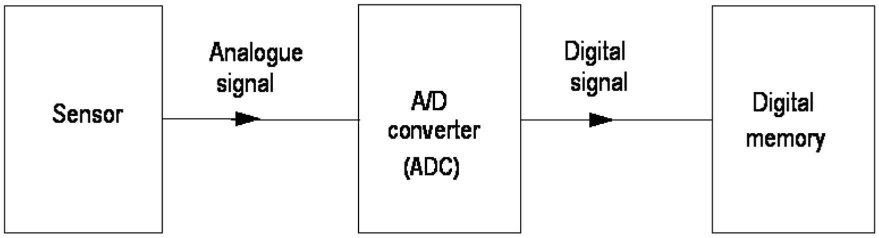
1. Types of Sensors Used for the Recording of Images

- **Photographic film** (not described in this compendium)

Advantages:	Cheap Independent of computer hard- and software (which changes rapidly)
Disadvantages:	Limited spectral range ($\lambda \leq 1\ \mu m$) Poor utilization of light Requires chemical processing Not directly accessible for computer storage and processing Can be recorded only once, after which it is a read-only medium

- **Image sensors which give a direct electrical signal**

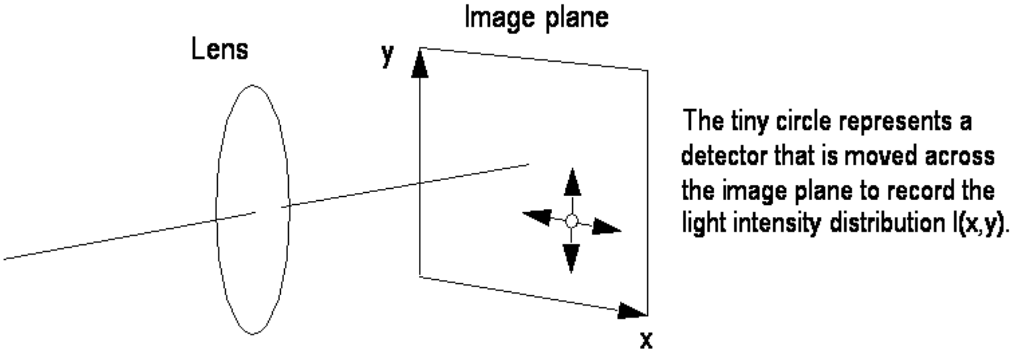
These sensors have good compatibility with computerized systems. The signal can easily be analog-to-digital (A/D) converted and stored in computer memory.



There are three principles for recording an image:

1. A point detector combined with two-dimensional (2D) scanning

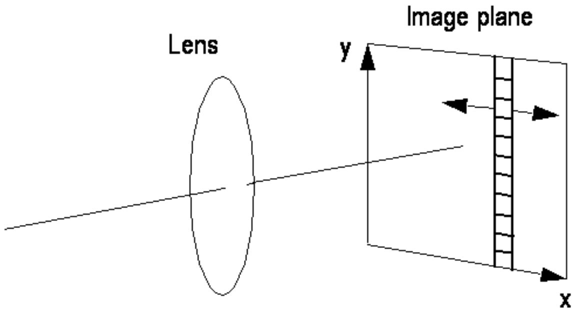
In this scanning mode a single, small-area ("point") detector is used for recording the light distribution in an optical image:


Lens
y
Image plane
x
The tiny circle represents a detector that is moved across the image plane to record the light intensity distribution I(x,y).

In practical instruments, it is more common for the detector to be stationary and for the image to move. (In some recordings, for example of the earth‘s surface using satellites, only one-dimensional scanning is necessary because the satellite itself moves over the surface of the earth.)
2D scanning using a single detector element is not so common today, but it is still used in some laser-scanning equipment like confocal microscopes. The type of detector used varies for different applications, but photomultiplier tubes are among the types used (described later in this compendium).

2. A linear sensor combined with 1D scanning

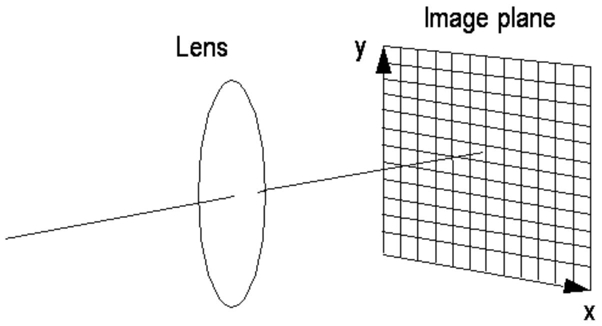
Typical number of elements: 5000 – 10 000. Often the linear sensor consists of a row a photodiodes, each having a size of approximately 5 μm square.



The ladder-like structure in the figure represents a row of detector elements. An entire picture column can be divided into pixels without any mechanical movement. To record an entire image, the detector is moved stepwise in the x direction.

Compared with 2D scanning, this is a much more rapid method for recording images, because multiple detector elements are exposed to light simultaneously. This type of scanning is used in document scanners. It is also used in satellite imaging of the earth, and in this case no mechanical scanning is necessary.

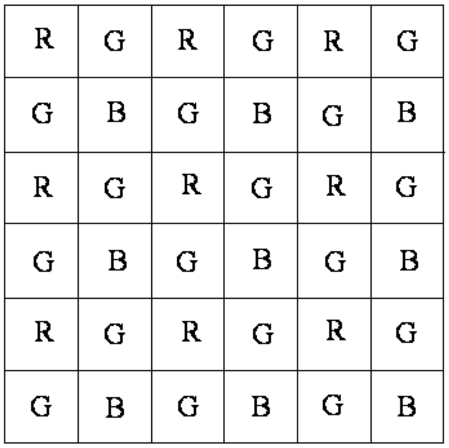
3. Area array sensor - no scanning necessary



The image plane is covered by detector elements. I(x,y) is obtained without the need for scanning.

