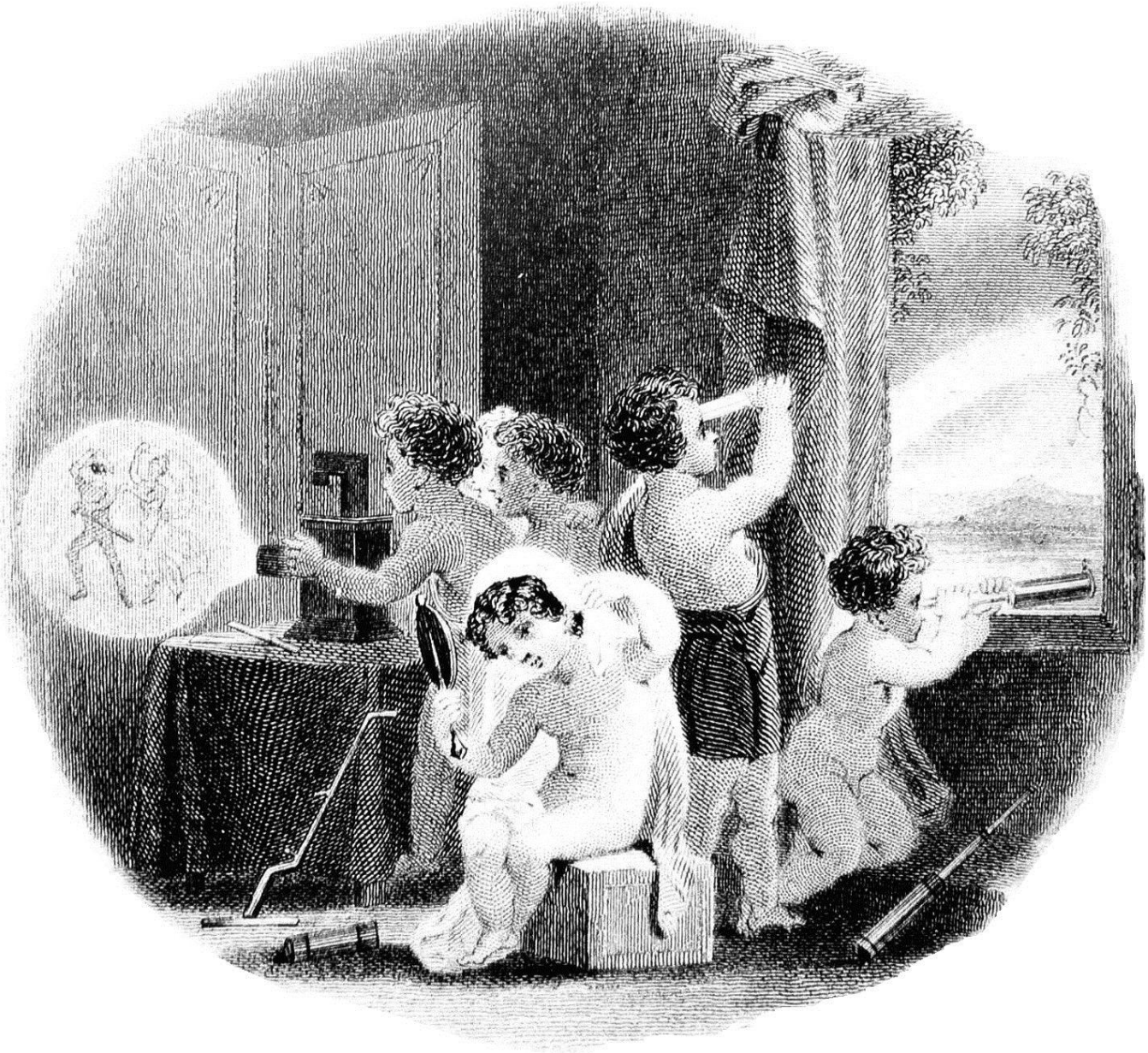
IMAGING PHYSICS



Kjell Carlsson

NOTE: The cover picture was kindly supplied by the illustrious society Ljusets barn (Children of Light). The picture clearly shows the many uses of optical imaging techniques, both for pleasure and for scientific purposes.

Contents

1. Types of Sensors Used for the Recording of Images

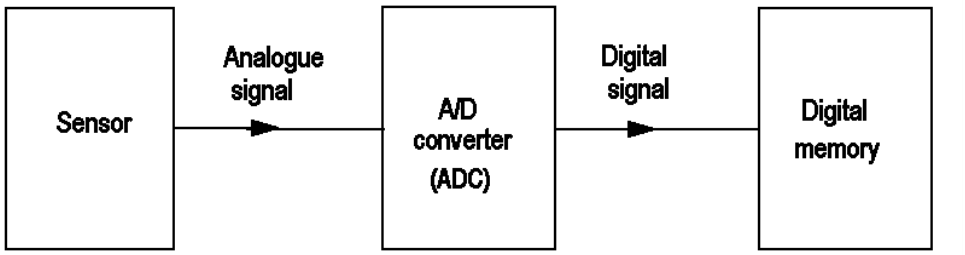
- **Photographic film** (not described in this compendium)

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**

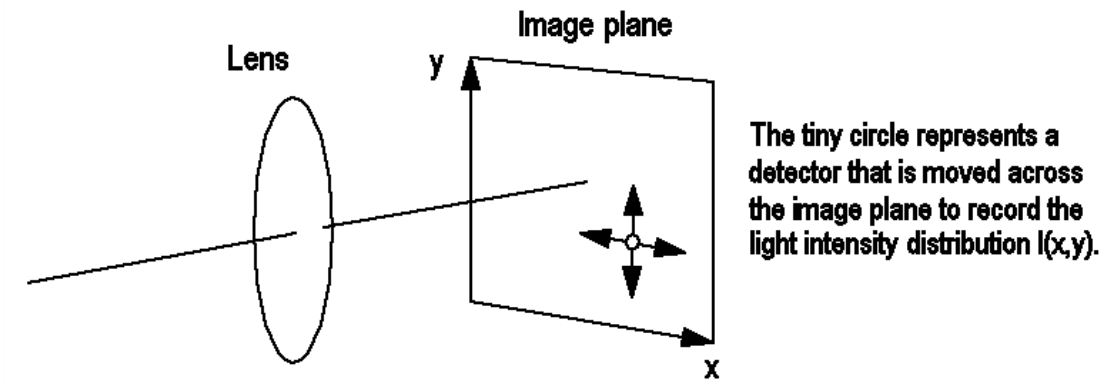
These sensors have good compatibility with computerized systems. The signal can easily be analog-to-digital (A/D) converted and stored in computer memory.



There are three principles for recording an image:

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

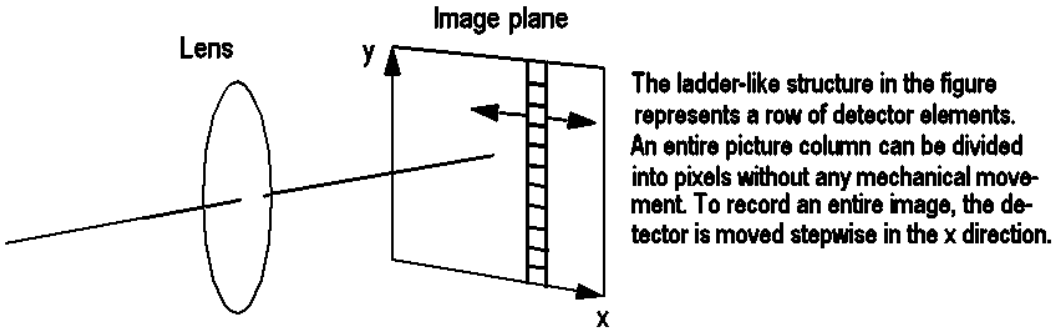
In this scanning mode a single, small-area ("point") detector is used for recording the light distribution in an optical image:



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).

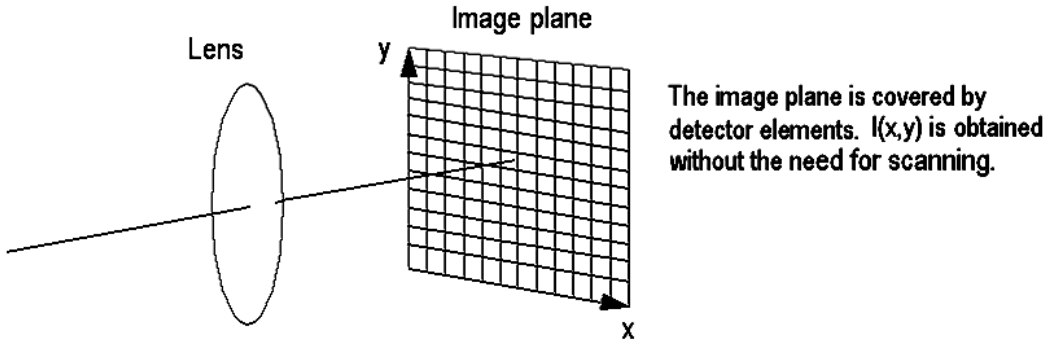
2. A linear sensor combined with 1D scanning

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.



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



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

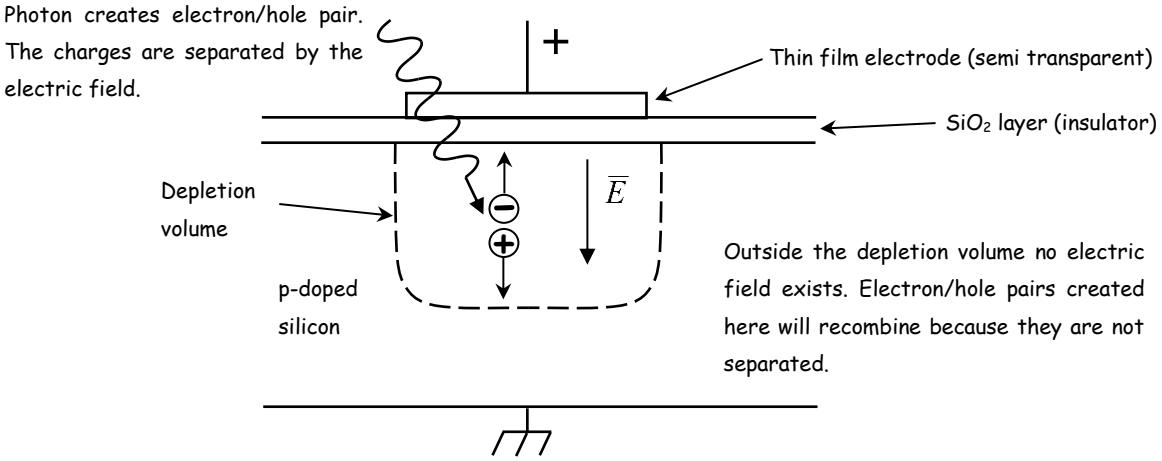
Layout of blue- green- and red-sensitive detector elements in Bayer mosaic pattern

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.

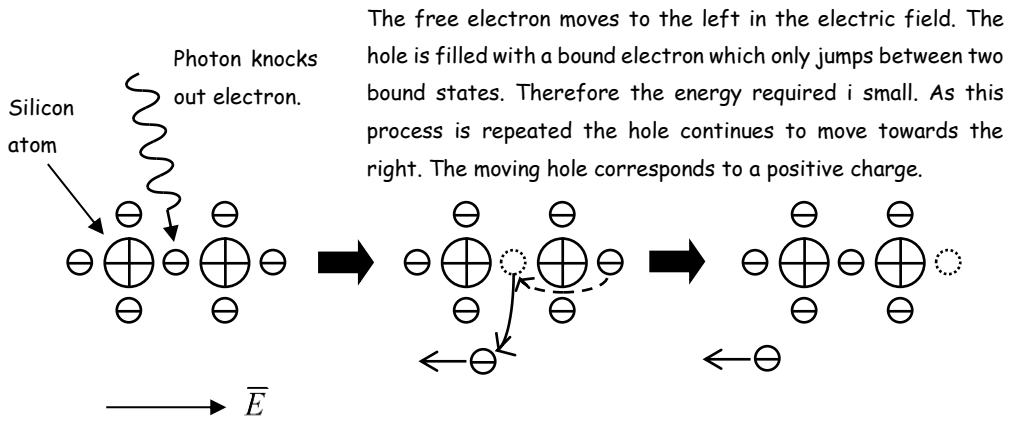
2. Semiconductor detectors

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.



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♣.



The more photons that are absorbed in the depletion volume, the more free electrons are created.

♣ 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.

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 be manufactured as linear arrays or matrices of detector elements is also a great advantage. It is

[♣] 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.

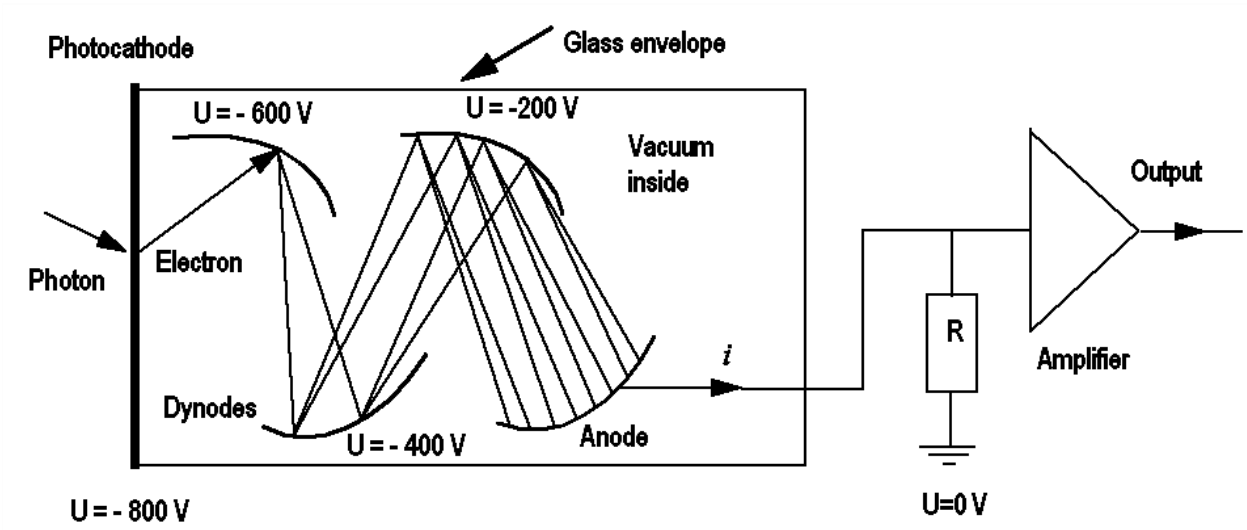
therefore natural to ask: **Is the semiconductor matrix detector the ideal detector which makes all others redundant?**
The answer to that question is **NO!** Examples of applications where other kinds of detectors are more suitable (at the moment):

- 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:

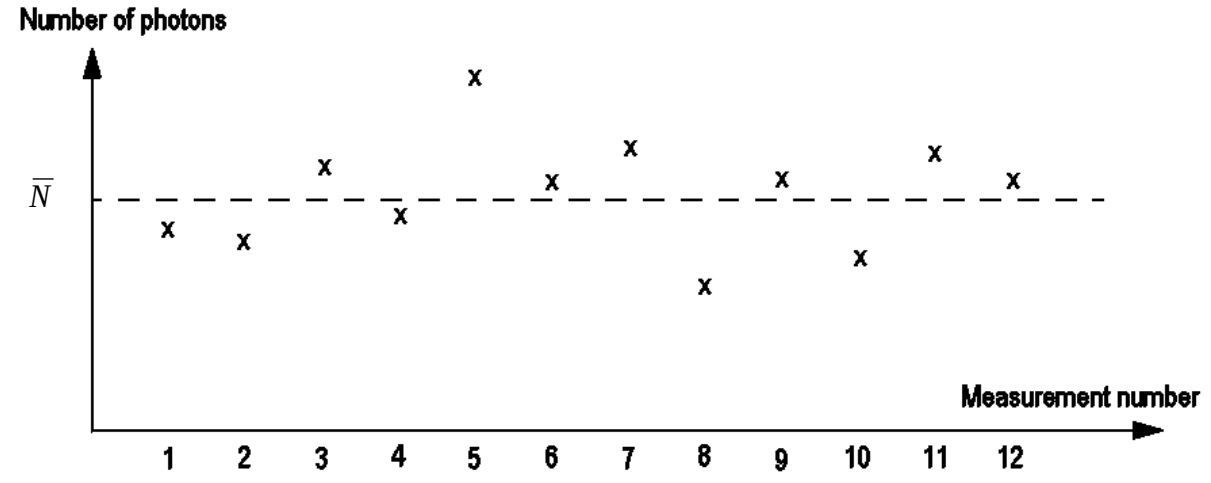
1. How long may the measurement take?

