too terrible in wideband applications, but it can make narrowband measurements and spectroscopy exciting. On the other hand, since there are real FETs between each CMOS pixel and its neighbors, CMOS imagers generally don't bloom.

3.9.14 Video Cameras

> Television is a vast wasteland.
>
> —Newton N. Minow (then Chairman of the US Federal Communications Commission)

Video cameras use CCDs or CMOS imagers, but most of them are poorly suited to precise measurements. They have black level and gain adjustments (usually automatic) that often foul up measurements by changing when we want them to keep still. In order to mimic the behavior of photographic film and vidicons, their response is deliberately

made nonlinear, sometimes with an adjustment for γ, the contrast exponent. There are cameras made for instrumentation use, and they work well enough but tend to cost a lot compared to the entertainment-grade ones. Since gain and black level adjustments are done on the full frame, putting an intensity reference in the field of view will sometimes allow you to fix it afterwards, but the γ problem is still there.

Cameras are useful in full field interferometric techniques (phase shifting, moiré, and holographic) and in structured light measurements (see Section 10.5.9). On the other hand, cameras are often overused by people whose primary strength is in software, and who want to get the data into digital form as early as possible. The importance of getting the detector subsystem right cannot be overemphasized, so this by itself is an inadequate reason for using video.

Sometimes video is necessary, for example, in machine vision systems or where the requirement for using commercially available hardware is more compelling than that for optimizing performance. In the long run, however, it's such a headache that it is vitally important to make sure you really understand why you're using video rather than one or a few silicon photodiodes. Its cost and complexity rise very quickly once you get past webcam quality, and even so the measurements are normally poor, due to the low SNR of image sensors, the 8 bit limit of the A/D converters, and the generally poor fidelity of the electronics used.

Color video is even worse than black and white. Color sensors are made by putting arrays of different-colored filters on top of a monochrome sensor.[†] These filters are generally arranged in groups of four, one pixel each of red and blue, and two of green. In order to prevent bad moiré effects, color sensors have "defuzzing filters," which are thin walkoff plates (see Section 6.3.5) mounted over the CCD to smear out the image over a four-pixel area. Of course, this reduces the resolution by a factor of 2 in each direction, and equally of course (marketing departments being what they are), the quoted resolution is that of the underlying sensor. When comparing color to monochrome, remember that it takes four *Marketing Megapixels*™ to make one real monochrome megapixel. Avoid color sensors for instrument use whenever possible.

3.9.15 Extending the Wavelength Range: CCDs + Fluors

The UV performance of a front-illuminated CCD detector can be dramatically improved by applying a thin coating of a fluorescent material to convert incident UV photons into visible light before they are absorbed. The quantum efficiency of this approach is around 10% for many fluors, which is not as high as we'd like but a lot better than nothing, which is what we'd get otherwise.

3.9.16 Electron Storage Materials

You can get near-IR sensor material for converting 0.8–1.6 μm light to visible. The best is Q-42 phosphor from Lumitek Corp (available as IR sensor cards from Lumitek and Edmund Optics, since Kodak went out of the business). It uses rare-earth oxides or sulfides in a calcium sulfide matrix and works by electron trapping: visible or UV excites the Ce^{2+}, Er^{2+}, or Sm^{3+} ions to a long-lived metastable state; when an IR photon

[†]There are honorable exceptions, in which dichroic prisms are used to form RGB images on three separate CCD chips, but you'll probably never see one.

comes along, it is absorbed, and a visible one emitted. The fluorescence dies away very rapidly (picoseconds to nanoseconds) when the IR stops, and the quantum efficiency is near unity. This material could probably be used to extend silicon CCD detectors out to 1.6 μm, in much the way as the UV fluors. It would obviously have to be pumped with a wavelength to which the CCD is insensitive, or refreshed during readout like dynamic memory. For near-IR applications at high power density, you can also get second-harmonic generating material, such as nitrofurazone (5-nitro 2-furaldehyde semicarbazone)[†] (5-nitro 2-furaldehyde semicarbazone, Aldrich Catalog #73340), which can be mixed with paint. Water and polar solvents deactivate it, but it survives well in anisole and NMP (n-methyl pyrridone).

3.9.17 Infrared Array Detectors

Platinum silicide arrays are made on silicon substrates, but other infrared arrays are odd hybrid devices, generally consisting of detector chips made of InGaAs (to 1.7 or 2.2 μm), InSb (to 5 μm), or HgCdTe (2.5–14 μm) bump-bonded to Si readout chip. Since the chip metal goes on the top, the chips are bonded face-to-face. Thus the light comes in through the substrate of the photodetector chip, which is the origin of the short-wavelength cutoff in most IR imaging arrays (e.g., the absorption edge at 850 nm due to the CdZnTe substrates used for many HgCdTe detector arrays). Nowadays more of these devices are being back-thinned, so it is possible to get InGaAs detectors with response as far down as the ultraviolet (<400 nm). The long-wavelength edge of HgCdTe is tunable from 2.5 μm to 14 μm by changing the alloy ratios, with the leakage becoming worse for longer wavelength cutoff, as the bandgap decreases. Near-IR sensors such as InGaAs as well as lower performance devices such as pyroelectrics and microbolometer arrays can work at room temperature, but all the others require cryogenic cooling.

IR arrays are much less linear than silicon CCDs, and at the current state of the art, their dark current and response nonuniformities are much worse. Thus calibrating an IR array isn't the simple matter it is with silicon CCDs (see Section 3.9.19). Even with pixel-by-pixel correction for gain and offset, the detection performance of IR arrays is usually limited by their residual fixed-pattern noise. Using calibrations taken at several radiance levels, and fitted with low-order polynomials or sums-of-exponentials, can greatly (20–30 dB) improve matters, at least within the range measured. Platinum silicide arrays have enormously better uniformity than InSb and HgCdTe arrays, so they often offer effectively lower noise in the mid-IR even though their QE is very poor. Pixels with bad $1/f$ noise are usually the limiting factor in the stability of the calibration; PtSi calibrations are good for days, but InSb and especially HgCdTe arrays require recalibration on timescales of minutes to hours, if the spatial noise is going to be below the temporal noise.[‡]

3.9.18 Intensified Cameras

Electron multiplication is often helpful in low light situations, to overcome circuit noise. It is natural to try applying it to imaging detectors such as vidicons and CCDs, to overcome

[†]Used as a topical antibiotic (Furacin), but also good for making 0.5–0.7 μm from 1.0–1.4 μm.