This method provides very rapid image recording, because all detector elements are exposed to light simultaneously. It is used in video cameras and cameras for digital photography. In the latter case the number of detector elements is typically about 10 megapixels in consumer products. The size of the entire detector matrix can vary from about 5 x 7 mm up to about 20

x 30 mm, or even more in professional cameras. Each detector element has a side-length of a few microns.
For the recording of color images, the area array sensor must perform some sort of color separation. In consumer products this is accomplished by covering different detector elements with different filters (usually red, green, and blue). A common color mosaic of this type is called Bayer pattern.



R	G	R	G	R	G
G	B	G	B	G	B
R	G	R	G	R	G
G	B	G	B	G	B
R	G	R	G	R	G
G	B	G	B	G	B

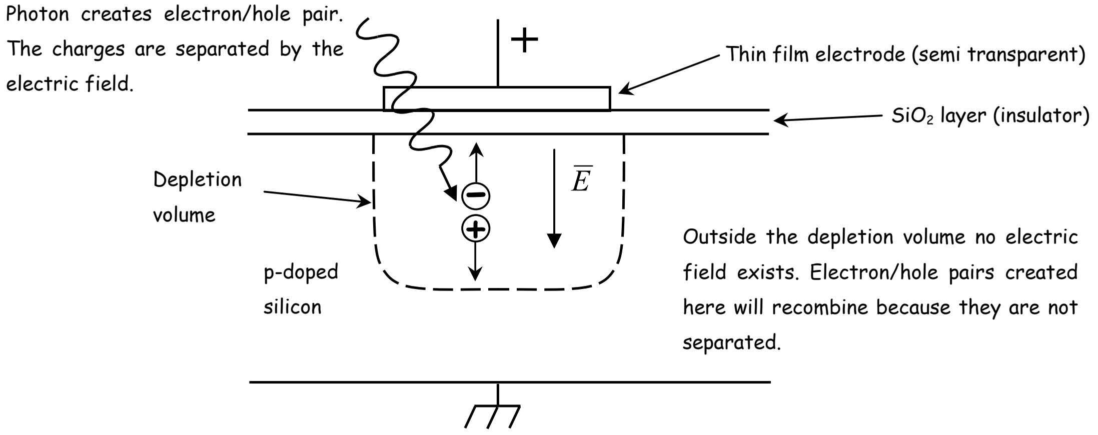
Layout of blue- green- and red-sensitive detector elements in Bayer mosaic pattern

Each detector in the matrix thus records only a limited part of the visible spectrum, and therefore it is not possible to determine the true image color at the single pixel level (we have to combine information from several pixels to determine the color). In reality the digital camera manufacturers use interpolation algorithms to calculate the most likely color for each pixel, so that they can present a full, say, 10 Mpixel image in color. One should keep in mind, however, that there is some guesswork involved in this, and that erroneous results can sometimes occur.

In the color mosaic pattern above there are twice as many green detectors as there are blue and red respectively. This bias reflects the human visual system, where most of the information concerning image detail and brightness comes from the green wavelengths. To produce high-quality images (both black-and-white and color) it is therefore most important to collect a sufficient amount of information in the green part of the spectrum.

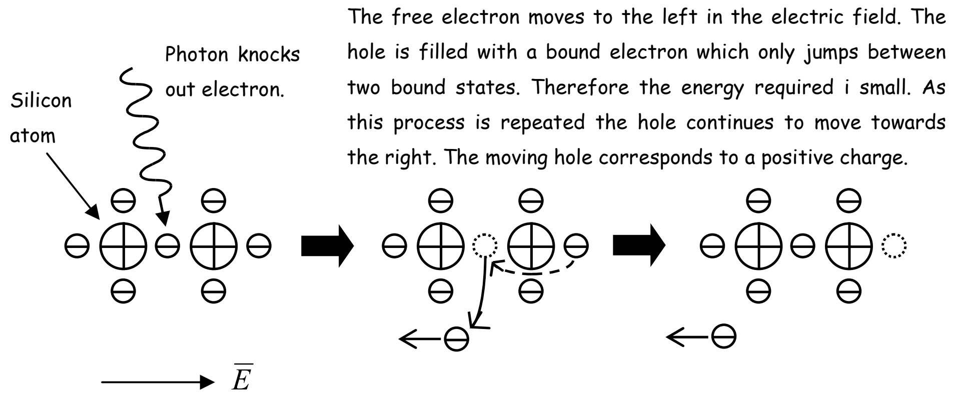
2. Semiconductor detectors

There exist many different types of semiconductor detectors, for example photodiodes, phototransistors and photogates. The basic light detection mechanism is more or less the same in all of these detectors, and therefore we will look at a photogate as a representative example of a semiconductor detector. Photogates are often used in area array sensors for scientific applications. A cross-sectional view of a photogate is shown in the figure on next page. Typically the size is somewhere in the range 3 to 5 μm.


Photon creates electron/hole pair. The charges are separated by the electric field.
+
Thin film electrode (semi transparent)
SiO_2 layer (insulator)
Depletion volume
$\overline{E}$
p-doped silicon
Outside the depletion volume no electric field exists. Electron/hole pairs created here will recombine because they are not separated.

The material of the photogate is p-doped silicon, on top of which there is an insulating layer of SiO_2. On top of this insulating layer there is a (semi)transparent thin film electrode. If a positive voltage is applied to the electrode, the positive charge carriers (holes) in the silicon are repelled. As a result, a depletion volume devoid of mobile charge carriers is formed below the electrode. The higher the voltage, the deeper this depletion volume will be. An electric field will form in this depletion volume. The photogate is now ready to detect light.