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

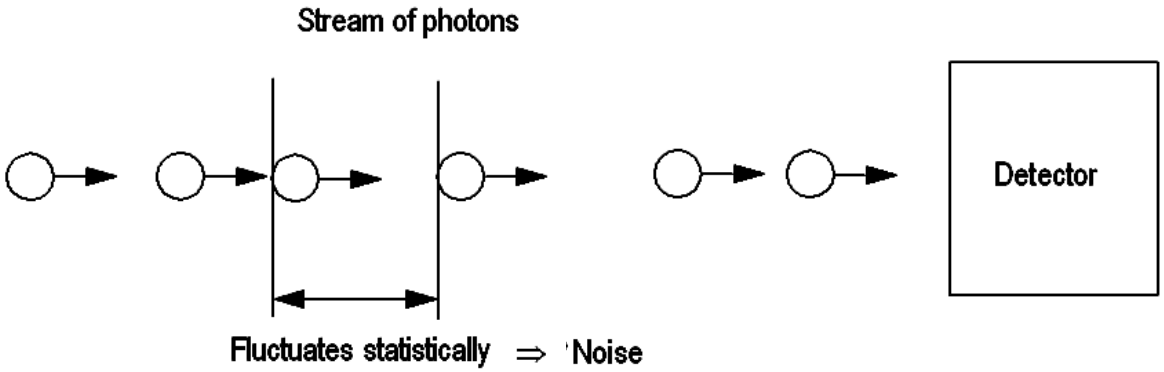
NOTE: The most significant contribution to the noise at low light intensities is not usually from the detector or the amplifier but from the light itself. We will now consider this noise, which is a characteristic of light, and is called "photon quantum noise".

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.

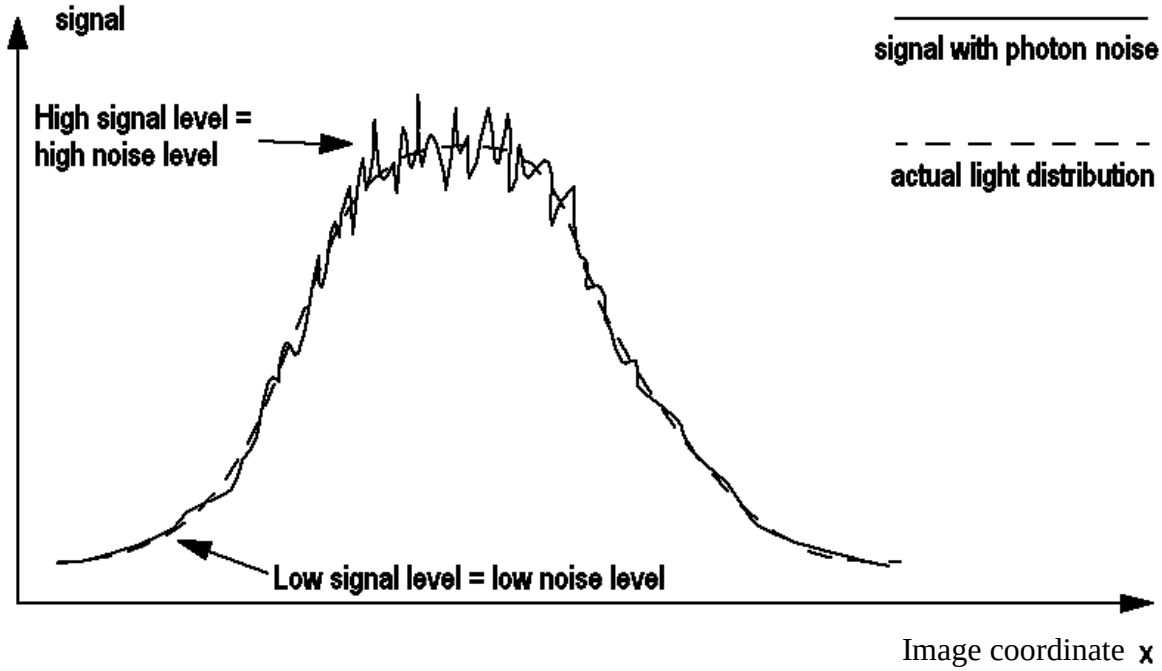


If we repeat the measurements a large number of times, we will obtain a **mean value** of $\overline{N}$ (i.e. the expected value) and a **standard deviation** of $\sqrt{\overline{N}}$. The mean value, $\overline{N}$, represents the magnitude of the signal and the standard deviation, $\sqrt{\overline{N}}$, 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: $SNR = \frac{\text{mean value}}{\text{standard deviation}} = \frac{\overline{N}}{\sqrt{\overline{N}}} = \sqrt{\overline{N}}$

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 $SNR = \frac{\text{mean value}}{\text{RMS noise}}$. 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).

7. Dynamic Range & Number of Significant Bits

A number often quoted by sensor manufacturers is the dynamic range. This is defined as: $\text{Dynamic range} = \frac{\text{Maximum output signal}}{\text{RMS}_{\text{noise,dark}}}$. $\text{RMS}_{\text{noise,dark}}$ denotes the noise level with the detector in complete darkness. The noise figure thus includes amplifier noise and dark signal noise, but not photon noise. In reality when the sensor is used at high light levels photon quantum noise strongly dominates, and the noise level becomes considerably higher. Therefore the maximum SNR obtainable for a sensor is usually much lower than its dynamic range. The dynamic range is, however, important when deciding the number of bits needed in the ADC. These things are investigated in Appendix 3, and some results of this are included in the example below.

Example: An area array sensor has a "well capacity" of 1.5×10^5 electrons, which means that the individual detectors can collect a charge corresponding to 1.5×10^5 electrons before saturation. Let's assume that the minimum RMS noise (corresponding to the combined effects of amplifier and dark current) is equivalent to 90 electrons.

a) What is the dynamic range of the sensor?
b) What is the maximum SNR obtainable?
c) How many bits should we use in the ADC?

Answers:

a) Using the definition above, the dynamic range will be $\frac{1.5\times10^5}{90} = 1.67\times10^3 \approx 1700$. The dynamic range is often expressed in decibels (dB), which in this case gives $20\log\left(1.67\times10^3\right) = 64\,\text{dB}$.

b) Maximum *SNR* is obtained at maximum exposure, i.e. 1.5×10^5 accumulated electrons (corresponding to an equal number of detected photons). This gives a standard deviation of $\sqrt{1.5\times10^5} = 387$ electrons, which is the RMS noise we get from the photon noise. Total RMS noise at maximum exposure is then given by $n_{tot} = \sqrt{387^2 + 90^2} = 398$ electrons. The $SNR = \frac{1.5\times10^5}{398} = 377$. This is very close to the value of 387, which would have been obtained if only photon quantum noise were present.

c) The number of bits required is related to the dynamic range of the sensor. We must be able to handle the highest light values from the sensor, and at the same time we must be able to handle subtle differences in grey value in the shadows of a scene. From the results in Appendix 3, we can conclude that a reasonable number of ADC levels is approximately $= \text{Dynamic range} = 1700$ in this case. This means that 11 bits are needed ($2^{11} = 2048$ levels). 11-bit ADCs probably don't exist, so we would use a 12-bit type.

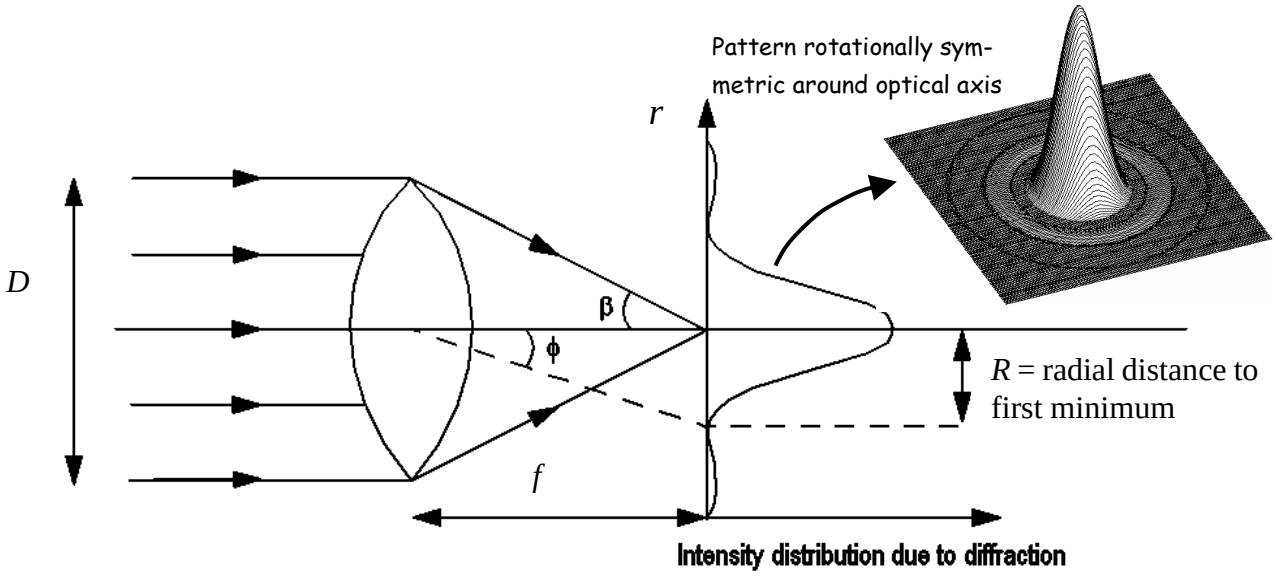
We have so far considered some aspects of photometry (the measurement of light) in connection with image recording. We have introduced quality measures such as SNR and dynamic range, and we have investigated the number of significant bits. We will now turn to other measures of image quality, namely those which describe how "sharp" (i.e. clear and detailed) the image is.

This property is often called (geometric) **resolution**♣. The term resolution is, however, not a precise term (there is, for example, no unequivocal definition), and we will thus soon proceed to other, more well-defined methods of determining image quality. This will lead us to "optical transfer functions", which were developed during the 1950s, and are based on mathematical models such as convolution and Fourier transforms. But first we will consider the simpler concept of resolution.

8. Geometric Resolution

Amateurs often believe that the degree of magnification is a good measure of the quality of, for example, a telescope or a microscope. The higher the magnification, the greater the detail that can be seen. Unfortunately, this is not so. Instrument makers became quickly aware that there was no point in pushing the magnification beyond a certain limit. The images became larger, certainly, but they also became more blurred. The result of this is that it is not worthwhile to magnify, for example, images by more than 1000 times in a light microscope.

The limit on how fine details can be seen in an image is determined by the aberrations of the optics, and by the diffraction of the light as it passes through the optics. Optics which have very low aberrations are often called **diffraction-limited optics**. In other words, it is the diffraction of the light which limits the resolution of the instrument. Let us consider a practical example. Assume that we are using a diffraction-limited circular lens to obtain an image of a star. The star can be regarded as a point object at infinite distance. According to geometric optics, we will obtain an infinitely small point image. Due to diffraction this is not true. Instead we get an intensity distribution with a maximum on the optical axis, and a number of small secondary maxima as we move away from the axis, as illustrated in the figure below.

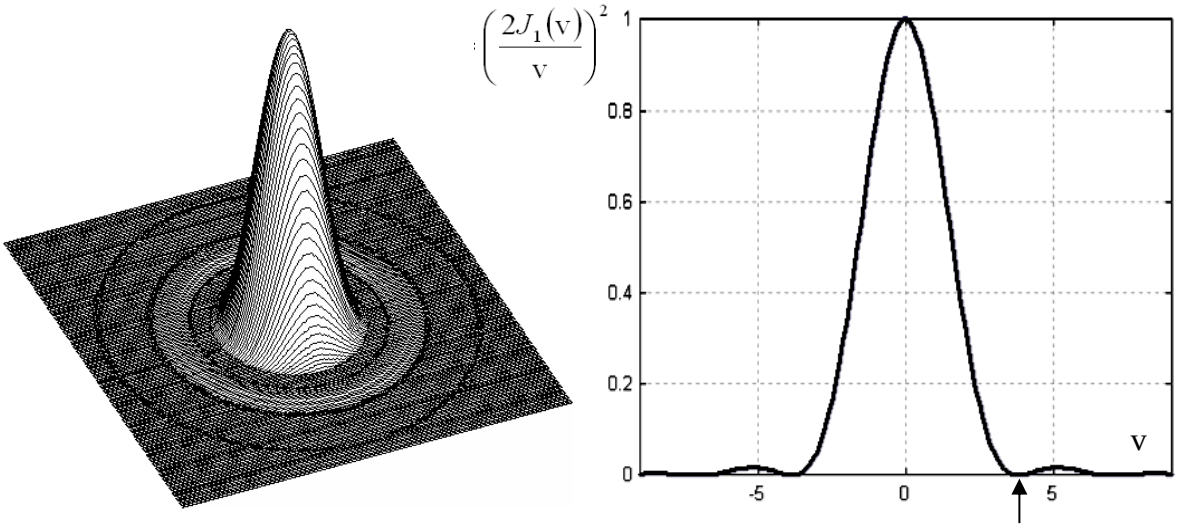


♣ The term resolution is often erroneously used when describing the number of pixels in digital cameras. This number has nothing to do with resolution, however.

Mathematically the light intensity distribution in the image plane can be described by the equation

$$I(r) = \left(\frac{2J_1(\mathrm{v})}{\mathrm{v}} \right)^2 ,$$

where J_1 is a Bessel function of the first order, and v is a normalized optical coordinate. This intensity distribution function is derived in many textbooks on optics. Assuming a small angle β in the figure on previous page, we get $\mathrm{v} \approx \frac{\pi D r}{\lambda f}$, where λ is the wavelength and *D*, *f* and *r* are defined in the figure on previous page. The intensity is a function of only a radial coordinate, *r*, which means that the pattern is rotationally symmetric around the optical axis. It consists of a central maximum surrounded by weak concentric rings as illustrated in the figure below.