[‡]Werner Gross, Thomas Hierl, and Max Schulz, Correctability and long-term stability of infrared focal plane arrays. *Opt. Eng*. **38**(5), 862–869 (May 1999).

their dark current and readout noise in the same way. Just now, the dominant type of intensified camera is the combination of an MCP image intensifier with a CCD sensor. Older types, such as the image orthicon, the silicon intensifier target (SIT) vidicon, and cameras based on image converter tubes and electron optics are now seldom used (see Section 3.9.10).

An MCP camera consists of a microchannel plate image intensifier whose output is coupled to a CCD imaging detector. These are often proximity focused; the phosphor screen is close to the CCD, making an extremely compact, simple, and robust system. MCPs have spatial resolutions of 10 μm or so, a good match to a CCD.

Such intensified cameras normally produce noisy output, because generally they can't improve the photon statistics of the incident light. You can't just use long exposures, because each primary may produce 2000 electrons in the CCD well, so you get only a few dozen counts before saturating. Such noisy signals often need frame averaging before they become useful, which limits the utility of intensified cameras. For instrument use, they're usually at a disadvantage compared with cooled CCDs using long integration times. On the other hand, image intensifier cameras are very suitable when the images are primarily intended to be viewed by eye; real-time imaging becomes more important then, and spatial averaging in the human visual system can take the place of frame averaging.

Another reason for using MCP intensified cameras is their very fast time-gating capability; an MCP has a rise time of 100 ps or so, and its gain can be turned on and off in a few nanoseconds, making it an excellent shutter as well as an amplifier. The two methods for doing this are to use an avalanche transistor circuit to gate the whole MCP bias voltage, or to use an MCP with a grid between photocathode and MCP array, which requires only a few tens of volts. Cameras exist that will take a few frames at 10^7 fps.

Aside: Night Vision Goggles. Direct-view image intensifiers, (e.g., night vision goggles) are helpful for a few other reasons. Photocathodes have about four times the QE of rod cells, plus a much wider wavelength band, especially toward the IR where the sky glow is brighter; they bring the cone cells into play, which have much higher resolution and higher response speed; and the collection system can have a much larger étendue than the eye, both in area and in solid angle.

3.9.19 Calibrating Image Sensors

Achieving imaging performance limited by counting statistics is a bit more of a challenge than this, because image sensors have characteristics that vary from pixel to pixel. Each output amplifier will have its own characteristic bias voltage; each pixel will have slightly different sensitivity and dark current. Separating out and correcting these effects is nontrivial.

What we initially measure is a raw data frame R, but what we want is the true image intensity I, corrected for gain and offset. Providing that the sensor is very linear (e.g., a properly operated CCD), the best way is to take long- and short-exposure dark frames, $D_l(t_l)$ and $D_s(t_s)$, and a flat field frame, $F(t_f)$, using the real optical system aimed at a featureless surface such as the dawn sky or an integrating sphere. A bias frame (just the bias voltage) B can be constructed as

$$B = \frac{t_s \cdot D_l - t_l \cdot D_s}{t_s - t_l}, \tag{3.18}$$

a normalized thermal frame (just the dark current in 1 s) as[†]

$$T = \frac{D_l - B}{t_l}, \tag{3.19}$$

and a sensitivity frame as

$$S = F - B - t_F \cdot T, \tag{3.20}$$

where S is normalized to unity at the radiance of the featureless surface. Note that all these frames should really be averaged across at least four separate exposures so that the noise of the calibration frames does not dominate that of the data you're going to take with the system. Make sure that the long-exposure dark frames are at least as long as your longest measurement, and that the flat fields are just below half-scale to guard against nonlinearity. The long-exposure frames will be slightly corrupted by cosmic ray events, so in the averaging operation, code in a check that throws out the highest value in each pixel if it's way out of line (e.g., four times the RMS noise).

When all this work is done, a properly normalized and calibrated true image I can be computed from a raw image $R(t_R)$ of exposure time t_R:

$$I = \frac{(R(t_R) - B - t_R \cdot T)}{S}. \tag{3.21}$$

This is a flexible and convenient calibration scheme, because you can make your exposures any length you need to. You'll need to experiment to determine how often it needs to be repeated. Watch out especially for sensor temperature changes, which will make these calibrations go all over the place. (See the problems for how many of each kind of frame you need.) Above all, make sure that your sensor is always operating in a highly linear regime: if your hot pixels saturate, you can't do good dark frames; if your flat fields are above half-scale and the antiblooming or MPP is on, you'll have the nonlinearity.

The flat field is wavelength sensitive, so make sure you take the flat field frames with a source of the same spectral characteristics as your typical data. Fringing in thinned, back-illuminated CCDs makes the wavelength variation of the nonuniformity sharper than you might expect; the normalized photoresponse nonuniformity (PRNU) of a good back-illuminated CCD is 1% or so in the red, rising to 5–10% in the UV and as much as 20% near the IR cutoff.

As we discussed earlier, IR arrays are much more difficult to calibrate, because they are not as linear, are more temperature sensitive, and are less uniform to begin with.

3.9.20 Linearity Calibration

CCDs are usually pretty linear up to about half the full well capacity, and many are very linear nearly to full well. On the other hand, it is much more comfortable knowing than hoping. One good way to get a linearity calibration is to use a diffused LED source (a few frosted LEDs all around the CCD, at some distance) that provides a nice uniform illumination across the whole chip (a few percent is OK, you can remove that

[†]If you're using integer arithmetic, you'll want to normalize to some longer time interval to avoid significance loss.

mathematically). Start with the CCD reset and shift the image out normally. Flash the LEDs once after each row is shifted out, and you get a nice intensity staircase signal that will tell you a great deal about your linearity curve. This is much faster than doing many, many full frame calibrations, and so can be used as an online linearity calibration. Note that LEDs generally have a temperature coefficient of output power near $-1\%/^{\circ}\mathrm{C}$, so for decent accuracy you have to temperature compensate them or use them in a closed-loop system with a photodiode.

3.10 HOW DO I KNOW WHICH NOISE SOURCE DOMINATES?

The most basic limit to the sensitivity of an optical measurement is set by the shot noise of the signal photons. Once other noise sources have been reduced below this level, further SNR improvements can come only from increasing the signal strength or narrowing the bandwidth. There is a certain cachet to "shot noise limited" or "quantum limited" measurements, which should not be allowed to obscure the fact that such measurements can still be too noisy, and that most of them can still be improved.