An incoming photon with sufficient energy can knock out an electron from a silicon atom in the crystal lattice. The result is that an electron/hole pair is formed. If this happens in the depletion volume, the electron and the hole will be separated by the electric field as illustrated in the figure below. The electron will move towards the electrode where it will come to rest just underneath the insulating layer. The hole, on the other hand, will move in the opposite direction and will leave the depletion volume. A photon energy of approximately 1.2 eV, corresponding to a wavelength of approximately 1 µm, is needed to create an electron/hole pair. As a result, a photogate of silicon has high sensitivity to both visible and near-infrared radiation[♣].


Photon knocks out electron.
Silicon atom
The free electron moves to the left in the electric field. The hole is filled with a bound electron which only jumps between two bound states. Therefore the energy required i small. As this process is repeated the hole continues to move towards the right. The moving hole corresponds to a positive charge.
$\overline{E}$

[♣] Digital consumer-type cameras incorporate a filter that blocks infrared radiation, because otherwise strange imaging effects would occur. In some (rare) cases these IR-blocking filters can be removed so that the camera can be used for infrared photography.

The more photons that are absorbed in the depletion volume, the more free electrons are created. These electrons will assemble below the positive electrode. But there is a limit to the number of electrons that can be collected in this limited area, because they mutually repel each other. Furthermore, their negative charge will reduce the effect of the positive electrode on deeper layers in the silicon substrate. As a result, the depletion volume will be reduced and ultimately it will vanish altogether. This situation, which can occur when the sensor is overexposed, means that electrons can start to spread to neighboring pixels in the area array sensor, an effect known as "blooming." In older sensor types the blooming phenomenon could spread and ruin a large image area. Nowadays the blooming effect is usually limited because of improvements in chip design. The limitation in the number of electrons that can be collected in a pixel remains, however. This maximum number is usually called "well capacity," because the collection of electrons in a pixel is often likened to the collection of water in a well. When the well is full it overflows. The area of each well (i.e. pixel) depends on the size of the electrode, and its depth depends (within limits) on the applied voltage. As a result, a large pixel area and a high voltage mean that more photons can be recorded during the exposure. The well capacity differs between different sensors, but it is often in the range $20\,000 - 100\,000$.

After the exposure to light has been completed, pixel data must be read out from the circuit. In a so-called CCD (charge coupled device) circuit the collected electrons are shifted between pixels until they reach an output register, where they are read out to external circuits. A CMOS (complementary metal oxide semiconductor) has additional electronic components (transistors, capacitors etc.) integrated in each pixel. Therefore charge is transformed into a voltage locally in each pixel before read-out. Another difference is that the individual pixels are addressable in a CMOS circuit, so that only the desired pixel values can be read out very quickly. This is a big advantage in some high-speed applications.

Electron-hole pairs can also be formed thermally, producing a so-called dark signal. Cooling reduces this problem, and it is a must for devices that use long exposure times (for example in astronomy).

The signal that is read out from an individual pixel after the exposure has been completed is proportional to the charge accumulated. As a result, the output signal is proportional to the number of photons detected during the exposure time[♣]. In this respect the semiconductor detector behaves quite differently from photographic film, which is highly non-linear. The semiconductor detector is therefore much better suited for quantitative measurements than photographic film.

Important concept: quantum conversion efficiency, η

η = the percentage of photons that produce an electron-hole pair.
For photodiodes (and other semiconductor detectors), η is often 50-90%.
For photographic film, $\eta \approx 1\%$ (the percentage of photons producing a chemical reaction).

Obviously semiconductor detectors are much better at detecting photons than photographic film, and they are also superior to most other detectors in this respect. The fact that they can

[♣] In consumer-type digital cameras gamma correction is often performed. As a result, the pixel values will no longer be proportional to exposure, see Appendix 4.

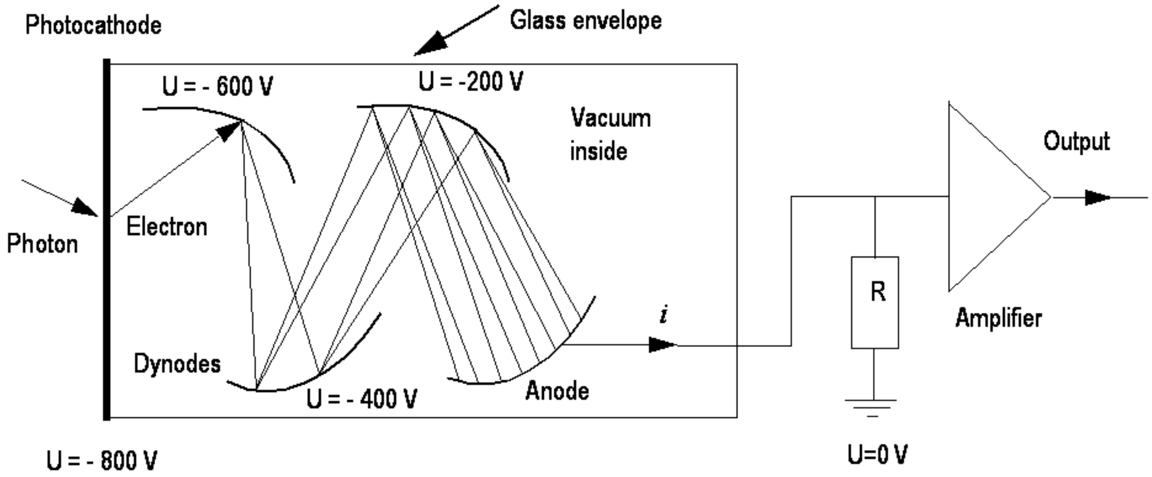
be manufactured as linear arrays or matrices of detector elements is also a great advantage. It is therefore natural to ask: **Is the semiconductor matrix detector the ideal detector which makes all others redundant?**
The answer to that question is **NO!** Examples of applications where other kinds of detectors are more suitable (at the moment):

- For wavelengths in the intervals <400 nm and >1 μm, other detectors are often more suitable.