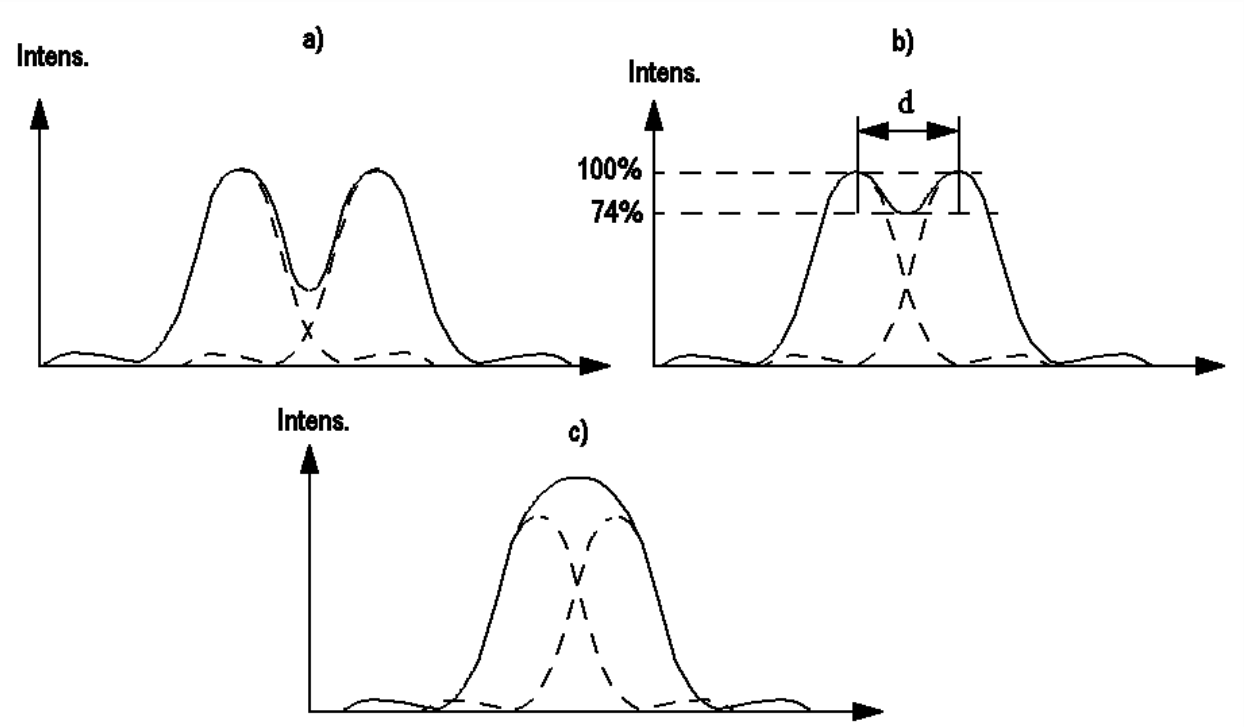
zero at 3.83

Since 84% of the light intensity is to be found in the central peak with radius v = 3.83 (called the Airy spot), this distance is often taken as a measure of the size of the diffraction pattern. v = 3.83 corresponds to a radial distance in the image plane of $r = \frac{1.22\lambda f}{D}$. We will denote this radial distance to the first intensity minimum by *R*. A point object is thus reproduced as a blurred disk or spot with a radius of approximately *R* . To obtain a small disk, we should use a short wavelength and a lens with a low value of $\frac{f}{D}$♣. (The ratio $\frac{f}{D}$ is referred to as "speed" in connection with photographic lenses, because it influences the intensity level in the image and

♣ From the equation it would appear that there is no theoretical lower limit to the size of the Airy spot; by reducing the ratio $\frac{f}{D}$ one could in principle get an arbitrarily small *R*. In reality the equation is not valid for very small ratios $\frac{f}{D}$. The exact equation is $R = \frac{0.61\lambda}{\sin\beta}$, where β is given in the figure on previous page. sinβ is called the (image side) numerical aperture of the lens.

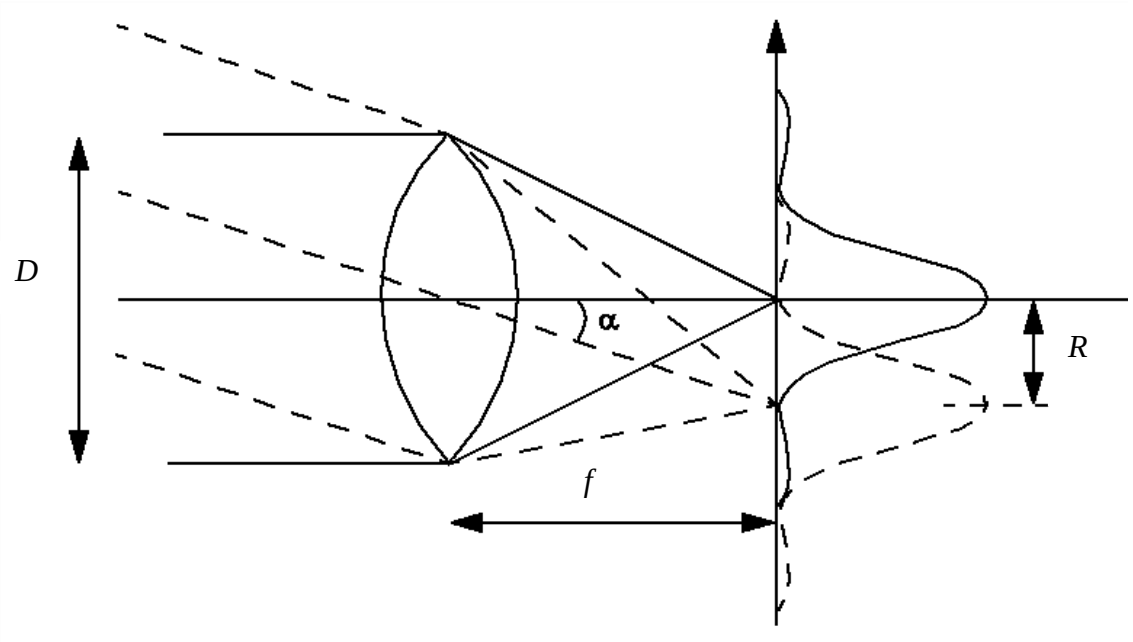
thus the shutter speed.) The intensity distribution in the image plane when reproducing a point object is called the **point spread function**, and is abbreviated ***psf***.

The resolution of an optical instrument is often defined as its ability to image two point objects of equal intensity that are located close together. In the astronomical example mentioned previously, we can imagine that we are looking at a binary star, whose components are separated by an angular distance α. In the image plane, we will get two overlapping diffraction patterns, one from each star. Depending on the angular distance, α, we can have three different cases according to the figure below. We assume that the imaging is incoherent, so that light intensities add linearly.



Case a) is resolved, there is no doubt that we can see two stars. In case b), we can just about see two objects, and in case c) we will see only one light spot. The resolution according to the Rayleigh criterion is equivalent to case b), with a reduction in intensity of 26% between the peaks. The choice of 26% is somewhat arbitrary, but it is used partly because it is approximately what the human eye needs to detect a reduction in intensity, and partly because it is mathematically convenient, as the distance d is then the same as the distance *R* to the first minimum in the *psf*.

Example: Calculate the resolution (expressed as an angular distance α) of a telescope with a lens diameter of 60 mm, for light of wavelength λ = 500 nm.



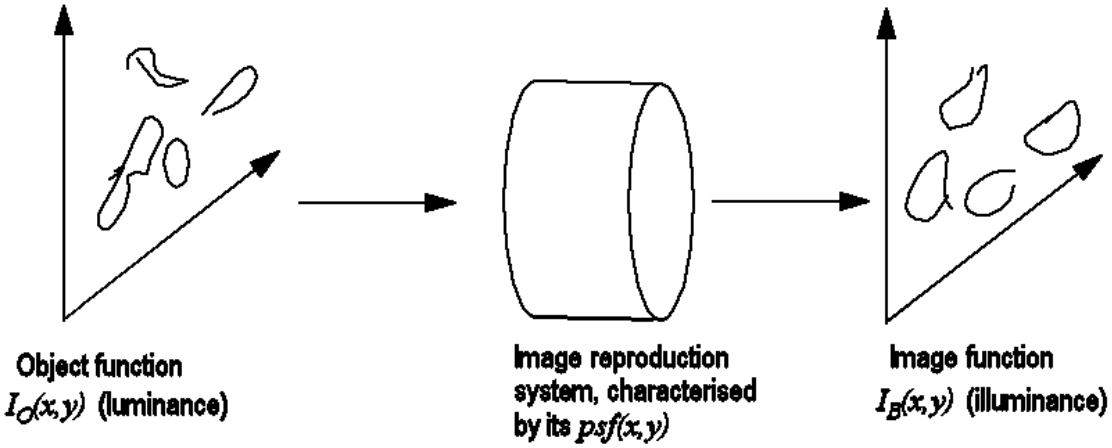
$\alpha \approx \frac{R}{f} = \frac{1.22\lambda}{D} = \frac{1.22 \times 500 \times 10^{-9}}{60 \times 10^{-3}} = 1.0 \times 10^{-5}$ radians, which is equal to 5.8×10^{-4} degrees or 2.1 arc seconds (approximately the angle subtended by a Swedish 1 kr coin, or an American quarter dollar, when viewed from a distance of 2.5 km!). The resolution, expressed as angle α, is thus determined solely by the wavelength of the light and the diameter of the lens. One may then wonder why telescopes magnify at all! The reason is that our eyes, which also have limited resolution, need a sufficiently large image to see all the fine details produced by the telescope lens.

The determination of the resolution of microscopes is a bit more complicated (the resolution is here expressed as the distance between two point objects in the specimen). The result is, however, a theoretical limit which is about $\lambda/2$, i.e. about 0.2 μm for visible light.

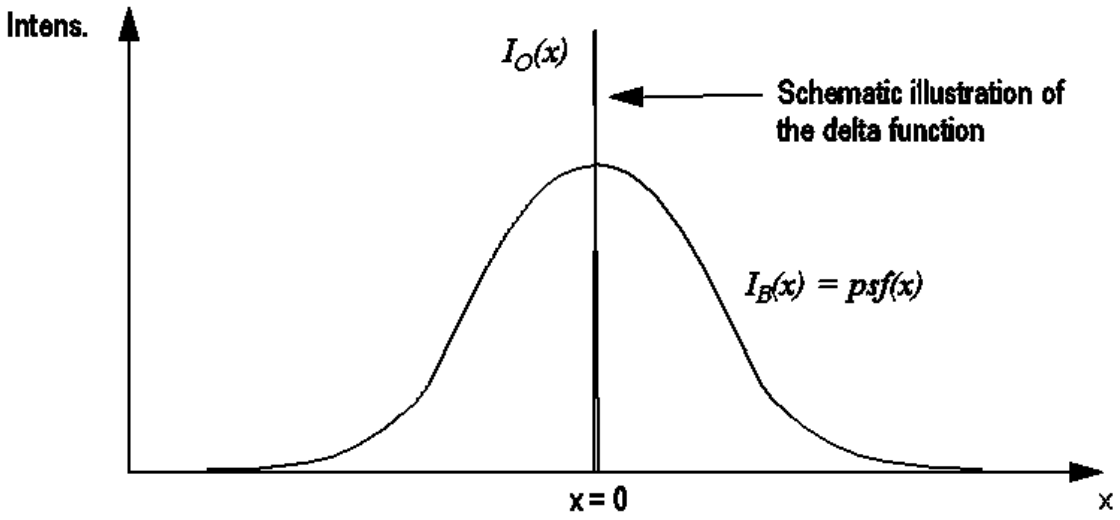
The expressions derived above for the resolution are valid for diffraction-limited optics, i.e. the optical aberrations are negligible. In reality, there will always be a certain degree of optical aberration which will degrade the resolution. The smallest aberrations can, of course, be expected from the most expensive optics.
Despite the fact that the resolution provides a measure of the performance of an optical instrument, the information it contains is somewhat limited. In principle, the only information it provides is how well the instrument can reproduce an image of two equally bright point objects. It provides no direct information on the image quality that can be expected for other kinds of objects. We will soon see, however, that the ***psf*** **contains all information on the imaging of an arbitrary object**. It is simply a case of using the information in a more intelligent way, than when determining the resolution. We will also see that it is possible to incorporate the effects of other factors on image quality, such as the detector characteristics, blurring due to motion, etc. We must first, however, introduce a mathematical description of the imaging process.

9. Mathematical Representation of the Image Reproduction Process

Let us assume that we are using an optical instrument to produce an image of an object with an arbitrary intensity distribution. Let us also assume that the point spread function, *psf*, is known. Our objective now is to calculate the intensity distribution of the image. The imaging process is schematically illustrated below.

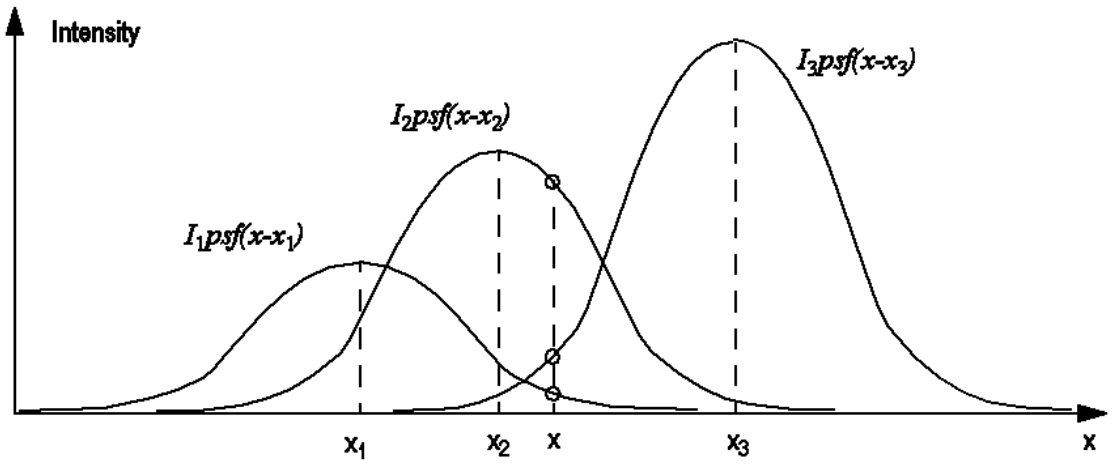


For the sake of simplicity, let us assume that the imaging scale is 1:1, and consider only one-dimensional functions $I_0(x)$, $I_B(x)$ and $psf(x)$. This is done only to simplify the expressions, and will not limit the applicability of the results. $I_0(x)$ represents the object luminance value, and $I_B(x)$ the image illuminance value. For convenience, we will in the mathematical treatment use the term ***intensity*** for both of these quantities. Let us first consider the case in which the object is a point, i.e. $I_0(x) = \delta(x)$ where δ is the delta function (Appendix 1). This will give the following function.



Assume now, that the object consists of an arbitrary number of point objects, located along the x-axis at $x_1, x_2, \ldots x_n$, with intensities I_1, I_2, etc. Each of these point objects will produce an image whose shape is identical to the *psf* shown in the figure above. The heights of the individual

curves, however, will vary depending on the intensities of the different point objects, see illustration below.



Assuming incoherent imaging, we can now calculate $I_B(x)$ at an arbitrary x-coordinate by adding the contributions from the images of the various point objects:

$$I_B(x) = I_1 \cdot psf(x - x_1) + I_2 \cdot psf(x - x_2) + I_3 \cdot psf(x - x_3) + \ldots = \sum_{k=1}^{n} I_k \cdot psf(x - x_k)$$

An arbitrary ***continuous*** object function, $I_O(x')$, can be regarded as the sum of infinitesimally close point objects. In the above expression, we will allow x_k to become x', which can take an arbitrary value, and I_k to become $I_O(x')$, which represents the intensity at x'. We can now write:

$$I_B(x) = \int_{-\infty}^{+\infty} I_O(x') \cdot psf(x - x')dx'$$

i.e. the image function is the convolution of the object function and the point spread function (see Appendix 1). Integration is carried out from $-\infty$ to $+\infty$ in the formula, but in reality the contribution to the integral will be negligible beyond a certain limiting value of x'. In order for the above expression for $I_B(x)$ to be correct, the *psf* must be uniform over the whole image plane. This is a simplification of the real situation, as the *psf* is often more diffuse at the edges of the image than in the center. As a consequence of this, it is sometimes necessary in practice to use slightly different expressions for *psf* depending on the x coordinate, i.e. where in the image field the image function is to be described. We will, however, not deal with this complication.

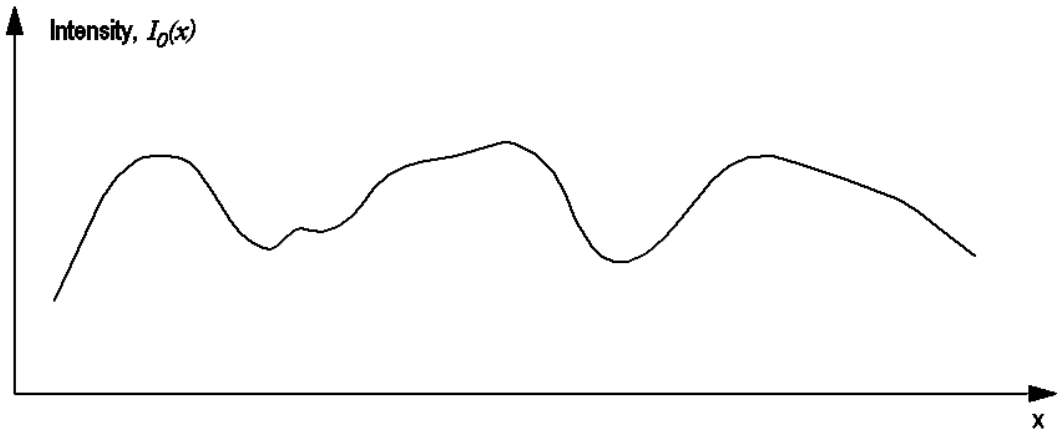
Another requirement in order to be able to calculate I_B from the convolution of I_0 and the *psf* is that the light intensities add linearly in the image plane. This is not always the case in practice, for example, if the object is illuminated by coherent laser light. We will, however, assume in the following that the image reproduction is incoherent so the demand on linearity is met.
In reality, the image function is, of course, two-dimensional, and can be described by a generalization of the earlier expression:

$$I_B(x,y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} I_O(x',y') \cdot psf(x-x',y-y')dx'dy'$$

We have now justified our earlier statement that if we know the *psf*, we can calculate the image function for an arbitrary object function. We have thus completely fulfilled our objective, namely to describe the quality of an image reproduction process. The above expressions are, however, not particularly suitable for practical purposes, such as lens tests in a photographic magazine. To find a more appropriate form for the quality information, we will apply Fourier transforms. But in order to understand the meaning of what we are doing, let us first start by looking at what a Fourier transform of a function actually represents.