The contributions of Johnson noise, signal and background shot noise, and thermal fluctuations are easily calculated from parameters given in typical data sheets; the one major imponderable is lattice (thermal) generation–recombination (G-R) noise in IR photodetectors, which depends on the carrier lifetime, a number that is not always easily available. In most cases, one must use a seat-of-the-pants method of estimating lattice G-R noise, such as taking the published noise specification and subtracting all the other noise sources, or rely on the detector manufacturer's assertion that a detector is background limited at given detector and background temperatures and field of view.

Formulas for noise contributions are often given in a form that explicitly includes the modulation frequency response of the detector. This seems unnecessary. Apart from $1/f$ noise, the detector noise has the same frequency response as the signal, and combining different authors' complicated formulas is cumbersome. The frequency response is a matter of deep concern to the designer, who is unlikely to be misled by, for example, someone describing shot noise as having a flat power spectrum.

The total noise contributed by a detector will depend on just how the detector is coupled to the external circuitry. Calculating this requires some sort of circuit and noise models of the detector. Determining the overall SNR of the signal emerging from the detector subsystem is a major topic of Chapter 18. Table 3.1 is a good starting point for figuring that out.

3.10.1 Source Noise

In many measurements based on externally applied illumination, source noise is the dominant contributor. Great ingenuity is expended on reducing and avoiding it. In laser-based measurements, especially bright-field ones such as interferometry, absorption spectroscopy, and transient extinction, laser residual intensity noise (RIN) is frequently the dominant contributor. It may easily be 60 dB above the shot noise level. Fortunately, it is usually tractable; see Sections 10.6.2 and 10.8.6 for how to handle it. Laser frequency noise must usually be treated by stabilizing the laser.

Noise from incoherent sources, especially arc lamps, is more difficult to deal with since it is strongly dependent on position and angle, so that compensating for it by

TABLE 3.1. Which Noise Source Dominates?

Detector Type	Noise Source	Dominates When[a]	Noise Spectral Density
Si, Ge, InGaAs photodiodes	Photocurrent shot	$i_s R_L > 2kT/e$ (50 mV @ 300 K)	$i_N = (2ei_s)^{1/2}$
	Background shot	$i_b R_L > 2kT/e$	$i_N = (2ei_b)^{1/2}$
	Johnson	$(i_s + i_b)R_L < 2kT/e$	$i_N = (4kTR_L)^{1/2}$
IR photodiodes	Photocurrent shot	$i_s(R_L \parallel R_{sh}) > 2kT/e$ (50 mV @ 300 K)	$i_N = (2ei_s)^{1/2}$
	Photon (background shot)	$i_b(R_{L\parallel R_{sh}}) > 2kT/e$	Eq. (3.10)
	Lattice generation/recombination	Only when reverse biased	Eq. (3.23)
	R_{sh} Johnson	Always unless cryogenically cooled	$i_N = (4kT/R_{sh})^{1/2}$
IR photoconductors	Shot (photogeneration/recombination)	$i_s G(R_L \parallel R_{sh}) > kT/e$ (25 mV @ 300 K)	$i_N = (4Gei_s)^{1/2}$ (recombination doubles variance)
	Lattice G-R	$V_{DC}\tau\mu > \ell^2 kT/e$	Eq. (3.23)
	Photon	BLIP when cryogenically cooled (believe manufacturer)	Eq. (3.10)
	Johnson	Always unless cryogenically cooled	$i_N = (4kT/R_{sh})^{1/2}$
Thermal detectors	Johnson	Nearly always	$v_N = (4kTR)^{1/2}$
	Thermal fluctuations		Eq. (3.24)
Avalanche photodiodes	Shot	Almost never	$i_N = (2Mei)^{1/2}$
	Multiplication	Almost always	$i_N = (2M^{1+x}ei)^{1/2}$
	Johnson	Only if M is too low	$i_N = (4kT/R_L)^{1/2}$

[a]Here i is the actual current from the device (after multiplication, if any).

comparing the measured signal in some way to a sample of the light from the source may not be sufficiently accurate to control the effect. It is possible to stabilize arcs using strong magnetic fields, but as this is awkward and expensive, it is seldom done; the problem remains a difficult one. Incandescent bulbs are much quieter than arcs and are often a good choice.

Source noise is multiplicative in character, since both signal and coherent background are proportional to the source intensity; the multiplication of the source noise times the coherent background level gives rise to the additive part of the source noise, while source noise multiplying the signal puts noise sidebands on the desired signal, a particularly obnoxious thing to do. See Chapter 2 for more on source noise and what to do about it.

3.10.2 Shot Noise

Shot noise is the easiest limit to calculate, but the hardest to improve upon. There is a fair amount of confusion about shot noise, where it comes from, and where it applies.

The formula, which states that if the arrival of electrons at a given circuit point is a Poisson process, the current i will have a noise current spectral density of

$$\langle i_N \rangle = \sqrt{2ei} \text{ A/Hz}^{1/2}, \tag{3.22}$$

is perfectly correct. However, much of the time the current is not Poissonian, so that the formula is inapplicable. The simplest case is in ordinary photodiodes, where all photocurrents and leakage currents exhibit exactly full shot noise, regardless of quantum efficiency.

It is easy to make currents with full shot noise, much less than full shot noise, or much more than full shot noise. A battery connected to a metal film resistor will exhibit much less than full shot noise, because any charge density fluctuations are smoothed out by electron scattering, which tends to reestablish the correlations and reduce the temperature of the electrons; the shot noise power is reduced by a factor of L/l, where L is the length of the resistor and l is the mean free path for electron–electron scattering.[†] Since $l \sim 100$ Å for disordered metals, and $L \sim 1$ mm, the suppression is pretty strong.

An avalanche photodiode's dark current will exhibit much more than full shot noise, since each electron generated in the junction gives rise to a large pulse; even if all N pulses arriving per second were exactly A volts tall, the RMS fluctuation in 1 s will be $A\sqrt{N}$. This is $\sqrt{M}$ times larger than the shot noise corresponding to the average output current.