- For very low light levels in combination with short measurement times. In this case semiconductor detectors produce an output signal that is often too low. (If you have plenty of time, the output signal from a cooled diode detector can be integrated for a long time. This is used, e.g., in astronomy.)

3. Photomultipliers

Photomultiplier tubes (PMTs or PMs) are used to measure very low light intensities in a short time. PMTs also work well at short wavelengths, down to ~ 100 nm. PMTs come in many different sizes and shapes, but in all cases they consist of a number of electrodes situated inside an evacuated glass envelope. The basic principle is that a photon impinging on the photocathode will (sometimes) knock out an electron from the material that covers the photocathode. An electric field in the PMT will accelerate the electron towards the closest electrode, which is called the first dynode, see figure. The electron will have a considerable speed when it hits the dynode, and as a consequence it will knock out several secondary electrons (typically 3-5). These secondary electrons will be accelerated towards the second dynode by an electric field, and will knock out perhaps 3-5 electrons each. This process is then repeated throughout the whole dynode chain, which often consists of something like ten dynodes. This process is a nice example of an avalanche effect (like a nuclear explosion, but less dramatic). The end result is that the single initial electron has multiplied to perhaps a million electrons that eventually hit the anode of the PMT, where they are detected by an external circuit.



The figure shows a simplified schematic representation of a photomultiplier tube. In reality there are more dynodes so that the current amplification is boosted (and the number of secondary electrons is usually higher than two as shown in the figure). The voltages given are typical, but can vary considerably depending on the type of PMT and what it is used for. Higher voltages will produce more secondary electrons and thus higher current amplification. Another thing not illustrated in the figure is that the number of secondary electrons varies statistically, a fact that gives rise to noise in the signal (multiplication noise). Typically PMTs have a current amplification of the order of 10^6. When used in a typical application, the current from the photocathode can be of the order of 10^{-12} A (impossible to amplify with external electronics as the signal will be lost in the noise). The corresponding anode current will then typically be 10^{-6} A = 1 μA, which is relatively easy to amplify with external electronics. Apart from light, thermal effects can also lead to the emission of electrons from the cathode, producing a dark current which flows also in the absence of light. Cooling reduces this problem.

The quantum conversion efficiency is often around 10% for a PMT, i.e. lower than for a semiconductor detector. **The greatest advantage with a PMT is that it provides a high current amplification with relatively low noise**.

Conclusion: Semiconductor detectors detect light very efficiently, but give a low output signal. PMTs detect light rather poorly, but are good at amplifying the signal.

Question: How low light levels can be detected, where is the lower limit?

In practice, it is the noise that limits the measurements. Two factors govern the lower limit of the light that can be detected:

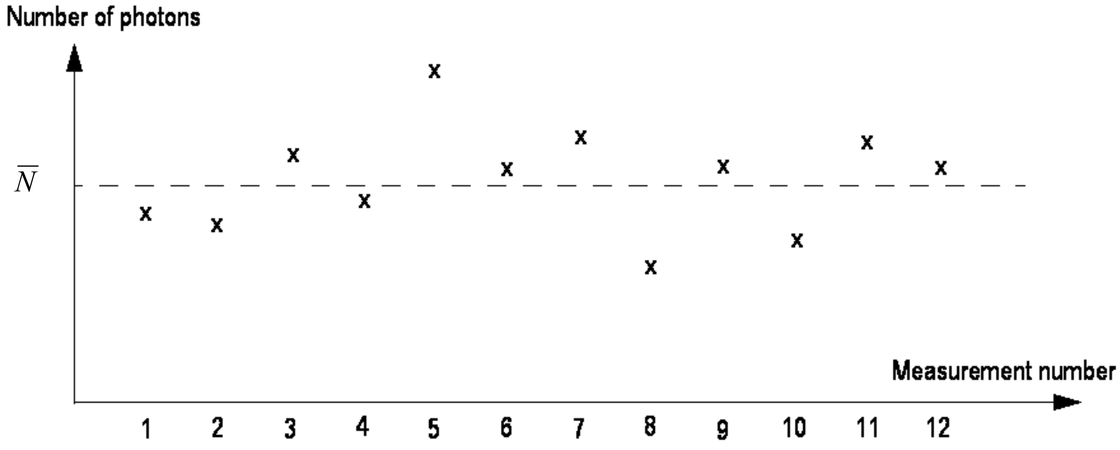
1. How long may the measurement take?

2. How "noise-free" must the measurement be?

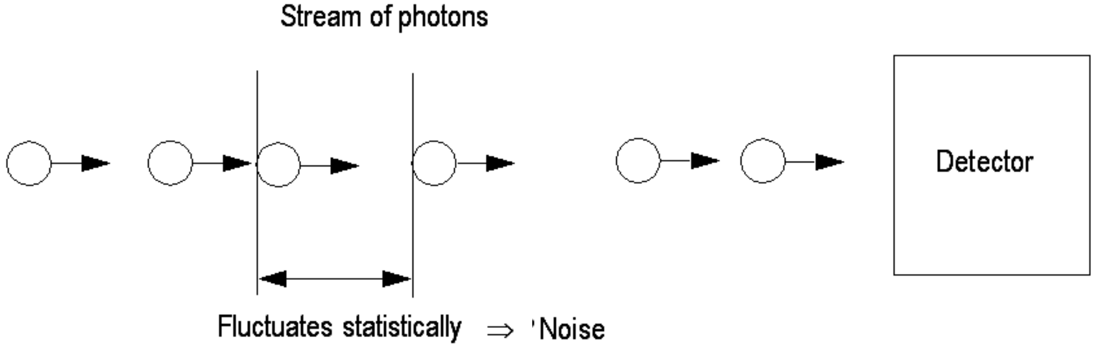
NOTE: The most significant contribution to the noise at low light intensities is not usually from the detector or the amplifier but from the light itself. We will now consider this noise, which is a characteristic of light, and is called "photon quantum noise".