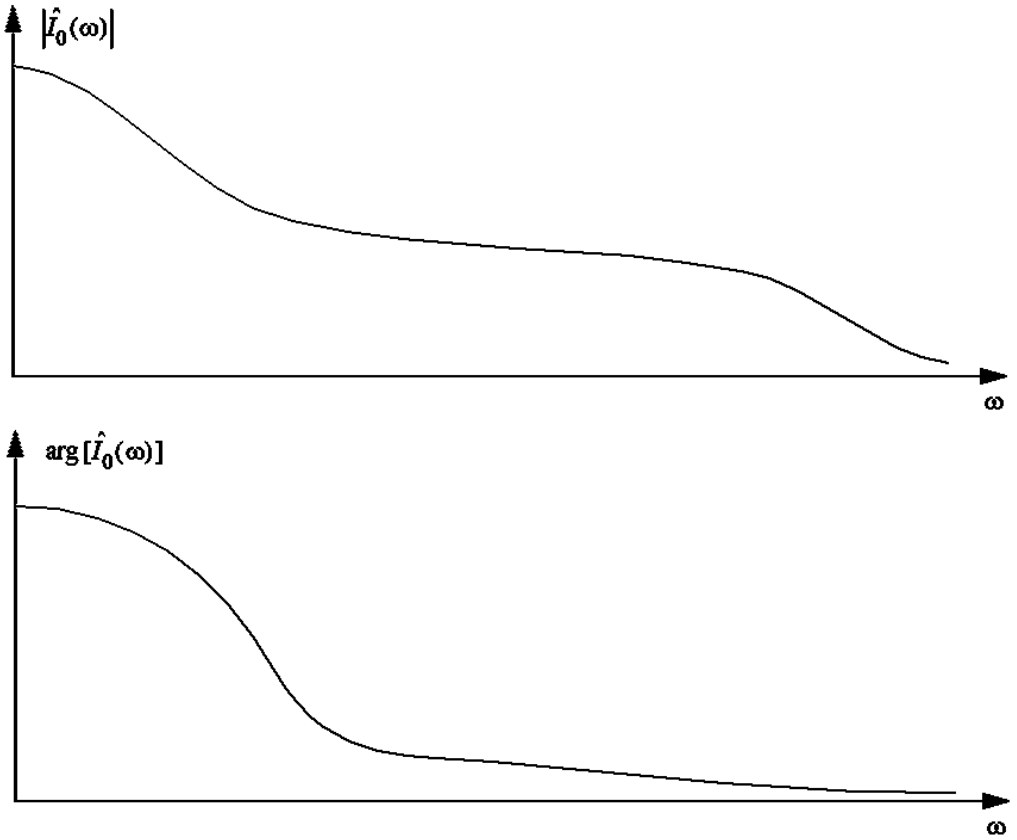
10. The Physical Interpretation of a Fourier Transform

For the sake of simplicity, we will still consider functions of one variable, but the interpretations below can also be extended to higher dimensions. Let us consider the function $I_O(x)$, which may, for example, represent the light intensity of an object to be imaged.



We obtain the Fourier transform of $I_O(x)$, which according to the definition is:

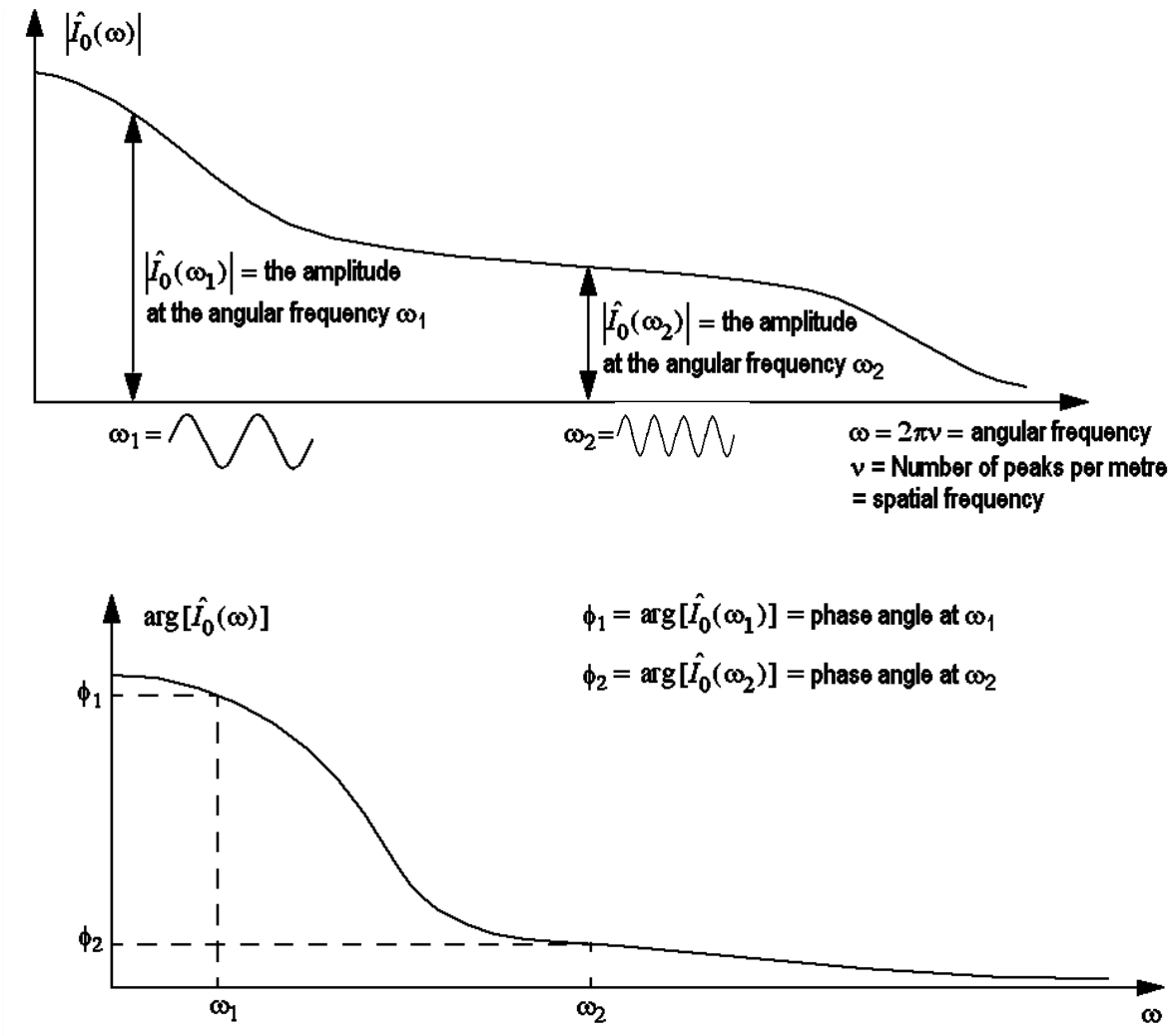
$\hat{I}_O(\omega) = \int_{-\infty}^{+\infty} I_O(x)e^{-i\omega x}dx$. The Fourier transform is usually a complex function, and we denote its absolute value, or modulus, $\left|\hat{I}_O(\omega)\right|$ and argument $\arg\left[\hat{I}_O(\omega)\right]$. Therefore, $\hat{I}_O(\omega) = \left|\hat{I}_O(\omega)\right| \cdot e^{i\arg\left[\hat{I}_O(\omega)\right]}$. The modulus and argument functions may have the following appearance.



These two functions have a simple physical interpretation:

We know from the theory for Fourier series that a periodic function can be described as a sum of harmonic oscillations of different frequencies (the fundamental tone and its harmonics)♣. Since we are dealing with functions that vary in ***space*** rather than in ***time***, frequencies are called spatial frequencies (unit m^{-1}). In the same way, a non-periodic function can be described as a sum (or, more correctly, an integral) of infinitely closely spaced harmonics. The Fourier transform provides information on the amplitude and phase angle of the harmonic components required to create the original function (the Fourier transform is often called the spectrum of the function). This is illustrated in the figures and the text on next page. (The description given is simplified. A more comprehensive and correct account is given in Appendix 1.)

♣ See Appendix 1.



To obtain $I_O(x)$ we sum all the harmonics:

$$I_O(x) = \left|\hat{I}_O(\omega_1)\right| \cdot \cos\left(\omega_1 x + \arg\left[\hat{I}_O(\omega_1)\right]\right) + \left|\hat{I}_O(\omega_2)\right| \cdot \cos\left(\omega_2 x + \arg\left[\hat{I}_O(\omega_2)\right]\right) + \ldots$$

or, expressed in a more mathematically correct way:

$$I_O(x) = k \cdot \int_0^\infty \left|\hat{I}_O(\omega)\right| \cdot \cos\left(\omega x + \arg\left[\hat{I}_O(\omega)\right]\right) d\omega$$

In the figures only the positive spatial frequencies are shown, while the Fourier transform is defined for both positive and negative frequencies. For a real function $I_O(x)$ (and we have no reason to work with any other kind)

$$\hat{I}_O(\omega) = \hat{I}_O^*(-\omega)$$

where the asterisk, *, indicates the complex conjugate. This means that the negative ω-axis contains exactly the same information as the positive ω-axis. In other words, it is sufficient to

consider the positive values of ω. A more comprehensive description of both Fourier transform and convolution is given in Appendix 1.

After this short description of the meaning of the Fourier transform, we are ready to return to optical imaging and to introduce the concept of the optical transfer function.

11. The Optical Transfer Function

We have previously seen that it is possible to write the image function as:

$$I_B(x,y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} I_O(x',y') \cdot psf(x-x', y-y')dx'dy'$$

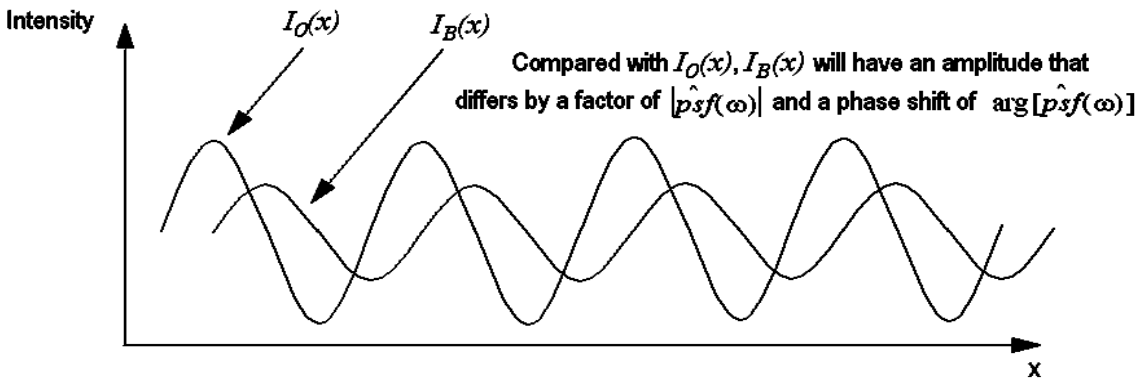
or, in short $I_B = I_O \otimes psf$, where the symbol $\otimes$ means convolution♣. If we take the Fourier transform (FT) of this expression, we obtain $\hat{I}_B = \hat{I}_O \cdot \hat{psf}$ ♥. Here we have used the fact that convolution is transformed into multiplication during Fourier transformation. Bearing in mind what was discussed previously regarding FT, it is easy to understand the meaning of $\hat{I}_B = \hat{I}_O \cdot \hat{psf}$. We note that:

$$\left|\hat{I}_B(\omega)\right| = \left|\hat{I}_O(\omega)\right| \cdot \left| \hat{psf}(\omega)\right|$$

and

$$\arg\left[\hat{I}_B(\omega)\right] = \arg\left[\hat{I}_O(\omega)\right] + \arg\left[\hat{psf}(\omega)\right]$$

This means that the amplitudes of all the frequency components in the object function are modified by multiplication by $\left|\hat{psf}(\omega)\right|$ to obtain the amplitude in the image function. In the image, these components will have a phase shift of $\arg\left[\hat{psf}(\omega)\right]$ relative to the equivalent component in the object function.



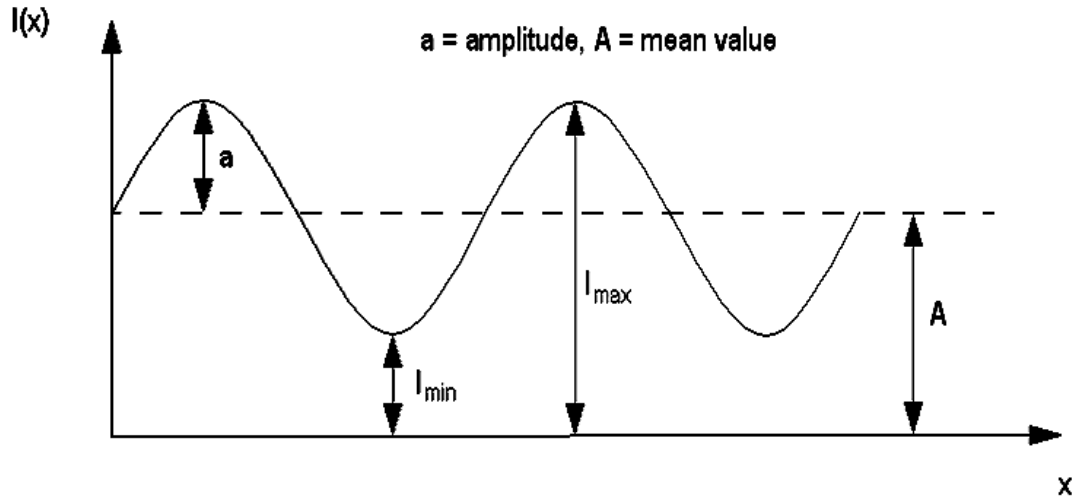
♣ In mathematics literature convolution is often denoted by the symbol $*$

♥ Using this equation, so-called deconvolution can be performed to compensate for image degradation caused by the *psf*. See Appendix 6

$\hat{psf}(\omega)$ is called the Optical Transfer Function, *OTF*, and its absolute value, or modulus, $|OTF(\omega)| = |\hat{psf}(\omega)|$ is called the Modulation Transfer Function, *MTF*. The argument, $\arg[OTF(\omega)] = \arg[\hat{psf}(\omega)]$ is called the Phase Transfer Function, *PTF*.

It is common practice to normalize *OTF* to the value 1 for $\omega = 0$.

After this normalization, the *MTF*, i.e. the absolute value (modulus) of the *OTF*, is a direct measure of how much the degree of modulation of a given spatial frequency in the image differs from that in the object. If, for example, the *MTF* is 0.50, this means that the modulation in the image is half that of the object. The definition of the degree of modulation is given in the figure below.