Shot noise in an optical beam can be viewed as the interference term between a noiseless optical signal and the zero-point *vacuum fluctuations* of the electromagnetic field. It affects both the amplitude and phase of a detected signal; by heroic preparation, in very special circumstances, the noise can be redistributed slightly between two conjugate components (sine and cosine, or amplitude and phase), but the product of the two components cannot be reduced. It is thus a fundamental feature of the interaction of light with matter.

Thermal light, at least at frequencies where $h\nu \gg kT$, is Poissonian to good accuracy; at lower frequencies, where the occupation number of the quantum states is higher, there is an additional term due to the Bose–Einstein statistics of thermal photons; this makes the variance of the photon count N go up by a factor of $1 + 1/[\exp(h\nu/kT) - 1]$.[‡] We are nearly always in the high frequency limit in practice.[§]

It is useful to concentrate on the statistics of the initially generated photocarriers, before any gain is applied, because in all quantum detectors this current exhibits full shot noise, and amplification does nothing whatever to improve the signal to shot noise ratio. Like quantum efficiency, this ratio must be kept in mind; it is a vitally important sanity check.

Carrier recombination is also Poissonian, so the shot noise variance is doubled. This double shot noise in photoconductors is sometimes called photocarrier

[†]There's been a lively literature on the details of this for the last 25 years, and there's a lot of interesting physics there: for example, see R. Landauer, *Ann. New York Acad. Sci*. **755**(1), 417–428 (1995).

[‡]For example, see E. L. Dereniak and G. D. Boreman, *Infrared Detectors and Systems*. Wiley, Hoboken, NJ, 1996, Section 5.2.2.

[§]Hanbury Brown and Twiss demonstrated that these classical fluctuations could be used to measure the angular diameter of hot thermal sources such as blue stars by cross-correlating the measured noise from two detectors whose spacing was varied—a so-called *intensity interferometer*. See Robert Hanbury Brown, *The Intensity Interferometer*. Halsted Press (Wiley), Hoboken, NJ, 1974. Hanbury Brown is one of the present author's technical heroes.

generation–recombination noise, but it passes the duck test[†] for shot noise. This renaming leads to the highly misleading statement that photoconductors do not exhibit shot noise.

A rule of thumb is that if the photocurrent from a photodiode is sufficient to drop $2kT/e$ (50 mV at room temperature) across the load resistor, the shot noise dominates the Johnson noise; in a photoconductor, the required current is reduced by a factor of 2 because of the increased shot noise from recombination.

Comparison of shot noise due to signal and background is easier; because both photocarriers generated by signal and background photons exhibit shot noise, the signal shot noise will dominate the background shot noise whenever the signal photocurrent is larger.

3.10.3 Background Fluctuations

In the absence of other modulation, the photon statistics of otherwise noiseless background light are the same as signal light. Background shot noise will dominate the Johnson noise any time that the background photocurrent drops more than $2kT/e$ across the load resistance for photodiodes, or $kT/$e for photoconductors.

In many instances, the background light is strongly modulated; examples include 120 Hz modulation in room light, and 15.75 kHz from television screens. Furthermore, in many measurements, the coherent background is very important; for example, the ISICL system (a short range coherent lidar) of Example 1.12 encounters unwanted reflections from the plasma chamber walls that are 10^6 time stronger than the desired signal, and which further exhibit strong modulation during scanning due to speckle. A combination of baffles, homodyne interferometry, and laser noise cancellation produce a stable measurement even so.

3.10.4 Thermal Emission

In the mid- and far-IR, say, from 5 to 20 μm, room temperature or thermoelectrically cooled quantum detectors are limited by the Johnson noise of their own shunt resistance, while a cryogenically cooled unit is generally limited by the fluctuations in the background thermal radiation, the so-called *BLIP* condition.[‡] For sufficiently strong signals, the shot noise limit may be reached, but this gets increasingly difficult as the wavelength gets longer, because the decreasing energy per photon makes the shot noise limited SNR higher, and because the thermal background gets stronger. The key design goal while using a BLIP detector is to reduce the detector area and field of view as much as possible, while keeping all the signal photons, and not doing anything silly in the electronics to add significant additional noise. Example 3.2 describes how to calculate the expected noise from an IR photodiode.

3.10.5 Lattice Generation–Recombination Noise

Photoconductors exhibit noise due to random fluctuations in their number of carriers, so-called generation–recombination noise. The last heading dealt with noise due to coupling to the fluctuations in the radiation field; noise also arises from coupling to the

[†]"If it looks like a duck and it quacks like a duck, it's a duck."

[‡]BLIP originally stood for "background-limited infrared photoconductor" but has come to be applied to any detector whose performance is limited by the thermal background.

fluctuations of the lattice vibrations. Since this is basically noise in the conductivity, it causes noise mainly in the bias (dark) current. Thus it does not strongly affect photodiodes, which are normally run without a bias current.

For a photoconductor with resistance R, made of a material with majority carrier lifetime τ and mobility μ, with a DC current I flowing in it, the lattice G-R noise voltage v_l is

$$\begin{aligned} v_l &= IR\sqrt{\frac{4\tau B}{\overline{N}}} \\ &= IR\sqrt{\frac{4\tau\mu BR}{\ell}}, \end{aligned} \tag{3.23}$$

Unfortunately, this calculation depends on parameters that are not readily extracted from most data sheets, such as the lifetime of each carrier species, which is the majority, and so on. In spectral regions (IR) where this is a significant noise source, it is usually necessary to rely on the manufacturer's assertion that a certain detector, at a given temperature and field of view, is BLIP, and go from there. This is unsatisfactory, since it makes it difficult to make trade-offs between, say, a cooled filter with significant passband losses or no filter at all. The cooled filter will be better if the lattice G-R noise and R_{sh} Johnson noise are low enough, whereas no filter is preferable if the signal loss is enough to bring the detector into the G-R or Johnson limits. It is usually best to set a lower limit on the noise using the Johnson noise of the published shunt resistance of the device, and then quiz the manufacturer as to how low the field of view can go before the detector ceases to be BLIP.

3.10.6 Multiplication Noise

APDs and photomultipliers exhibit multiplication noise, which appears as gain fluctuations. In a PMT, the variance of the electron gain at the first dynode is the principal contributor to this noise, while in an APD, the effect is distributed. This noise source is normally specified in the manufacturers' data sheets and must not be overlooked when designing detector subsystems, as it is often the dominant noise source toward the bright end of the system's dynamic range.