4. Photon Quantum Noise

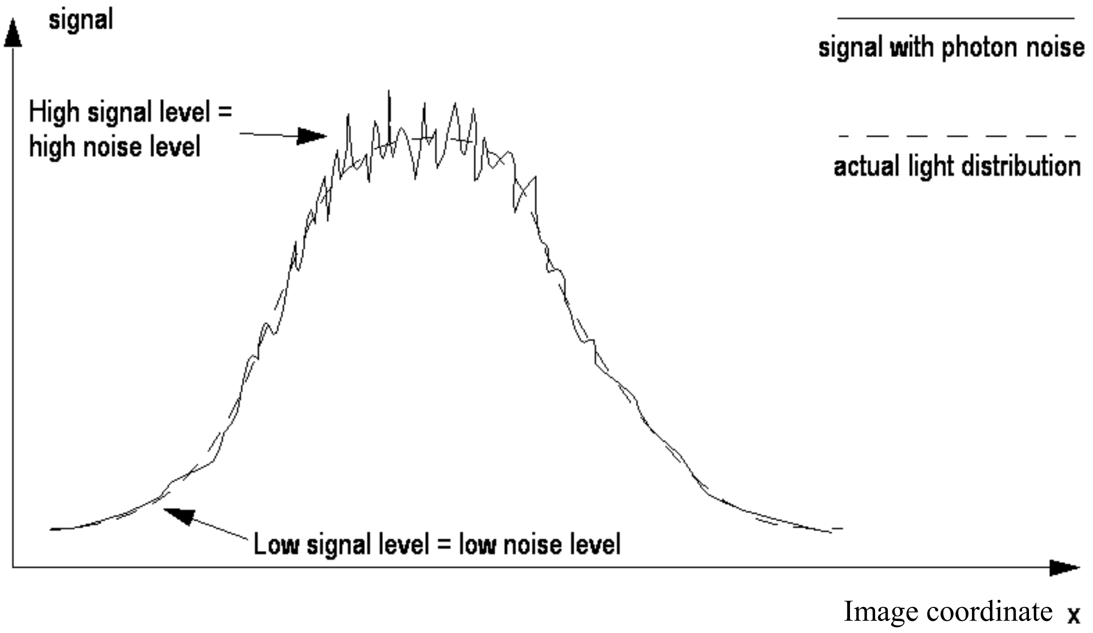
Assume that the intensity of the light is such that we expect $\overline{N}$ photons to arrive during the chosen duration of the measurement. Assume also that we have a perfect detector, which simply counts the exact number of photons arriving during the measurement period. Repeated measurements will give varying photon numbers N, e.g. the results shown in the figure on next page. The spread in the results is described by a Poisson distribution.



If we repeat the measurements a large number of times, we will obtain a **mean value** of $\overline{N}$ (i.e. the expected value) and a **standard deviation** of $\sqrt{\overline{N}}$. The mean value, $\overline{N}$, represents the magnitude of the signal and the standard deviation, $\sqrt{\overline{N}}$, the noise. This noise is not due to errors in the measurements, but is an intrinsic characteristic of the light itself. The emission of a photon from a light source is a statistical process. One can never know when the next photon will be emitted, only the probability that it will occur within the next, say, picosecond. As a result, the stream of photons arriving at the detector will not be equally spaced. Instead there will be a random fluctuation in the distance between consecutive photons as shown in the illustration below. This means that the number of photons arriving at the detector during a measuring period will also vary.



What about photon quantum noise in systems that do not use photon counting (for example the PMT circuit shown previously, which produces an analog output signal)? It can be shown that also in such cases the noise, expressed as root-mean-square (RMS), increases as the square root of the signal level. This means that if the output signal is a current, *i*, as from a PMT, we get $i_{noise} = \lim_{T \to \infty} \sqrt{\frac{1}{T}\int_0^T \left(i - i_{average}\right)^2 dt} = K \cdot \sqrt{i_{average}}$, where i_{noise} is the RMS noise value, $i_{average}$ is the current averaged over a long time period (i.e. the current we would get in the absence of noise) and *K* is a constant. In a practical situation, the integration is usually performed over a total time, *T*, which is long compared with the statistical fluctuations in the signal. Photon quantum noise differs from many other types of noise in that it increases when the signal level increases. This can often be seen when looking at the output signal from an image scanning system, as in the illustration below.



5. The Signal-to-Noise Ratio

Referring to the photon counting measurements described on the previous page, we define the signal-to-noise ratio (*SNR*) as: $SNR = \dfrac{\text{mean value}}{\text{standard deviation}} = \dfrac{\overline{N}}{\sqrt{\overline{N}}} = \sqrt{\overline{N}}$

This is true if photon quantum noise is the only source of noise. Obviously, the only way to improve the *SNR* is to record more photons. This can be achieved by increasing the light intensity and/or extending the measuring time. It is often difficult to increase the light intensity (e.g. in astronomy) and the only alternative is then to extend the measuring time. For images, the *SNR* usually refers to repeated light measurements from the same pixel.