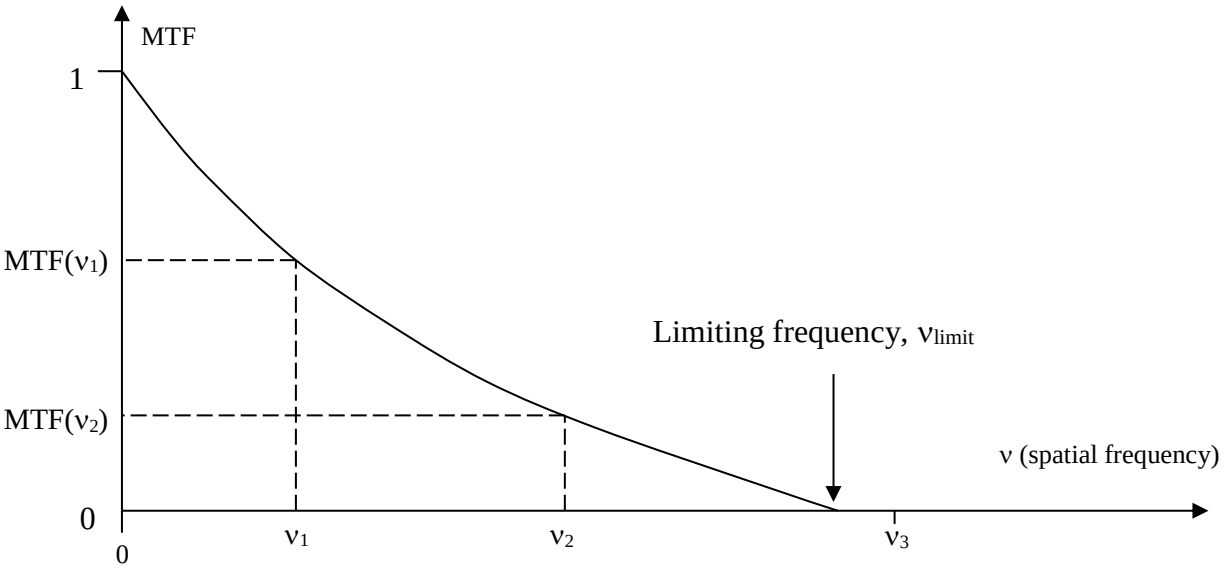


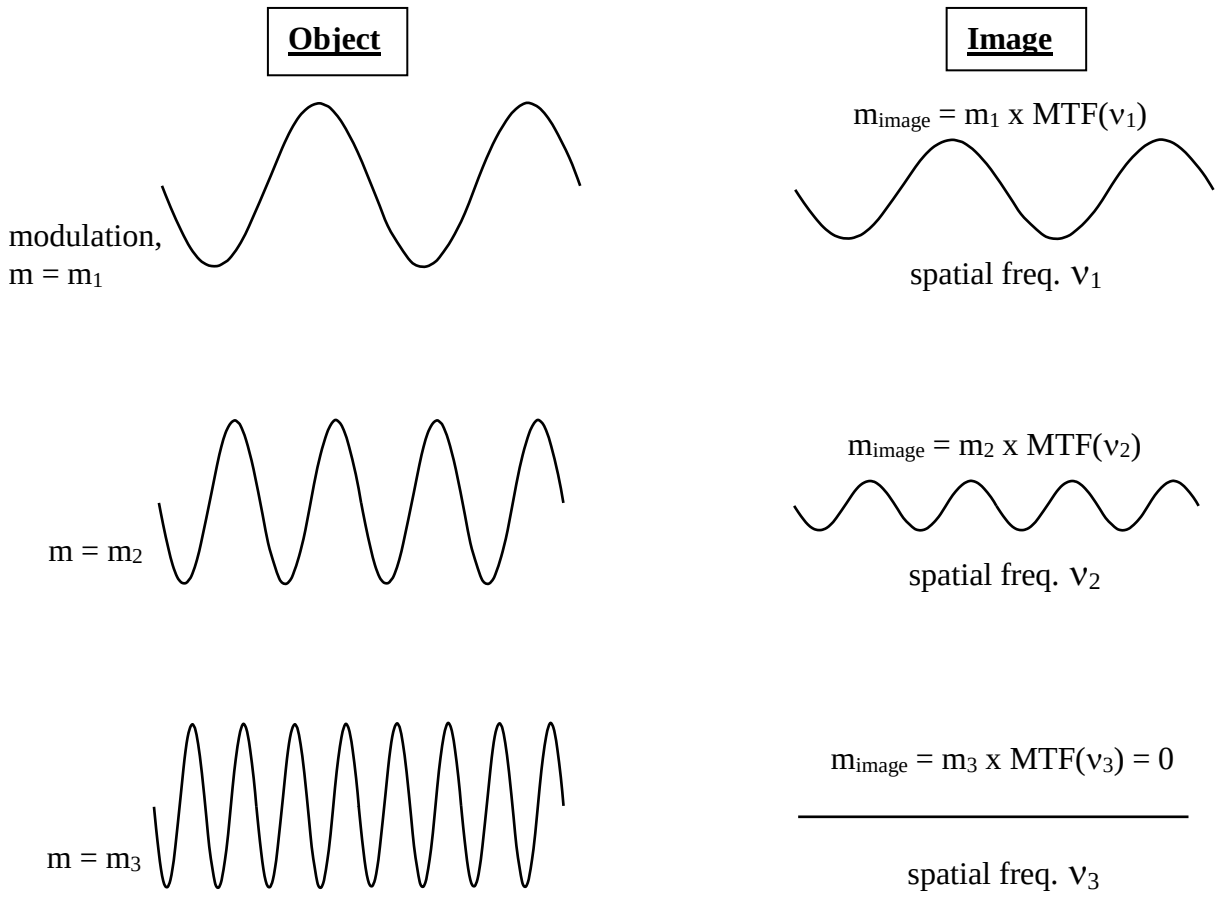
Degree of modulation $= \dfrac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} = \dfrac{a}{A} = m$

In layman‘s terms, one can say that the degree of modulation is a measure of the contrast in the pattern. A high degree of modulation gives a pattern with high contrast, while a low degree of modulation gives a pattern with low contrast.

The quality of imaging systems is often illustrated with the help of *OTF* (*MTF* and *PTF* curves). It is more common to measure the *MTF* than the *PTF*, as the latter is more difficult both to measure and to interpret. However, this involves some loss of information on the imaging quality.

An example of an *MTF* curve is shown on next page, together with the effects it will have on the imaging of patterns with different spatial frequencies. Instead of the angular frequency, ω, one usually uses the spatial frequency $\dfrac{\omega}{2\pi}$, which is denoted by ν (a letter in the Greek alphabet, which is pronounced “new”).





If the degree of modulation in the object is not 100%, but has a lower value *m*, then the degree of modulation in the image will be $m \cdot MTF(\nu)$. A perfect image would be obtained if $MTF(\nu)$ were equal to 1 for all spatial frequencies, as this gives the same contrast in the object and the image for all spatial frequencies. A perfect *PTF* is equal to zero for all spatial frequencies. (Actually it can be shown that $PTF(\nu) = K \cdot \nu$, where *K* is a constant, is also a perfect *PTF*. The only difference is that for $K \neq 0$, the image is shifted laterally, i.e. sideways)

When developing the *OTF* theory, we assumed that the imaging scale was 1:1. This is, of course, usually not the case. One may then ask whether the spatial frequencies in *MTF* and *PTF* curves relate to the object or image planes. The answer is that nearly always they relate to the ***image*** plane. This is the case, for example, for *MTF* curves for photographic lenses (one exception is microscopic lenses, where the spatial frequencies refer to the specimen plane). So, unless anything else is said, we can usually assume that spatial frequencies refer to the image plane.

It is often necessary to transform spatial frequencies in the object to spatial frequencies in the image plane, or vice versa.

Example: We are using a 50 mm photographic lens for imaging, at a distance of 5.0 meters, a periodic pattern with a period length of 1.0 cm (spatial frequency 1.0×10^{2} m^{-1}). What spatial frequency will the image of this pattern have in the image plane of the camera?

Solution: The object distance, 5.0 meters, is much larger than the focal length. Therefore, the imaging scale is, to a good approximation, given by $\frac{\text{focal length}}{\text{object distance}} = 1.0\times10^{-2}$. This means that the image of the pattern has a period length of 0.10 mm, and therefore the spatial frequency is 1.0×10^{4} m^{-1} or 10 mm^{-1} (for photographic lenses, spatial frequencies are often given in units of mm^{-1} rather than m^{-1} to avoid excessively large numbers). After this scale transformation, we can easily use the *MTF* curve of the photographic lens to find out the amount of modulation loss that we get for the pattern that we are imaging.

12. The *OTF* for a Diffraction-Limited Lens

As we have seen earlier, the *OTF* is the Fourier transform of the *psf*. The *psf* and *OTF* for a diffraction-limited lens is derived in many optics textbooks. It is found that the *OTF* is real and positive, i.e. *OTF* = *MTF* (*PTF* = 0). It can be described by the following equation:

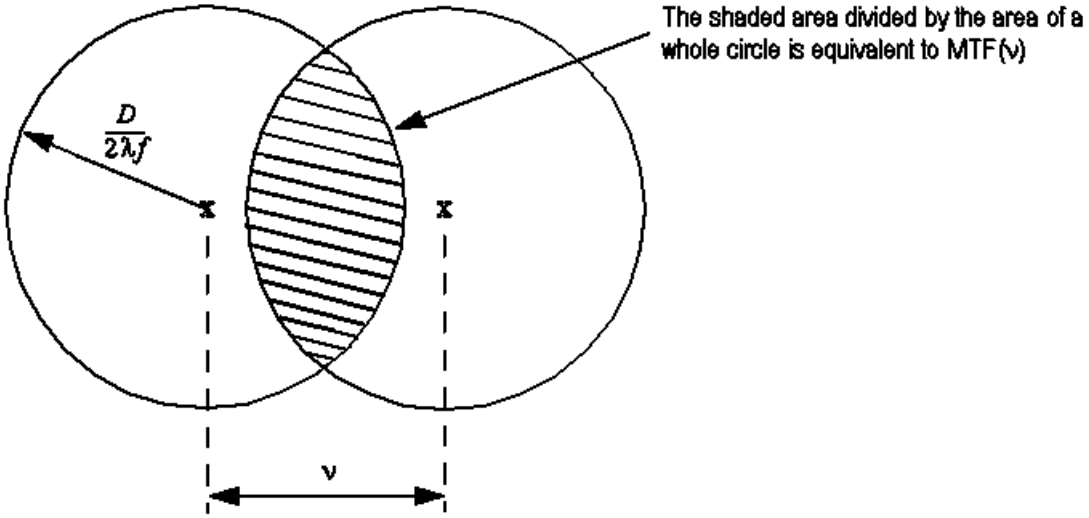
$$MTF(x) = \frac{2}{\pi}\left[\arccos(x) - x\sqrt{1-x^{2}}\right]$$

$x = \frac{\nu}{\nu_{\text{lim}}}$, where ν is the real spatial frequency and ν_{lim} = limiting frequency = $\frac{D}{\lambda f}$.

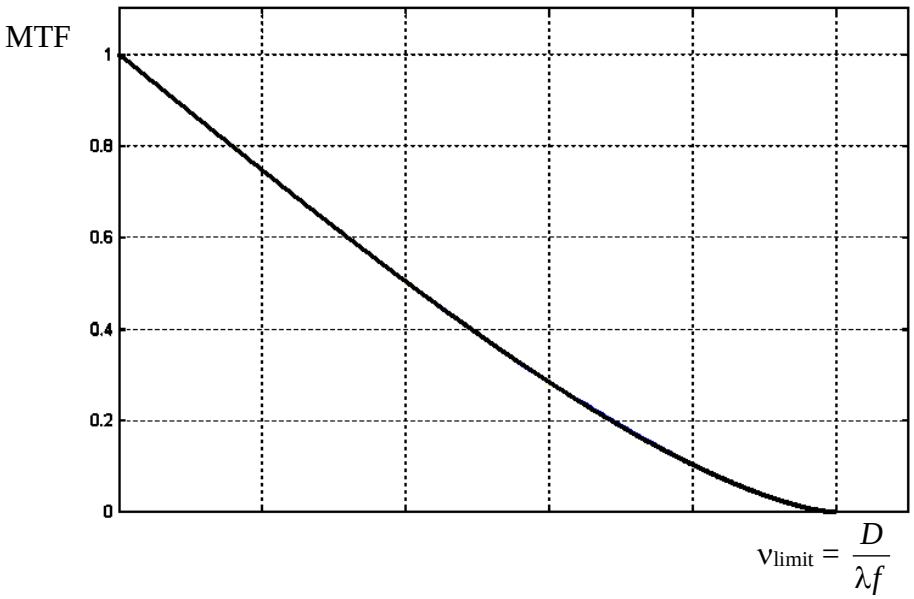
It is also found that the *MTF* can be calculated using the following simple geometrical method:

Draw a figure of the same shape as the lens opening, but scaled by $\frac{1}{\lambda f}$, where λ is the wavelength of the light and *f* is the focal length. That is to say, for a circular lens with a diameter *D*, we will obtain a circle with a diameter $\frac{D}{\lambda f}$ (unit m^{-1}, i.e. the same as for spatial frequency).

If we wish to find the *MTF* at a spatial frequency ν, we draw another circle of the same diameter, shifted laterally by ν. We then determine the area of overlap of the two circles and divide this by the area of a whole circle (in order to normalize *MTF* to 1 at ν = 0), see illustration below.



The result is shown in the figure below:

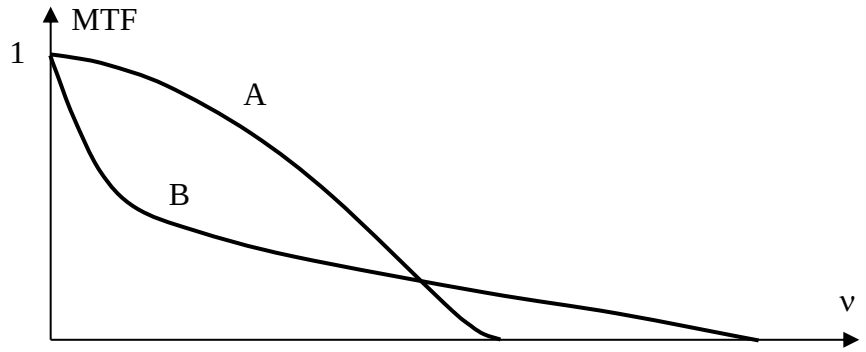


The highest spatial frequency that can be reproduced by the lens (the limiting frequency) is $\nu_{limit} = \frac{D}{\lambda f}$ ♣. All optical systems have such a limiting frequency, above which the *MTF* value is identical to zero. The resolution limit according to the Rayleigh criterion gives approximately the same information as the limiting frequency of the *MTF* curve, ***but*** the Rayleigh criterion tells us nothing about the modulation loss at lower frequencies. High *MTF* values at

---

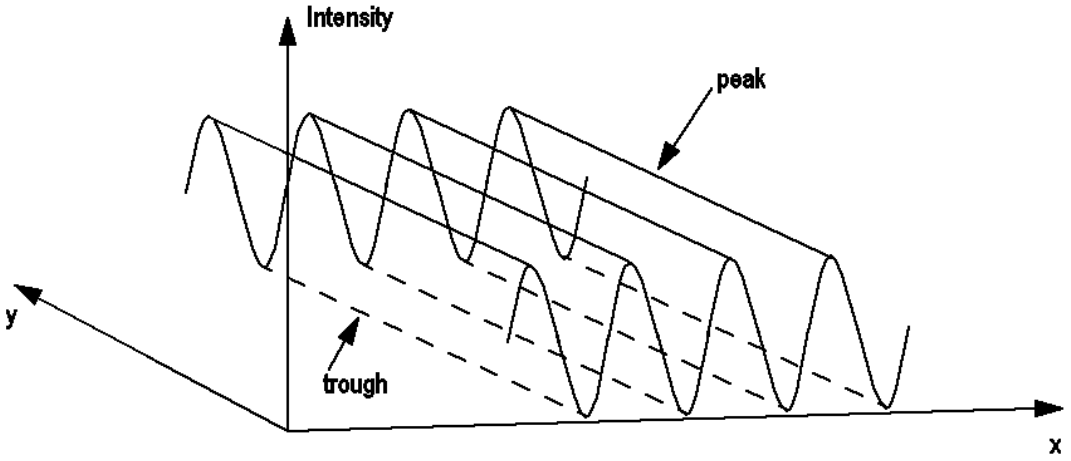
♣ The equation is an approximation which gives good results for $\frac{D}{f} < \mathrm{approx.} 0.7$. The exact equation is $\nu_{limit} = \frac{2\sin\beta}{\lambda}$, where β is defined in the figure on page 17.

comparatively low frequencies turn out to be very important for our impression of image sharpness, contrast and overall quality. In most practical cases this is more important than to have a high limiting frequency. An example of two *MTF* curves representing different optical systems is given in the figure below. For normal imaging purposes system A is the preferred one although it has a considerably lower limiting frequency. If we only had access to Rayleigh resolution data for the two systems, we would conclude that system B was the superior one. This shows how limited, and even misleading, the information can be that we get from the resolution number. We need the entire *MTF* curve to judge the imaging quality.