3.10.7 Temperature Fluctuations

A small detector weakly coupled to a thermal reservoir exhibits local fluctuations in thermal energy density, which are sometimes described as temperature fluctuations. This is a poor name, since the idea of temperature is well defined only in the limit of large systems, but it is unfortunately entrenched. Thermal fluctuations depend on the thermal mass of the detector, and on the thermal resistance between it and the reservoir, but not directly on the area; this makes it one of the few intrinsic additive noise sources for which D^* is an inappropriate measure. For a detector connected to a reservoir at temperature T through a thermal conductance G, the RMS thermal noise power spectral density is given by

$$\langle \Delta P^2 \rangle = 4kT^2G. \tag{3.24}$$

If the two surfaces are connected by thermal radiation, then for a small temperature difference, the thermal conductance is given by the derivative of the Stefan–Boltzmann

formula (2.4), with an emissivity correction:

$$G_{\text{rad}} = 4\sigma T^3 \frac{\eta_1 \eta_2}{\eta_1 + \eta_2 - \eta_1 \eta_2}, \tag{3.25}$$

so the the fluctuation due to radiation thermal conductance is

$$\left\langle \Delta P_{\text{rad}}^2 \right\rangle = 4k\sigma T^5 \frac{\eta_1 \eta_2}{\eta_1 + \eta_2 - \eta_1 \eta_2}. \tag{3.26}$$

For $\eta = 1$ and $T = 300$ K, the thermal conductance due to radiation is about 2 W/m^2/K, which is equivalent to that of 1.3 cm of still air (0.025 W/m/K). Thus for well-insulated thermal detectors, radiation can be an important source of both thermal forcing and fluctuation noise. Low emissivity surfaces and cooling can help a lot.[†]

3.10.8 Electronic Noise

A good detector can easily be swamped in electronic noise from a poorly designed or poorly matched preamp. It is vitally important to match the detector's characteristics to those of the amplifier if the best noise performance is to be obtained.

One common way in which amplifiers are misapplied is in connecting a simple transimpedance amp to a detector whose shunt capacitance is significant. This leads to a large noise peak near the 3 dB cutoff of the measurement system, which (although data sheets and application notes often describe it as inevitable) is easily avoided with a few circuit tricks. See Section 18.4.4 for details.

3.10.9 Noise Statistics

It is not enough to know the RMS signal-to-noise ratio of a measurement; without knowledge of the noise statistics, it is impossible to know what the effects of noise on a given measurement will be. The noise sources listed in this section are Gaussian, with the exception of most kinds of source noise and some kinds of electronic noise. Section 13.6.12 has a detailed discussion of this and other noise and signal detection issues.

3.11 HACKS

This section contains only optical hacks, but there are a number of circuit tricks listed in Chapters 10, 15, and 18 as well, which should be considered when choosing a detection strategy. It is important to keep the complete subsystem in mind during selection of a detector element and detection strategy.

[†]See, for example, Lynn E. Garn, Fundamental noise limits of thermal detectors. *J. Appl. Phys*. **55**(5), 1243–1250 (March 1, 1984).

3.11.1 Use an Optical Filter

If your measurement is limited by the noise of the background light, it can often be improved by a filter. In mid- and far-infrared systems using cooled detectors, you usually have to cool the filter too, because it emits radiation of its own in its stopbands. Interference filters may present less of a problem; they mostly reflect the stopband light, so the detector may see a reflection of itself and its cold baffles in the out-of-band region. Make sure you mount a room temperature interference filter with its interference coating facing the detector, or else the colored glass backing will radiate IR into your detector. This helps with narrow filters, which drift a long way when cooled.

3.11.2 Reduce the Field of View

In background-limited situations, the background level can often be reduced by limiting the field of view (FOV) of the detector. Typical ways of doing this are descanning the detector in a flying-spot measurement, or by using baffles and spatial filters to reject photons not coming from the volume of interest. In the mid- to far-IR, the situation is complicated by the thermal radiation of the baffles themselves, which must often be cooled in order to afford a signal-to-noise improvement. For BLIP detectors, if the background radiation is isotropic, the (electrical) noise power will scale approximately as the solid angle of the field of view, which is a very worthwhile improvement. To control stray background light and reduce the thermal load on these cold shields, the inside should be black and the outside shiny.

3.11.3 Reduce the Detector Size

As the FOV is reduced, there will come a point at which the background ceases to dominate other noise sources, so that further reductions are no help. If the shot noise of the signal is the next-largest effect, then only the collection of more photons will improve the measurement; most of the time, however, the next-largest effect will be Johnson or lattice G-R noise, which scale as the detector area.

The spatial coherence of the incident light will set a minimum étendue $n^2 A\Omega'$ for the detector, but if this has not yet been reached, it is possible to focus the light more tightly (larger FOV) on a smaller detector. This strategy has the advantage of reducing all the noise sources, while keeping the signal strength constant; as a bonus, smaller detectors tend to be faster and cheaper. The limits on this approach are set by the available detector sizes, by working distance restrictions at higher NA, and by approaching the shot noise level, which does not depend on detector area and FOV.

3.11.4 Tile with Detectors

Gain isn't everything. To pull really weak optical signals out of significant amounts of background light (too much for photon counting), consider using detectors as wallpaper. If your measurement is limited by background noise statistics, and the signal and background have the same spatial, angular, and spectral distribution, then the tricks of reducing detector size or FOV won't help any more. There's only one way forward: collect *all* of the light.

As you increase the detection solid angle, the background noise grows as $\sqrt{\Omega}$ but the signal goes as Ω. Sometimes you can just line a box with photodetectors, such as

CCDs or solar cells, and improve your measurement statistics. In this sort of case, consider especially whether your detector really needs imaging optics. Would a nonimaging concentrator or just putting the detector up close be better?

3.11.5 Cool the Detector

Cooling a silicon photodiode below room temperature doesn't accomplish much, though cooling does reduce the dark current of CCDs quite a bit. In the IR, where we're stuck with narrow bandgap materials, cooling helps detector noise in two ways. The main one is that it reduces leakage by sharply cutting the rate of thermal carrier generation (in photodiodes) or thermionic emission (in photocathodes); this effect is exponential in the temperature. In a photodiode, this leads to an enormous increase in the shunt impedance of the device, which reduces the Johnson noise current as well as the G-R noise.