NOTE: Due to the square root dependence, a 10 times better *SNR* requires a 100 times longer measuring time.

In non-photon-counting systems, like the PMT circuit shown on page 10, we define $SNR = \dfrac{\text{mean value}}{\text{RMS noise}}$. All systems incorporate some means for signal integration (= low-pass electrical filtering). If the signal integration time is τ, we get $SNR = \sqrt{\overline{N}}$, where $\overline{N}$ is the expected number of photons detected by the system during τ. Again we have assumed that photon quantum noise is the only source of noise (In data sheets from manufacturers, the *SNR* under most favorable conditions, i.e. close to saturating light intensity, are usually quoted.). We can also get a digital output from the PMT circuit by connecting the output signal to an ADC (cf. page 5). This will produce an output in the form of integer numbers, just like in the photon-counting case. A difference from the photon-counting case, however, is that the digital numbers obtained do not (in general) correspond to the number of photons detected during the integration time τ. The digital values will be influenced by, for example, PMT voltage and amplifier gain. The *SNR* in this case is given by $SNR = \frac{\text{mean value}}{\text{standard deviation}} = \sqrt{\overline{N}}$, where the mean value and standard deviation of the digital numbers from the ADC are inserted. ***Note that $\overline{N}$ is the number of detected photons, not the mean value of the digital numbers from the ADC.***

If the quantum conversion efficiency is less than unity, we will lose some photons, and then the SNR will be $\sqrt{\eta\overline{N}}$.

Example: A PMT with η = 0.10 gives a *SNR* of about 30% of the theoretical maximum, while a diode detector with η = 0.80 gives a *SNR* of about 90% of the theoretical maximum (assuming that other sources of noise are negligible).

Despite the higher value of η for the semiconductor detector, a PMT is often more suitable in practice for low light intensities as the amplification of the signal is less noisy. The ideal solution would be to combine a high value of η with virtually noise-free amplification, but this has proved difficult to achieve. When detecting extremely low intensities, both semiconductor detectors and PMTs must be cooled. This is necessary to prevent the signal from "drowning" in the background noise caused by thermal effects.

One may ask if this is not only of academic interest, and that in reality values of $\overline{N}$ are very high.
The answer to that is NO! In many cases, measurements of light are limited by photon noise.

Below are two examples of such situations.

- In astronomy it is common to study very faint objects. In some cases the situation may be improved by employing very long measuring times (hours). In such cases, cooled semiconductor area array sensors are a good choice. (Liquid nitrogen is sometimes used as a coolant.) These sensors have replaced photographic film, because film has a low quantum conversion efficiency.
  Compared with a PMT, an area array sensor has the advantage that detection is parallel, i.e. light is collected on all pixels simultaneously. When using a PMT it is necessary to scan the image in two dimensions. An area array sensor can thus collect much more light than a PMT in the same time. However, semiconductor detectors are not suitable in some wavelength regions, or when measurements must be made quickly. Then other kinds of detectors, e.g. PMTs, must be used.

- Fluorescence microscopy is often used to study very faint objects with a limited lifetime (cf. astronomy where objects are usually quite long-lived). Furthermore, dynamic events are often studied, and therefore the measuring time is limited. In such cases PMTs are often the correct choice. In practice, a pixel value in fluorescence microscopy is often based on about 100 detected photons. The kind of noise level associated with such a signal can be appreciated by considering that the probability of a single measurement being less than 90 or greater than 110 is 32%. This level of noise will cause the image to appear grainy.

6. Sources of Noise Other Than Photon Noise

In addition to photon quantum noise, there are also other sources of noise present. Examples of such noise include:

Amplifier noise: Random fluctuations in the output signal caused by thermal effects etc. in the electronics. This type of noise can be reduced by reducing the speed with which data are read out from the sensor. For example, reduced frame rate in a video camera means less amplifier noise.

Multiplication noise in PMTs: The source of this noise is statistical fluctuations in the number of secondary electrons emitted from the dynodes.

Fixed pattern noise: In linear and area array sensors the sensitivity of the individual detector elements varies somewhat. This is due to imperfections in the manufacturing process. The result is a seemingly random variation in pixel value in the recorded images. In reality, however, these variations follow a fixed pattern (hence the name) and can therefore be compensated for. Since we are not dealing with random variations, and since the defect can be compensated for, it is questionable if it should really be called noise.

Dark signal noise: This noise is caused by statistical fluctuations in the dark signal mentioned in connection with semiconductor detectors and PMTs. Although the average value of the dark signal can be subtracted from the measurements, the statistical variations, i.e. the noise, will still remain. The higher the average value for the dark signal, the higher the noise will be (analogous to photon quantum noise). To reduce this type of noise, the detector can be cooled. This reduces the dark signal and thereby also the noise associated with the dark signal.

Quantization noise: This type of noise is caused by the discrete output levels of the analog-to-digital converter (ADC). As a result, analog inputs within ± 0.5 of an ADC level will all result in the same digital output. This will give a standard deviation of $\frac{1}{\sqrt{12}}$ ADC levels. See Appendix 3 for details.

If the noise levels from several different sources are known, the total noise level, n_{tot}, is given by:

$n_{tot} = \sqrt{n_1^2 + n_2^2 + \ldots}$, where n_1, n_2 etc. are the noise levels of the individual sources. For digital signals, the n-values represent standard deviation, and for analog signals they represent RMS noise (the use of the term RMS is not very strict, however, and it is sometimes used to denote standard deviation).