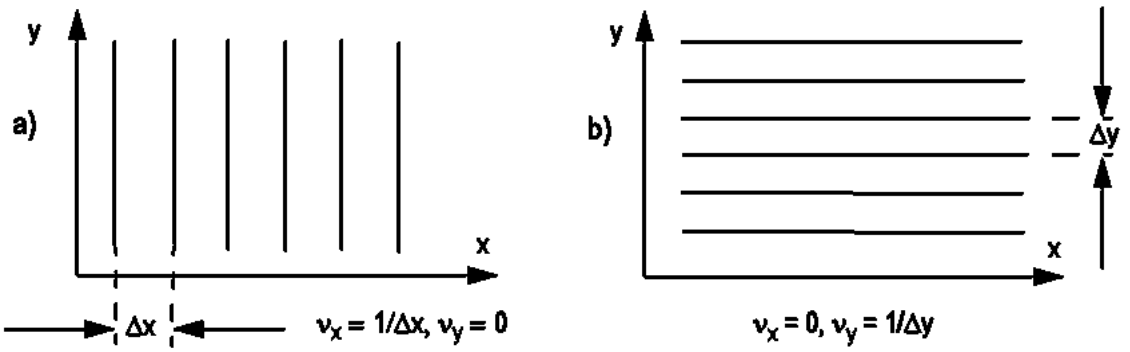


13. The Two-Dimensional *OTF*

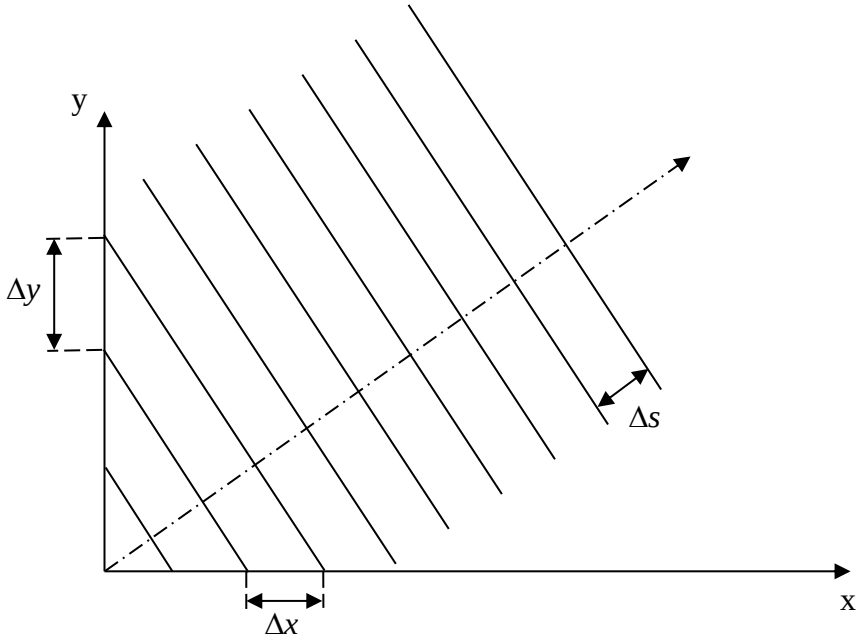
The object and image functions, as well as the *psf*, are functions of two spatial variables, x and y. This means that the *OTF* is given by the two-dimensional Fourier transform (see chapter 14) of the *psf*. The *OTF* is therefore a function of two spatial frequency variables, ν_x and ν_y. How can these variables be physically interpreted? For the sake of simplicity, let's look at a simple sinusoidal intensity pattern that varies in the x-direction, and is constant in the y-direction, see illustration below.



To simplify the following figures, we will show only the xy plane, seen from above, and the waves will only be illustrated by the straight lines representing the peaks. The figure above is then replaced by figure a) on next page. A similar pattern, but with variations in the y-direction, is shown in b).



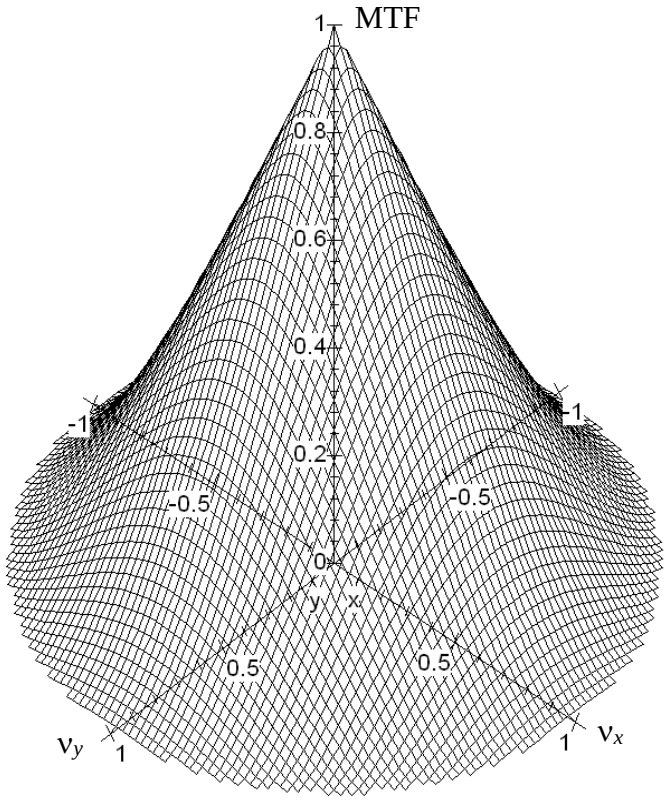
When describing patterns with a variation in two dimensions (x,y) we use two spatial frequency components, ν_x and ν_y. In figure a) we have a sinusoidal pattern whose variation is in the x-direction only (if, for a constant value of x, we travel in the y-direction we will see a constant intensity value). Therefore, the period length in the y-direction is infinite, which corresponds to a spatial frequency of zero, i.e. $\nu_y = 0$. In the x-direction we have assumed a period length of Δx, which corresponds to a spatial frequency of $(\Delta x)^{-1}$, i.e. $\nu_x = (\Delta x)^{-1}$. Figure b) shows a similar pattern that varies in the y-direction only. Turning to a more general case, we look at the figure below.



The spatial frequency components in the x and y directions are given by $\nu_x = \frac{1}{\Delta x}$ and $\nu_y = \frac{1}{\Delta y}$. In a direction indicated by $-\cdot-\cdot-\cdot\rightarrow$, perpendicular to the lines in the pattern, the spatial frequency is given by $\nu = \frac{1}{\Delta s} = \sqrt{\nu_x^2 + \nu_y^2}$. In the figure above the "direction vector" $-\cdot-\cdot-\cdot\rightarrow$ of the wave pattern is pointing in the positive x and y directions. Both ν_x and ν_y are then positive. If instead the "direction vector" points like this $\searrow$ ν_x will be positive, but ν_y will be negative.

A pattern with arbitrary direction and period length can thus be described completely by the two spatial frequency components ν_x and ν_y. From these components we get both period length and direction of the pattern. A more mathematical description of this subject is given in chapter 14.

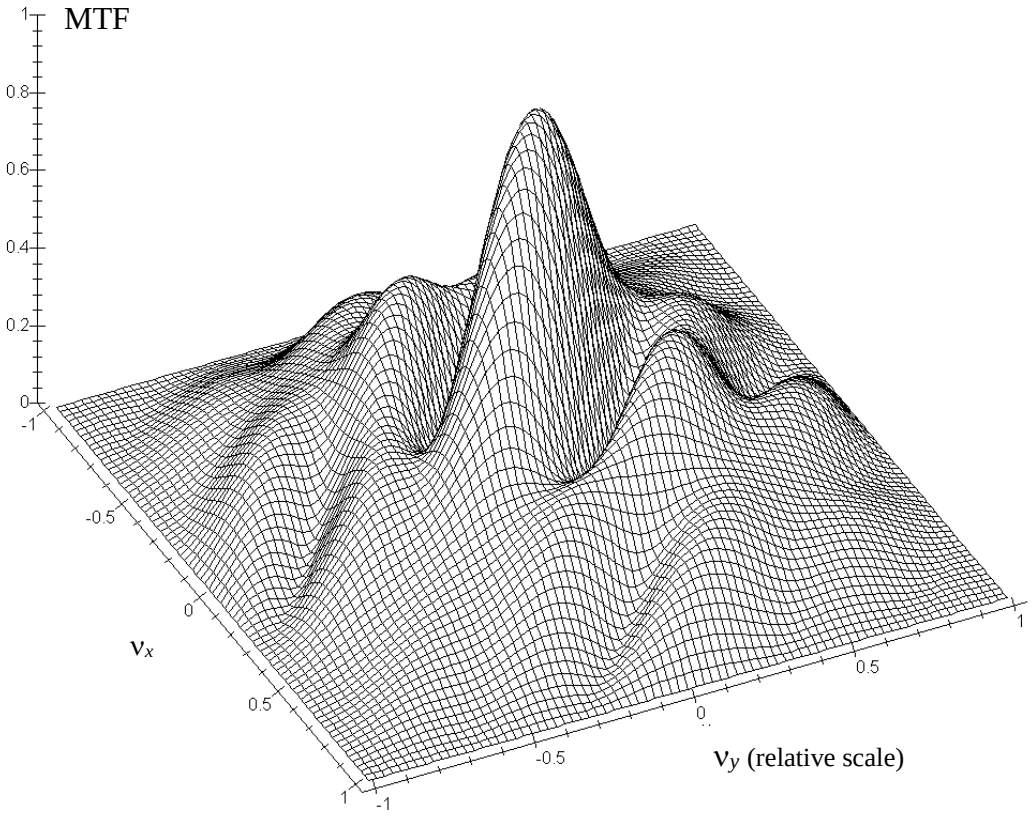
The meaning of the 2-dimensional (2D) *MTF*, is that the degree of modulation for a pattern with spatial frequency components ν_x and ν_y will be $MTF(\nu_x,\nu_y)$ times lower in the image compared with the object. The phase shift between image and object for the same pattern is given by $PTF(\nu_x,\nu_y)$. If the *psf* is rotationally symmetric, the *MTF* will also be rotationally symmetric and the $PTF = 0$. This is the case for a diffraction-limited lens with circular aperture. In this case the *MTF* will look like a "circus tent" as shown in figure below. The spatial frequency components ν_x and ν_y are given in units of $\frac{D}{\lambda f}$. The one-dimensional *MTF* shown in chapter 12 is simply a vertical cut going through the origin of the 2D *MTF*.



For rotationally symmetric *psf*s, producing rotationally symmetric *MTF*s, there is obviously no need to plot *psf* or *MTF* in 2D. All information can be obtained from an ordinary *MTF* curve like the one shown in chapter 12. For non-symmetric *psf*s, it is necessary to consider the 2D *MTF* to understand how patterns of different orientations are reproduced in the image. A non-symmetric *psf* will, to some extent, occur in all optical systems as one leaves the image center and moves towards the edge of the image. The reason for this is that off-axis aberrations like astigmatism and coma, producing non-symmetric *psf*s, cannot be completely eliminated even in high-quality optics. In the image center, however, the *psf* is usually rotationally symmetric even for lenses that are not diffraction-limited. The reason for this is that on-axis aberrations

(spherical aberration and longitudinal chromatic aberration) are rotationally symmetric around the optical axis.

An example of a 2D *MTF* that is not rotationally symmetric is seen in the figure below. It was obtained by roughly simulating the *psf* obtained in the presence of coma, i.e. the *psf* looks like a tiny comet with a bright head and a diffuse tail extending in one direction. The non-symmetric shape of the *MTF* means that pattern orientation, not just frequency, will determine how well it is reproduced in the image. The *PTF* is not equal to zero in this case, which furthermore complicates the imaging properties.



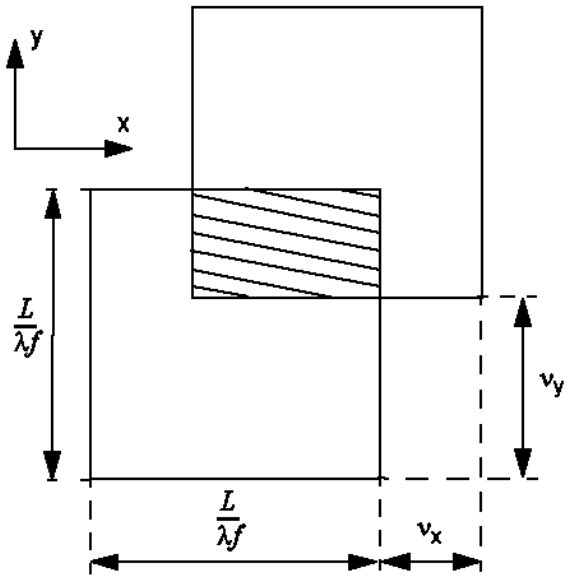
Even in cases where the 2D *MTF* is not rotationally symmetric, it still has a pronounced symmetry in the $\nu_x\nu_y$ plane as seen in the figure above. This is a mathematical property of Fourier transforms of real (i.e. non-complex) functions; changing sign of both ν_x and ν_y will change the transform into its complex conjugate value. It is therefore sufficient to display *MTF* in one half of the $\nu_x\nu_y$ plane. For example one can skip all negative ν_x or ν_y values.

As we shall see in chapter 15, the influence of the detector can also be described in terms of *psf* and *MTF*. For a detector these functions are, in general, not rotationally symmetric, and therefore the total *MTF* for the imaging system♣ (including both optics and detector) is usually not rotationally symmetric even in the image center.

♣ This will be described in chapter 16.

For diffraction-limited lenses the 2D *MTF* can be calculated using the same geometrical method that was described in chapter 12. We will illustrate this with an example of a lens having a square aperture.

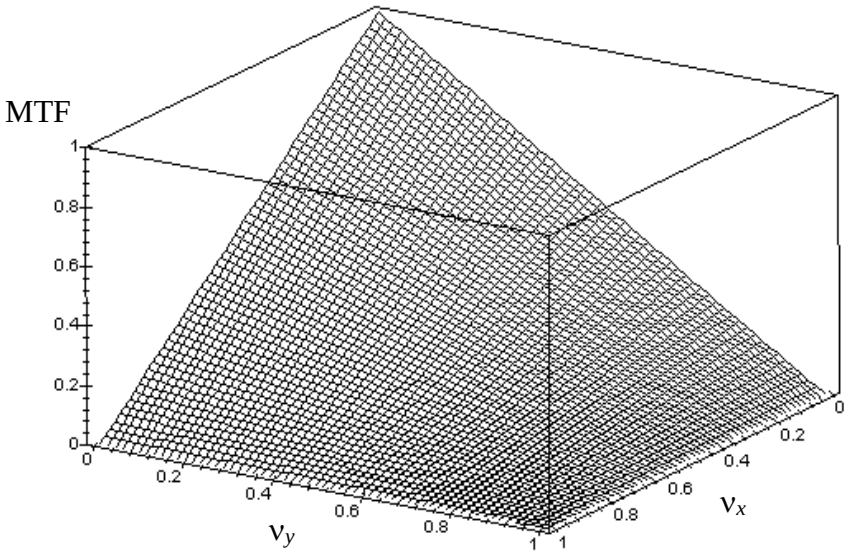
Example: Determine $MTF(\nu_x,\nu_y)$ for a diffraction-limited lens with a square aperture, L x L, and a focal length f, where the sides of the aperture are parallel to the x and y directions.