The other way cooling helps is that the Johnson noise power of a resistor is proportional to its temperature, so that even with a fixed impedance, the noise current goes down; this effect is only linear, and so contributes less to the overall noise reduction. Transistor amplifiers running at room temperature can have noise temperatures as low as 30 K (see Section 18.5.3), so that it is usually unnecessary to cool the amplifier if the detector is run at 77 K (liquid nitrogen) or above.

Cooling schemes divide into thermoelectric (TE) and cryogenic. Neither is free, but TE coolers (TECs) are much cheaper than cryogenic ones. Multistage TECs can achieve trouble-free ΔTs of 130 °C, and single stage ones 60 °C, provided care is taken not to "short-circuit" them with heavy wires or mounts. This is adequate for work at 3.5 μm and shorter, or with strong signals at longer wavelengths.

Getting BLIP performance at $\lambda \gtrsim 5$ μm requires cryogenic cooling, which is much more involved. For lab use, LN_2 cooling is usually best, because simplest; in a field instrument, where LN_2 is hard to come by, some sort of mechanical cooler, such as a Joule–Thompson or Stirling cycle device, will be needed. Both alternatives are expensive.

In the extreme IR (beyond 20 μm), choices are more limited: often the choice is between a helium cooled extrinsic photoconductor such as Ge:Zn, or a room temperature bolometer or pyroelectric detector.

3.11.6 Reduce the Duty Cycle

The signal-to-noise ratio of a detection scheme can also be improved by concentrating the signal into a shorter time, as in a pulsed measurement with time-gated detection. Assuming the average optical power remains constant, as the the duty cycle[†] d decreases the electrical SNR improves as $1/d$, because the average electrical signal power goes as $1/d$, and the average noise power is constant, because the noise bandwidth goes as d^{-1} but the detection time goes as d. The limit to this is when the shot noise of the signal is reached—see Sections 10.8.2, 13.8.10, and 15.5.6.

3.11.7 Use Coherent Detection

By far the quietest and best-performing signal intensification scheme is coherent detection. It exploits the square-law properties of optical detectors to form the product of the

[†]Duty cycle is the fraction of the time the signal is active: a square wave has a 50% duty cycle.

signal beam with a brighter beam (often called the *local oscillator (LO) beam*, by analogy with a superheterodyne receiver). If the signal and LO beams have time-dependent vector electric fields $\mathbf{E}_s$ and $\mathbf{E}_{\mathrm{LO}}$, respectively, the photocurrent is given by

$$\begin{aligned} i(t) &= \mathcal{R}\left\{\iint_{\mathrm{det}} |\mathbf{E}_{\mathrm{LO}}(t) + \mathbf{E}_s(t)|\, dA\right\} \\ &= i_{\mathrm{LO}} + i_s + 2\ \mathrm{Re}\left\{\iint_{\mathrm{det}} \mathbf{E}_{\mathrm{LO}}(t)\mathbf{E}_s^*(t) dA\right\} \\ &= (DC) + 2R\sqrt{i_{\mathrm{LO}} i_s}\iint_{\mathrm{det}} W(t)\cos(\theta(t)) dA, \end{aligned} \tag{3.27}$$

where $\mathcal{R}$ is the responsivity, i_s and i_{LO} are the photocurrents generated by the signal and LO beams alone, W is the ratio of the local value of $|E_{\mathrm{LO}}E_s|$ to its average, and θ is the optical phase difference between the two beams as a function of position. If the two beams are in phase, perfectly aligned, and in the same state of focus and polarization, the integral evaluates to 1, so that the signal photocurrent sees a power gain of (i_{LO}/i_s). The shot noise is dominated by the additive noise of the LO beam, but since the amplification ratio is just equal to the ratio of the LO shot noise to the signal shot noise, the resulting total signal to shot noise ratio is equal to that of the signal beam alone, even with an E_s equivalent to one photon in the measurement time—a remarkable and counterintuitive result. This effect can overcome the Johnson noise of a small load resistor with only a milliwatt or two of LO power. This remains so for arbitrarily weak signal beams, so coherent detection offers an excellent way to escape the Johnson noise limit.

If θ is not 0 everywhere on the detector, the value of the integrals in Eq. (3.27) will be reduced. Even slight differences in angle or focus between the two beams will give rise to fringes, which will dramatically reduce the available amplification, and hence the signal-to-noise ratio. This seriously restricts the field of view of a heterodyne system, which may be undesirable. In some instances this restriction is very useful, as it allows rejection of signals from undesired locations in the sample space.

When the light beams are at exactly the same frequency, this is called homodyne detection, and when their frequencies differ, heterodyne. The SNR for heterodyne detection goes down by a factor of 2 because the signal power is averaged over all ϕ, and the average value of $\cos^2\phi$ is 0.5. Another way of looking at this is that a heterodyne detector receives noise from twice the bandwidth, since an upshifted optical beam gives the same beat frequency as a downshifted one (see Section 13.7.2). Temporal incoherence between the beams will spread the interference term out over a wide bandwidth, reducing the gain available as well (see Section 2.5.3). Examples of the use of this technique are heterodyne confocal microscopes, measuring interferometers, and coherent cw lidars.

3.11.8 Catch the Front Surface Reflection

You can get a signal-to-noise boost, in really tight spots, by arranging the first photodiode at 45° to the incoming beam and putting another one normal to the reflected light, wiring the two in parallel so that their photocurrents add (Figure 3.8). That way, a sufficiently

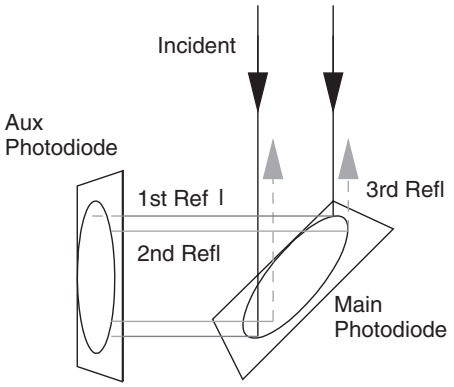

Figure 3.8. Catching the front-surface reflection from photodiodes. This trick improves detection efficiency and can enormously reduce the effects of etalon fringes in the photodiode windows.

low NA beam has to make three bounces off a photodiode before it can escape. (Using a smaller angle will increase the number of bounces but require bigger photodiodes.)