17. Sampling

In practical imaging, it is obviously impossible to store the value of $I_R(x, y)$ for every real pair of coordinates (x,y). This would require, for one thing, infinite storage capacity. In addition, for linear and area array sensors, the distance between the detector elements defines the minimum distance between measuring points. This leads us to the topic of ***sampling***. Image sampling implies the measurement of light intensity at a number of detector positions in the image, normally distributed in a uniform pattern.

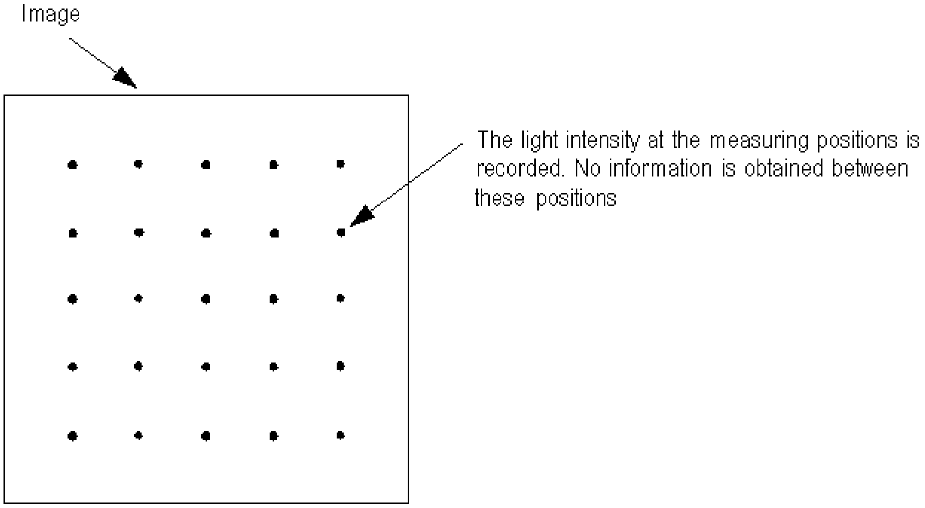
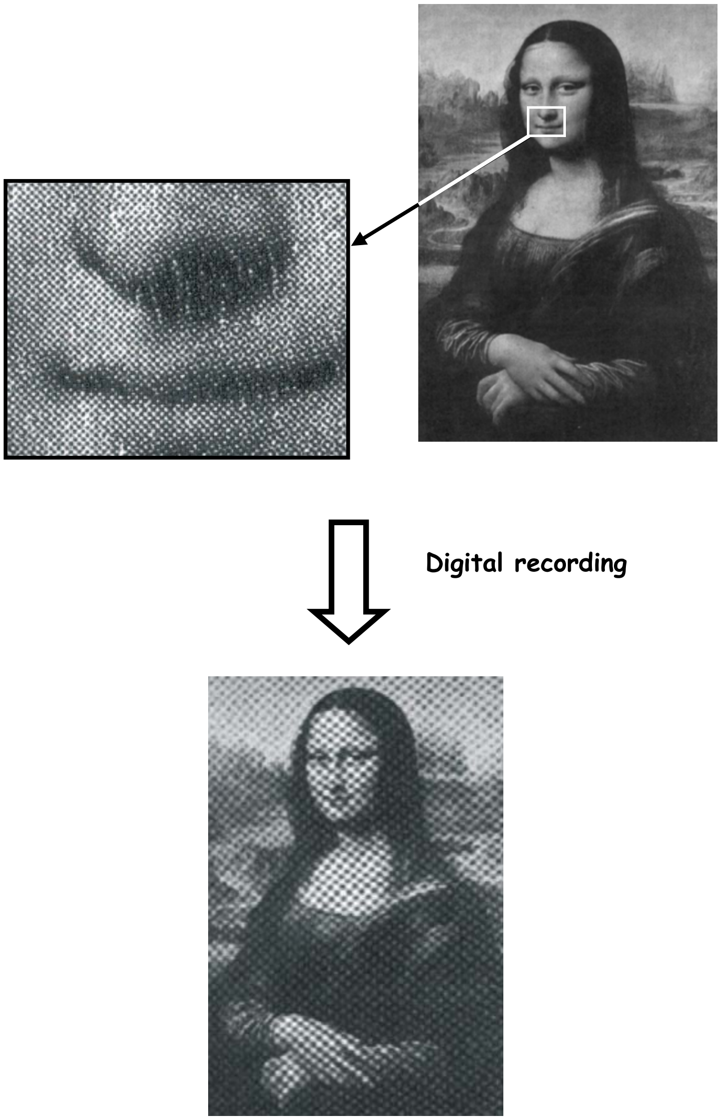


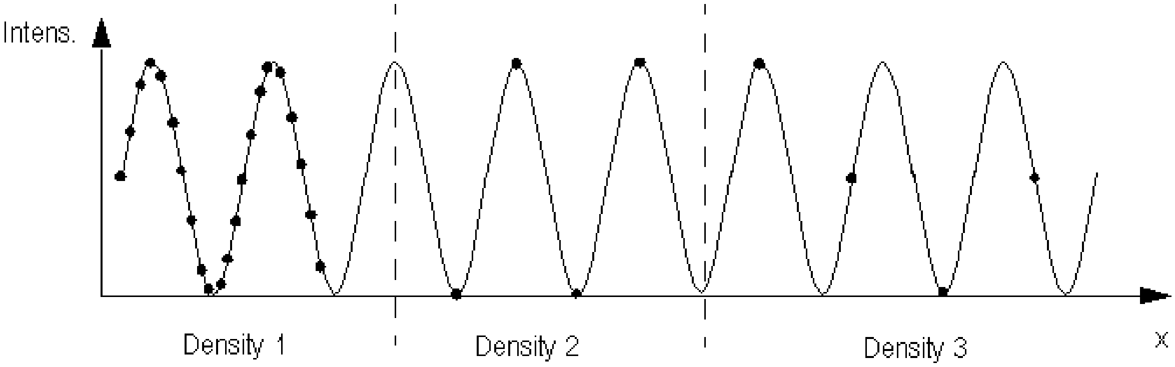
Image
The light intensity at the measuring positions is recorded. No information is obtained between these positions

The measurement values, which are normally stored digitally, are called pixels (from “picture cells“). Depending on the application, digital images may consist of anything from just a few pixels up to many million pixels. For comparison it can be mentioned that an ordinary television image consists of about 0.25 Mpixels. The fewer the number of pixels, the less information we get in the image (NOTE: The number of pixels recorded by, for example, a digital camera is often erroneously referred to as resolution).