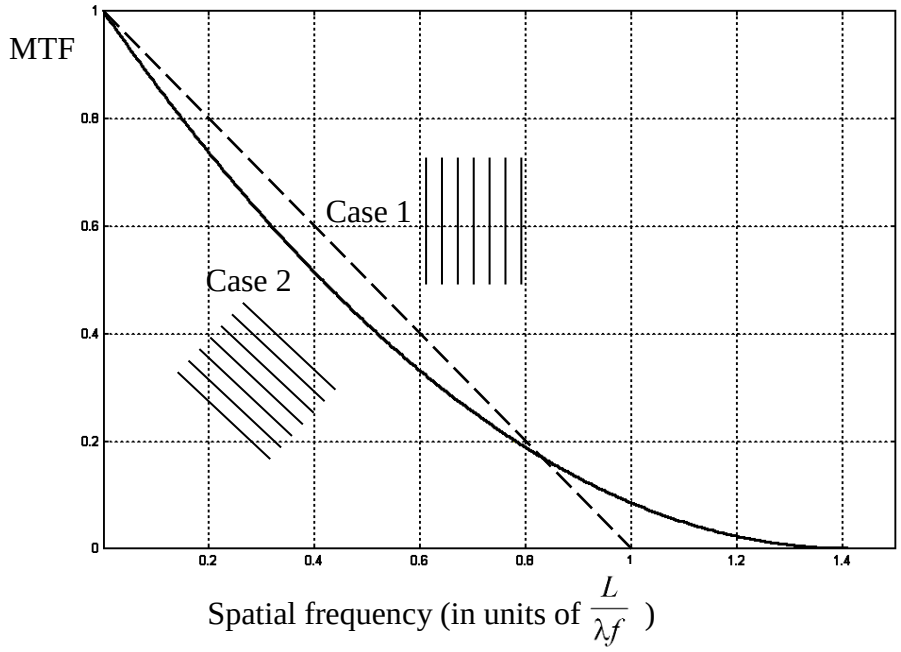
Solution: Two lens apertures, each scaled by a factor of $1/\lambda f$, are drawn. One is displaced ν_x in the horizontal direction and ν_y in the vertical direction relative to the other aperture. The area of overlap of the two apertures, divided by the area of one whole aperture, is the MTF value for a pattern with spatial frequency components ν_x and ν_y. The result is:

$$MTF(\nu_x,\nu_y) = 1 - \frac{\lambda f}{L}(\nu_x + \nu_y) + \left(\frac{\lambda f}{L}\right)^2 \nu_x \nu_y$$

The *MTF* for the square, diffraction-limited lens is shown below for positive ν_x and ν_y. The vertical axis represents the *MTF* value, and the two horizontal axes represent the spatial frequency components ν_x and ν_y, expressed in units of $L/\lambda f$. As stated previously, the degree of modulation for a pattern with spatial frequency components ν_x and ν_y will be $MTF(\nu_x,\nu_y)$ times lower in the image compared with the object.

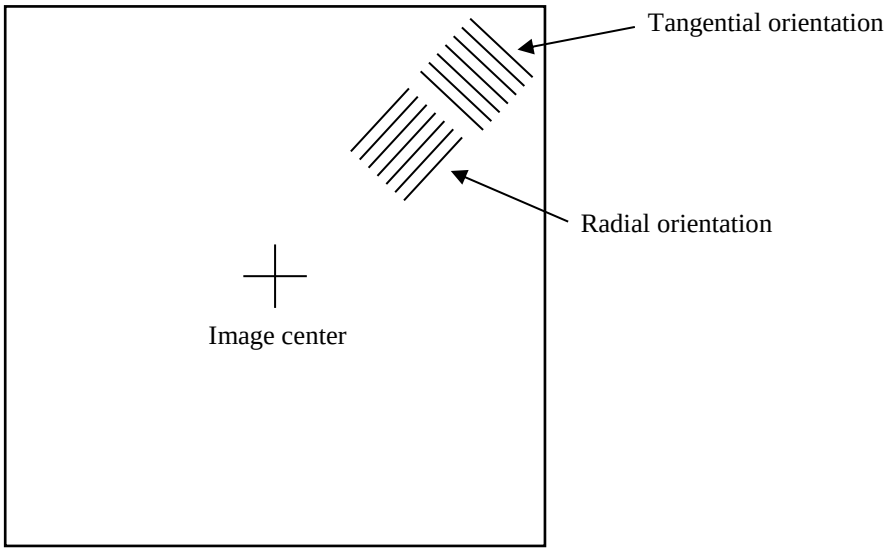


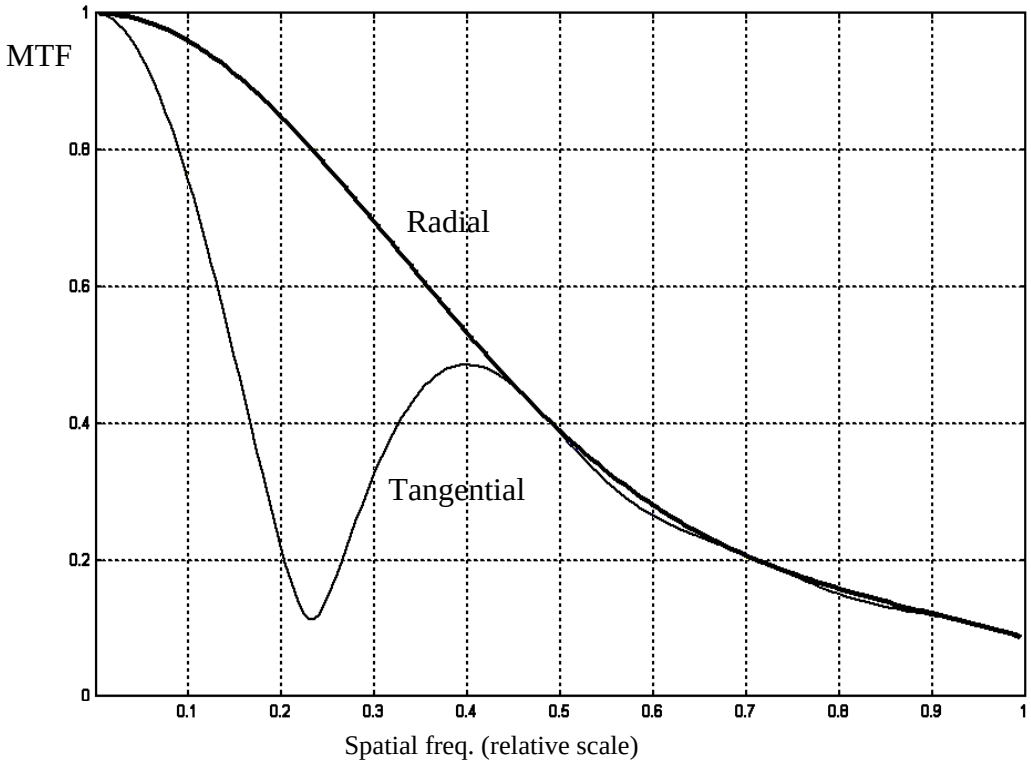
Even though images are 2-dimensional, and the 2D *OTF* is usually needed to fully characterize the imaging properties, it is not so common that 2D *OTF*s or *MTF*s are presented. There are two reasons for this. First, it would be very time-consuming to measure the full 2D *MTF* for an optical system (*PTF*s are seldom measured at all), especially since one usually wants to know the *MTF* for several positions in the image field (the *psf* varies over the image field due to aberrations). Second, it is often difficult to visualize and interpret a 2D *MTF* displayed as a surface plot, and furthermore quantitative comparison is difficult. Therefore, one often displays two 1D *MTF* plots for different pattern orientations instead of the full 2D *MTF*. For example, in the case with square aperture on the previous page, we can display *MTF* curves for the cases (compare page 32): $\nu_y = 0$, $\nu_x = \nu$ (Case 1) and $\nu_x = \nu_y = \frac{\nu}{\sqrt{2}}$ (Case 2). We get the following results (for clarity, the pattern orientations are displayed)



Looking at the figure above, it is much easier to quantitatively see the differences in *MTF* for the two pattern orientations than looking at the 2D plot on previous page.

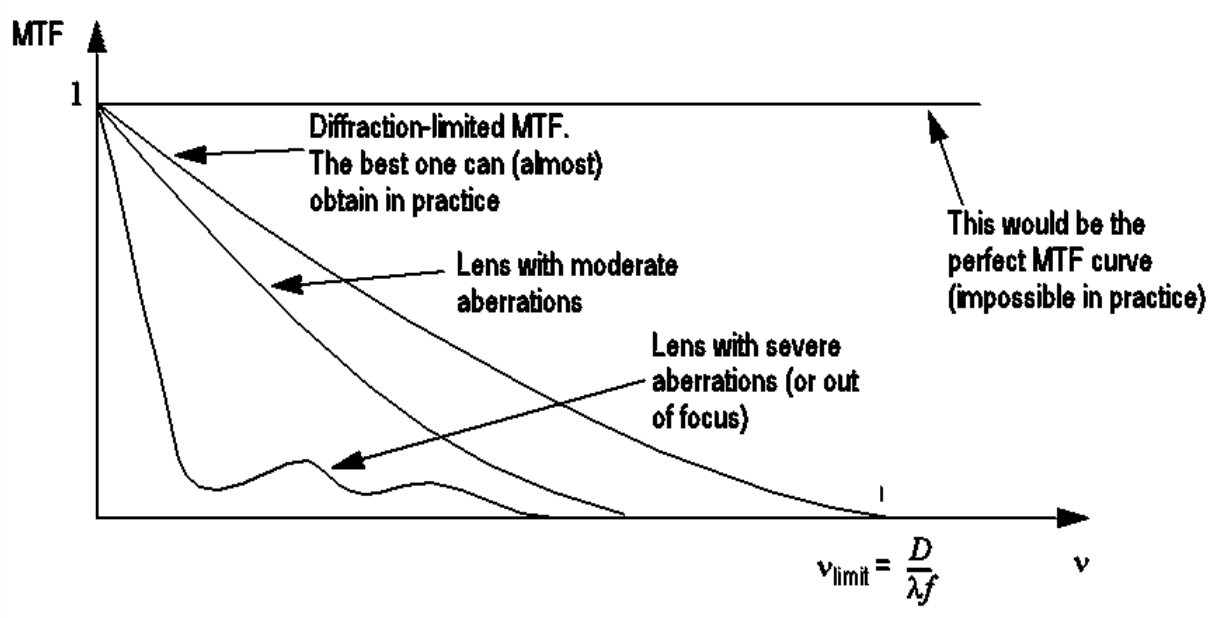
When optical systems are tested, it is common to present three *MTF* curves. One shows the performance in the image center. The other two curves show the performance some distance away from the image center, and for two perpendicular line pattern orientations. For the image center one curve is sufficient because, as previously said, here the *psf* is rotationally symmetric and therefore performance is independent of pattern orientation. Off-center the *psf* is usually not rotationally symmetric; often it is more elongated in the radial direction than in the tangential direction. Therefore *MTF* curves for both radial and tangential line patterns are often displayed. In the illustration on next page these pattern orientations are illustrated. Also illustrated on next page are the radial and tangential *MTF* curves for the "coma simulation" case on page 34. Note how much easier it is to compare quantitatively the two curves on next page compared with extracting the same information from the surface plot on page 34.





It should be noted that there is some confusion about the terms “radial” and “tangential” in connection with off-axis *MTF*. Sometimes these terms refer to the orientation of the lines (as in the figures above), whereas in other cases they refer to the “direction vector” mentioned on page 32.

As we have seen in both the one- and two-dimensional *MTF* curves presented so far, the general tendency is that higher spatial frequencies mean lower *MTF* values. This continues up to the limiting frequency, above which the *MTF* value is zero. Such is the behavior for diffraction-limited lenses. For real lenses with aberrations, the *MTF* values are lower than for a diffraction-limited lens, and it may well happen (especially when the aberrations are large) that the curve displays oscillations, see illustrations on previous page and this page. The perfect *MTF* curve would, of course, be one that has a value of unity for all spatial frequencies. This would produce an image that preserves all the object details with full contrast up to the highest spatial frequencies, but, alas, no real lens can live up to this, because it would require a *psf* that is a δ function, and this is impossible due to diffraction.



14. On Two-Dimensional Fourier Transforms

The transition from one- to two-dimensional Fourier transforms may seem trivial from a mathematical point of view; one needs simply to expand all the equations by one more dimension. However, this may call for an explanation and some discussion on the physical interpretation of the results.

Definition
A one-dimensional Fourier transform of $f(x)$ is given by:

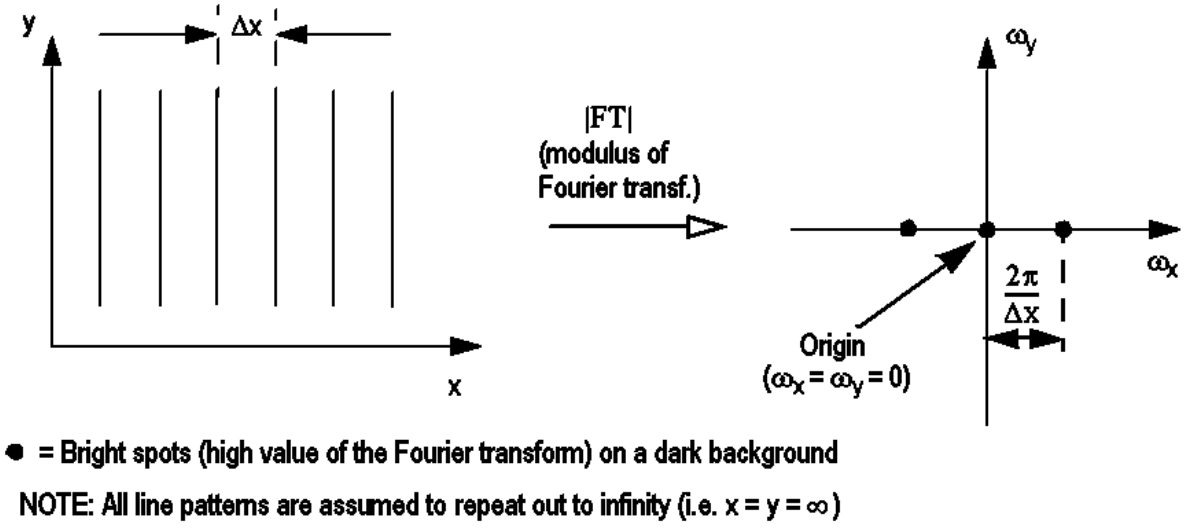
$$\hat{f}(\omega) = \int_{-\infty}^{+\infty} f(x) \cdot e^{-i\omega x} dx$$

A two-dimensional Fourier transform of f(x,y) is given by:

$$\hat{f}(\omega_x, \omega_y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) \cdot e^{-i(\omega_x x + \omega_y y)} dxdy$$

The function $f(x,y)$ may represent the distribution of light in an optical image, for example. How should $\hat{f}(\omega_x,\omega_y)$ be physically interpreted, and how can it be presented (we cannot simply draw a curve)? We will start with the last question, and restrict ourselves to studying the modulus of the Fourier transform $\left|\hat{f}(\omega_x,\omega_y)\right|$, for the sake of simplicity.