The front-surface reflection from an uncoated silicon photodiode is about 40%, so this trick can result in a gain of almost 4 dB in electrical signal power (2–4 dB in SNR), a highly worthwhile return from a small investment. With coated photodiodes, the gain will be smaller, but even there, this is a simple and inexpensive way to pick up another several tenths of a decibel in signal strength. Another benefit is that by collecting most of the reflected light, you can greatly reduce the signal strength drifts caused by temperature and wavelength sensitivity of etalon fringes in the photodiode windows. (This effect, which often limits the attainable accuracy of CW optical measurements, is discussed in Section 4.7.2.) Adding a third photodiode to bend the light path out of the page by 90° eliminates polarization dependence completely, and the two extra bounces improve the light trapping even more. This approach is used in high accuracy radiometry.

3.11.9 Watch Background Temperature

In mid- and far-infrared systems, the detector is normally the coldest thing, and so (although its emissivity is very high) its thermal radiation is weak. A convex lens surface facing the detector will reflect a demagnified image of the detector and its surroundings. In an imaging system, this will result in a dark region surrounded by a lighter annulus, the *narcissus effect*, from Ovid's story of the boy who fell fatally in love with his own reflection. Narcissus is actually a good effect—it becomes a problem when the image of the detector moves, or when its magnification < 1 as in our example, so that it has hot edges and a large spatial variation. Since the detector's emission is wideband, in general narcissus is not an etalon issue. Good baffles and careful attention to silvering are needed to ensure that radiation from the rest of the optical system doesn't dominate the detected signal.

Besides narcissus and other instrumental emission, nonuniformity in background temperature can mask weak infrared sources, as skylight does stars. The chopping secondary mirror of Example 10.5 is one way of fixing this.

3.11.10 Form Linear Combinations

Many measurements require addition and subtraction of photocurrents. This is best done by wiring the detectors themselves in series (for subtraction) or parallel (addition). Doing this ensures that both photocurrents see exactly the same circuit strays and amplifier gain and phase shifts; this makes the addition or subtraction extremely accurate and stable, without tweaks (see the subtraction trick of Section 18.6.1) and the differential laser noise canceler of Section 18.6.5.[†] It is a bit confusing at first, but since the far ends of the photodiodes are connected to very low impedance bias points (which are basically AC ground), the series and parallel connections are equivalent for AC purposes; the noise sources and capacitances appear in parallel in both cases.

3.11.11 Use Solar Cells at AC

One problem with good quality silicon photodiodes is their cost per unit area. A 5 mm diameter photodiode can easily run $100, although some are available more cheaply (down to $5). There are lots of applications in which more area is better, but cost is a problem. If you have such an application, consider using solar cells. A 25×75 mm amorphous silicon solar cell costs $5 in unit quantity, has a quantum efficiency of 0.5, and responds well throughout the visible. It is very linear at high currents, and surprisingly enough, if you use the cascode transistor trick (Section 18.4.4), you can get 3 dB cutoffs up to 20 kHz or so. Some smaller cells work at 100 kHz. Because of leakage, you can't usually run much reverse bias, so if you're using an NPN cascode transistor with its base and collector at ground potential, bias the solar cell's anode at -0.6 to -1 V. Besides large capacitance and leakage, solar cells have serious nonuniformity—they often have metal stripes across their faces, to reduce lateral voltage drops. On the other hand, for photoelectrons per dollar, you can't beat them.

3.11.12 Make Windowed Photodiodes into Windowless Ones

One good way of avoiding etalon fringes in photodiode windows is to use windowless photodiodes. Many types of metal-can photodiodes can be used without windows, but procuring such devices can be very difficult and expensive in small quantities. For laboratory and evaluation use, it is frequently convenient to remove the windows from ordinary devices. The methods used most are filing or cutting using a lathe. These methods often lead to metal chips or cutting oil being left behind on the die, possibly causing short circuits, scratches, or $1/f$ noise and drift.

A much more convenient and safe method is to use a big ball-peen hammer, although this may seem odd initially. Hold the diode in a vice by the leads, with the base resting on top of the jaws, and tap the glass gently with the peen (the rounded side). It will turn to powder, which can be removed by turning the diode over and tapping it against the side of the vice. The protruding face of the ball makes the blow fall on the glass, but the gentleness of its curvature ensures that it will be stopped by the metal rim of the case before any glass dust is ground into the die.

Because the glass is clean and nonconductive, it does not lead to any long-term degradation of the optical performance of the detector, and because any glass falling

[†]There's lots more on noise cancelers in Philip C. D. Hobbs, Ultrasensitive laser measurements without tears. *Appl. Opt*. **36**(4), 903–920 (February 1, 1997).

on the die does so reasonably gently, no scratches result. The only problem with this approach is that not all diodes are adequately passivated for windowless operation. The passivation layer used in ordinary IC chips is usually a thick layer of silica glass produced by a sol-gel process or by sputtering. Because the glass and the chip have very different refractive indices (1.5 vs. 3.4 to 4), it is not so easy to AR coat a diode processed this way, especially since the thickness of the passivation layer may be poorly controlled; for best performance, it may be necessary to AR coat the die, passivate it, and then AR coat the top of the passivation layer. Understandably, this is not often done, so that a diode with excellent performance in a hermetically sealed package may degrade very rapidly when the window is removed. The usual symptoms are a gradually increasing dark current, together with rapidly growing $1/f$ noise and occasional *popcorn* bursts. The otherwise good Hamamatsu S-1722 used in the example is in this class.

Aside: Hermetic Seals. Most ICs and other active devices are packaged in Novolac epoxy. Lots of optoelectronic parts such as LEDs and photodiodes are encapsulated in clear polycarbonate. CCD windows are often glued on with epoxy. All these plastics are great, but there's one thing they aren't: hermetic. Water vapor diffuses readily through plastic. The air inside a CCD package with an epoxied window will respond to humidity changes outside with a time constant of a week or two; this can lead to condensation inside the package in service, especially in cooled setups. If you're building a system with a cooled detector, insist on a glass-to-metal or frit-bonded seal, or else work without a window and fight the dust instead.

3.11.13 Use an LED as a Photodetector

Direct bandgap devices such as GaAs diodes have very steep long-wavelength cutoffs, which can reduce the need for short-pass filters. This can be used to good account in highly cost-sensitive applications, at least when these detectors can be had at low cost. Unfortunately, most such detectors are relatively expensive. One exception is ordinary AlGaAs LEDs. These devices are inefficient as detectors; their quantum efficiency is low, they have small areas, and the optical quality of their packages is extremely poor. Nevertheless, their long-wavelength cutoff is very steep, and it can be selected to some degree by choosing a device of the appropriate emission color and package tint. Where spectral selectivity is needed and every nickel counts, they are sometimes just the thing.