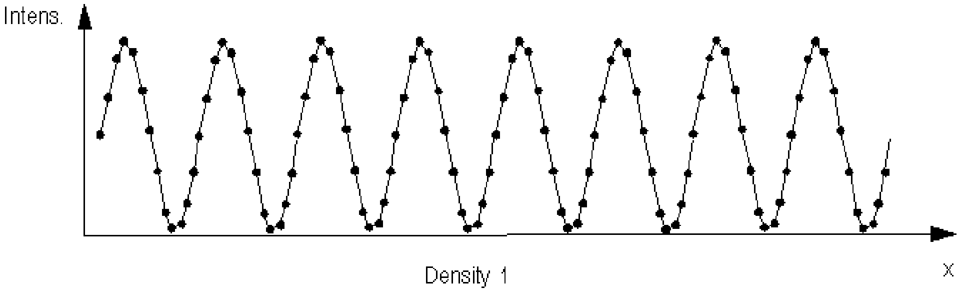
Not only do we get less information if the sampling points are spaced far apart, we may also get imaging artifacts introducing false information in the recorded images. A commonly seen example of this is when printed raster images are scanned in a document scanner, see illustration below. The coarse spotted pattern appearing in the recorded image is called aliasing, and this phenomenon will be investigated in this chapter. To simplify the description, we will start with the one-dimensional sampling case. The two-dimensional case will be treated in chapter 18.

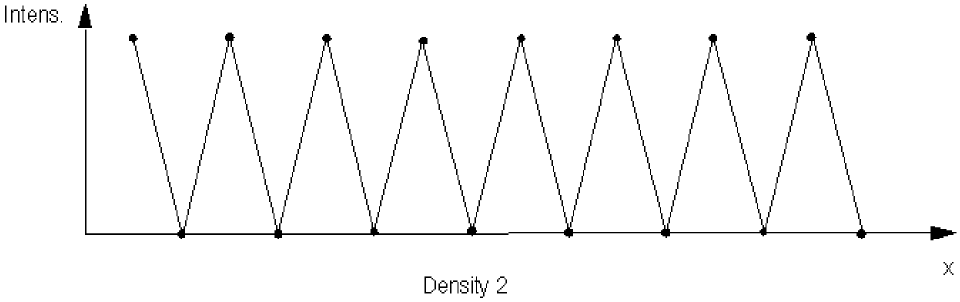


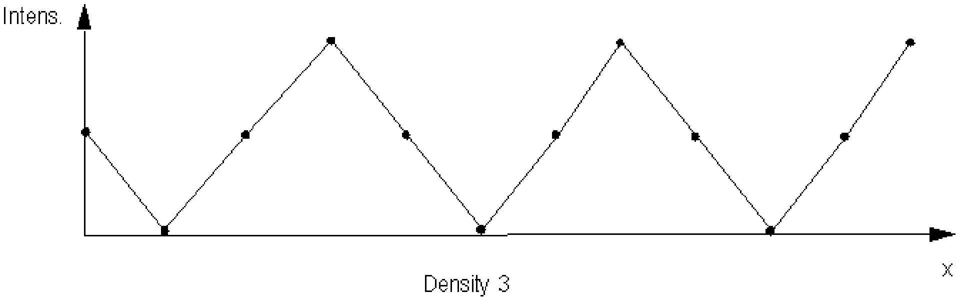
Let us now study sampling in more detail, and express it quantitatively. We will also see how sampling and the *MTF* together affect the image quality. Assume that we have a sinusoidal variation in intensity in the x-direction of the image. Let us consider the results of various sampling densities.



Sampling density 1 corresponds to "many" sample points per period of the sine wave. Density 2 corresponds to exactly two sample points per period, and density 3 to less than two sample points per period. Let us make simple image reconstructions (linear interpolation) from the sampled values for the three sampling densities:







Density 1 gives a rather good representation of the original signal. Density 2 preserves the correct spatial frequency, but the shape of the curve is not correct. Density 3 preserves neither frequency nor shape of the original curve. Furthermore, at low sampling densities it is important how the samples are placed in relation to the peaks and troughs of the sine curve (if, for density 2, we had moved the sample points a quarter of a period along the x axis, we would not have recorded any modulation at all!). These results seem to imply that the higher the sampling frequency (the smaller the distance between sampling points) the more accurate the result will be. This seems intuitively correct, and we would expect that when sampling an optical image we can never exactly reconstruct the original image from a finite number of samples in x and y. This conclusion, however, is ***wrong***. We will soon show that in order to ***exactly*** reconstruct an image which contains spatial frequencies up to ν_{max} (remember that the optics always has a ν_{limit}) the sampling frequency must be at least $2\nu_{max}$. This is called the sampling theorem, and a frequency of half the sampling frequency is called the Nyquist frequency. In our examples density 2 just barely fulfils the sampling theorem, i.e. at least two sampling points per period. (The spiky appearance in the reconstruction can be removed by using a mathematically correct reconstruction process instead of drawing straight lines between the points. This is studied in more detail later.)