3.11.14 Use an Immersion Lens

Although the minimum étendue cannot be reduced, remember that it contains a factor of n^2. If you contact a hemisphere of index n to the photodiode, you can reduce its area by n^2, thereby reducing the capacitance by the same factor and increasing the effective D^* too. This of course works only for n up to the refractive index of the photodiode material, but this is 3.5 for Si and 4 for Ge. Plastic package photodiodes are a good candidate for this, because their indices are similar to glass, so that UV epoxy or index oil can be used easily. Thermoelectrically cooled HgCdTe devices really need this treatment.

3.11.15 Use a Nonimaging Concentrator

The immersion idea can be extended by using a nonimaging concentrator, as shown in Figure 3.9. Repeated bounces off the sides of the cone cause the angle of incidence to

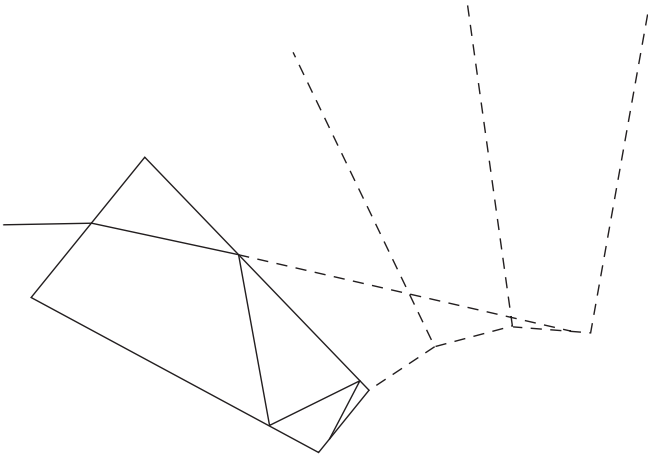

Figure 3.9. Nonimaging concentrator for improving photon collection with a small diode. The simplest kind is the cone concentrator, shown along with its unfolded light path; unfolding is an easy way to see which rays will make it and which won't. Due to near-normal incidence, the bottom of the cone will need silvering.

increase. As shown in the figure, TIR cannot be relied on near the bottom of the cone, so it will probably have to be silvered. Don't silver the whole cone unless you have to, since TIR is more efficient. There are better ways to make a concentrator than this, for example, the compound-parabolic concentrator, which can achieve the thermodynamic limit.

3.11.16 Think Outside the Box

There are fashions in the detector business as elsewhere, and the received wisdom about how to do things is always a mixture of actual engineering experience and "professional judgment." Sometimes it's right and sometimes it's wrong. For example, most solid state thermal cameras are built in lithographically defined arrays much like CCDs or CMOS imagers. They may be cryogenically cooled (InSb and HgCdTe) or run at room temperature (PZT, lithium tantalate, or microbolometers), but their basic physical outline is an array of small pixels on a solid surface. This has significant effects on the performance and economics of the devices—they tend to cost between $2000 and $40,000 and have maximum NEΔT of a bit below 0.1 K, in array sizes from 256 to 500,000. The small pixel sizes require well-corrected lenses, which are very expensive in the infrared. Not all applications absolutely need that many pixels, and for those uses, there's an alternative method as shown in Figure 3.10.

This sensor uses large pixels (3×5 mm) made of carbon ink applied to a 9 μm film of PVDF by screen printing (T-shirt lithography). The film is freestanding in air, leading to very low thermal conductance. Interestingly, a photon budget shows that the best SNR is achieved by insulating the pixels (which makes them slow but sensitive) and recovering the bandwidth by digital filtering afterwards, as we'll do in Example 17.1. The reason is that the insulation slows the sensor down by *increasing* the low frequency sensitivity, rather than *decreasing* the high frequency sensitivity. The big pixels and low resolution (96 pixels) mean that a simple molded polyethylene Fresnel lens works well. The multiplexer is a little more of a problem, but it turns out that an array of ordinary

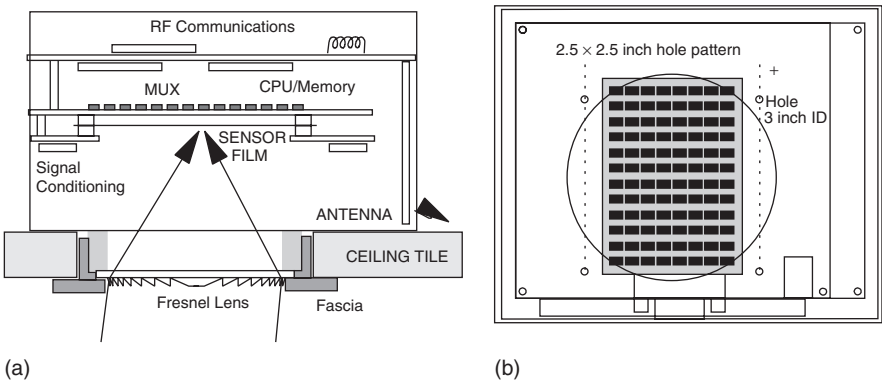

Figure 3.10. Footprints thermal infrared imager: (a) a cross-section view of the development version shows its 50 mm diameter, 0.7 mm thick HDPE Fresnel lens, and 9 μm PVDF sensor film freestanding in air; (b) a front view without the lens shows the 8 × 12 array of 3 × 5 mm carbon ink pixels screen-printed on the PVDF.

display LEDs used as switches (see Section 14.6.1), one per pixel and driven by 5 V logic lines, does an excellent job, leading to a sensor costing about $10, whose NE$\Delta$T is about 0.13 K. There are some other signal processing aspects we'll look at in Section 18.7.2 and Example 17.1.[†]

[†]For more details, see Philip C. D. Hobbs, A $10 thermal infrared sensor. *Proc. SPIE* **4563**, 42–51 (2001) (http://electrooptical.net/www/footprints/fpspie11.pdf) for the gory technical stuff, and Philip C. D. Hobbs, Footprints: a war story. *Opt. Photonics News*, pp. 32–37 (September 2003) (http://electrooptical.net/www/footprints/fpwaropn.pdf) for the war story.