As $\left|\hat{f}(\omega_x,\omega_y)\right|$ is a function of two variables, it is usually presented as an image where ω_x is the horizontal image coordinate, ω_y the vertical coordinate, and the value of the function is represented on a grayscale (a high value being a light area, and a low value a dark area). Below are some examples of simple patterns and their Fourier transforms. The lines shown in the original images (on the left) represent peaks in a sinusoidal intensity pattern (cf. page 31)



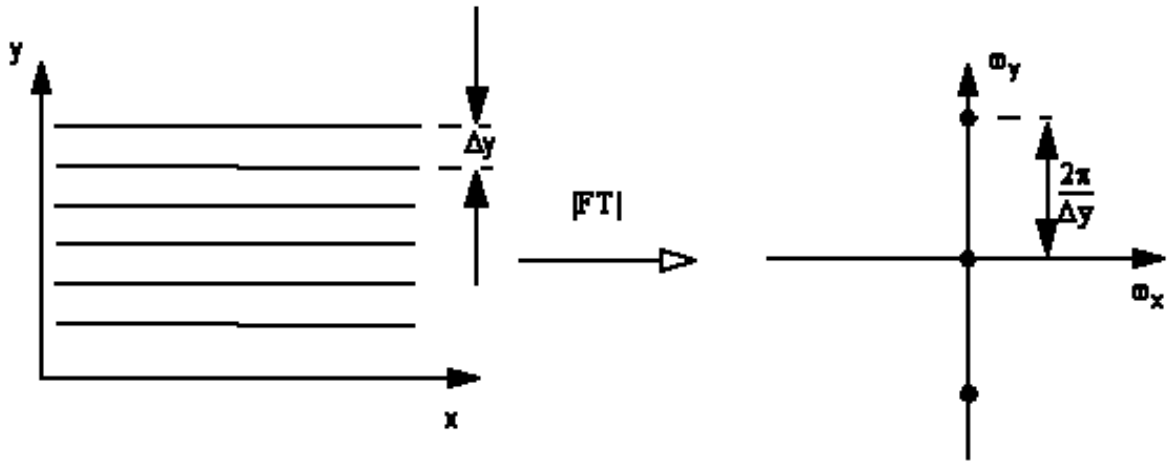
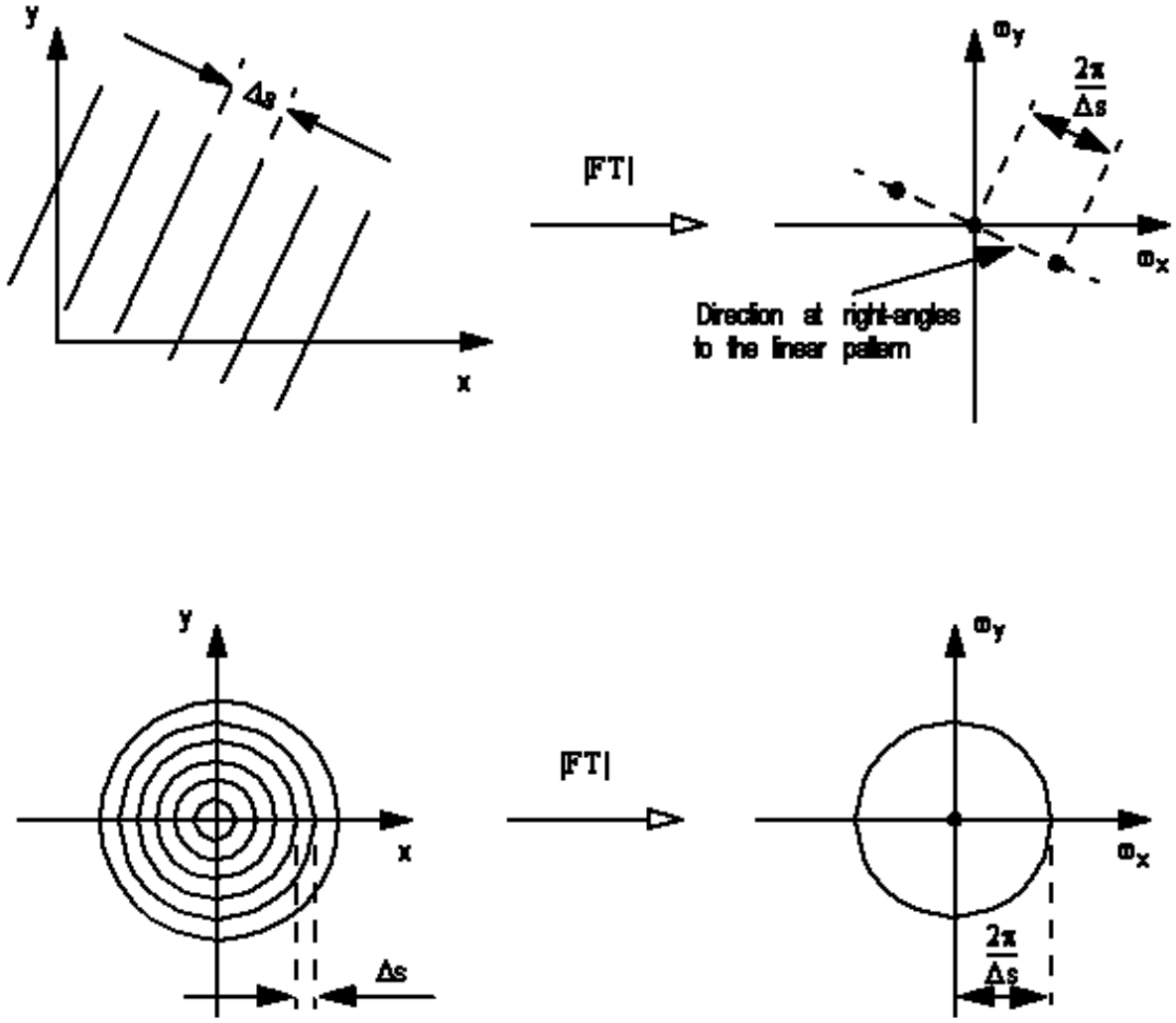


Figure continued on next page!



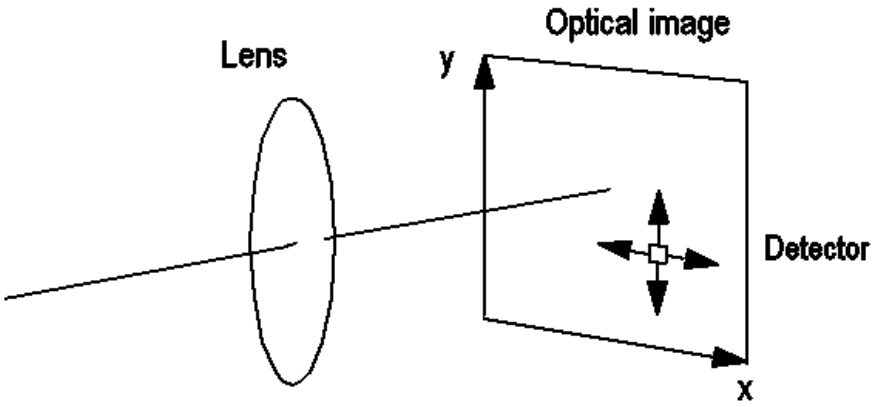
The bright spot at the center of the Fourier transforms in the figures represents a spatial frequency of zero, and provides information on the mean intensity of the original image.

The physical interpretation of the two-dimensional Fourier transform, is that it is possible to re-create the original image by summing a large number (in the general case an infinite number) of 2-dimensional harmonic oscillation patterns with different spatial frequencies, directions, amplitudes and phase angles. All this information is contained in the two-dimensional Fourier transform (although in the figures above, we have not illustrated the phase information).

15. The *OTF* of the Detector

Apart from the optics, there are other factors which limit the imaging quality. One such factor is the detector. In order for photons to impinge on the detector, it must have a certain area.

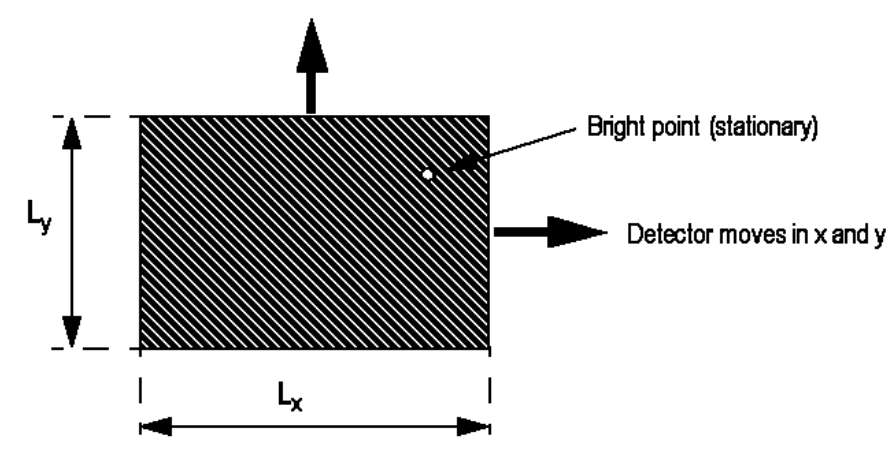
Let us assume that we have a single rectangular detector with the dimensions L_x and L_y (in a linear sensor or an area array sensor, each detector element is often only a few microns in size). We further assume that this detector is used to scan an optical image, i.e. to record the light intensity level (or more accurately, illuminance level) as a function of x and y.



The recorded image function $I_R(x,y)$ will differ from the actual intensity distribution in the optical image $I_B(x,y)$. This is because the area of the detector blurs the details of the original image. Mathematically, this can be expressed in the same way as for optical imaging, namely through convolution:

$$I_R = I_B \otimes psf_{\text{detector}}$$

where psf_{detector} is the recorded image function we would have obtained if the optical image had been a perfect point (δ function). We can obtain $psf_{\text{detector}}(x,y)$ by measuring the output signal as a function of the x and y coordinates when moving the detector relative to an infinitely small optical point image (δ function), see figure below.



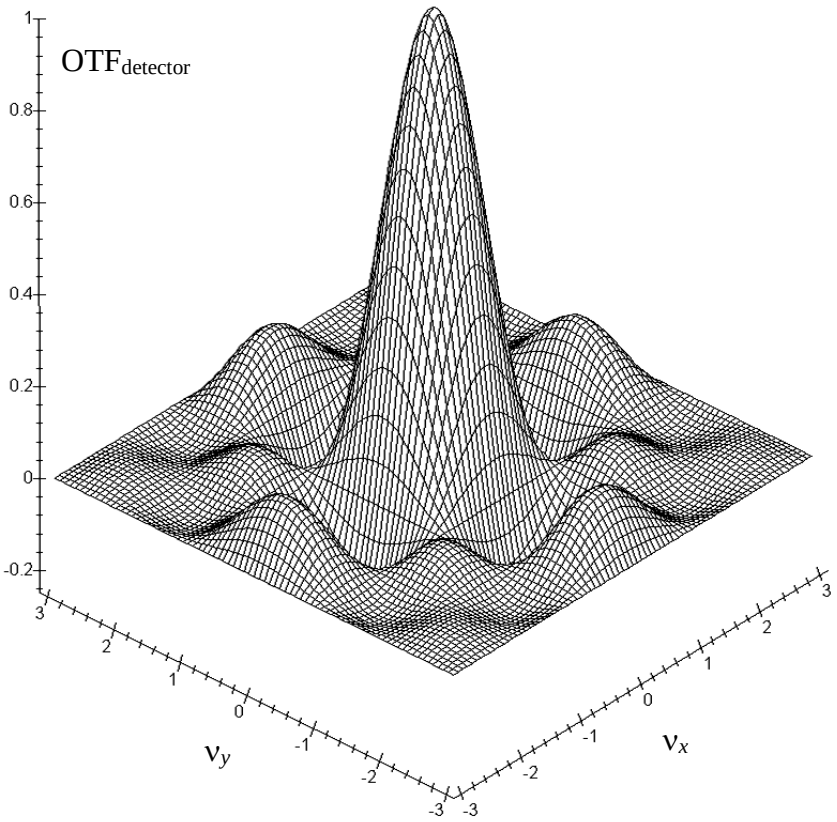
If the sensitivity of the detector is constant over the whole area, the output signal (i.e. psf_{detector}) can be written:

$$psf_{\text{detector}}(x,y) = rect\left(\frac{x}{L_x},\frac{y}{L_y}\right)$$

where *rect* is a "rectangular" function. This function is equal to 1 if both condition $-\frac{L_x}{2} \le x \le \frac{L_x}{2}$ and condition $-\frac{L_y}{2} \le y \le \frac{L_y}{2}$ are fulfilled simultaneously, and zero otherwise.
We can now calculate

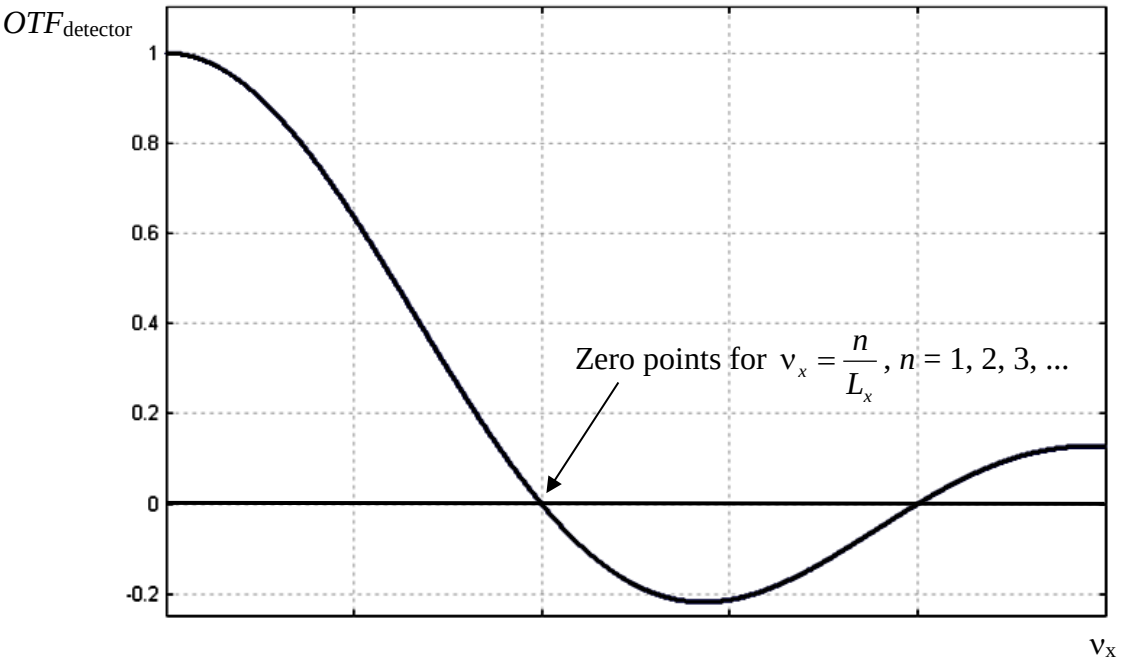
$$OTF_{\text{detector}} = \hat{psf}_{\text{detector}} = FT\left[rect\left(\frac{x}{L_x},\frac{y}{L_y}\right)\right] = \frac{\sin(\pi\nu_x L_x)}{\pi\nu_x L_x}\cdot\frac{\sin(\pi\nu_y L_y)}{\pi\nu_y L_y}$$

This function is illustrated in the figure below. The spatial frequencies ν_x and ν_y are given in units of $1/L_x$ and $1/L_y$ respectively.

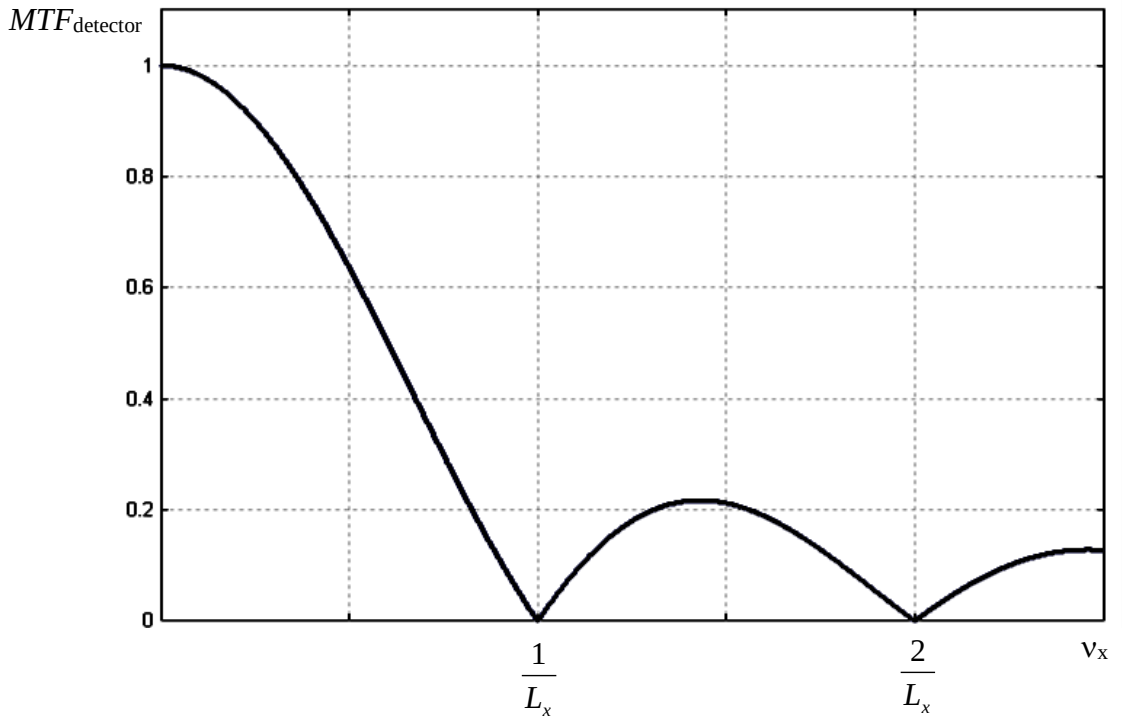


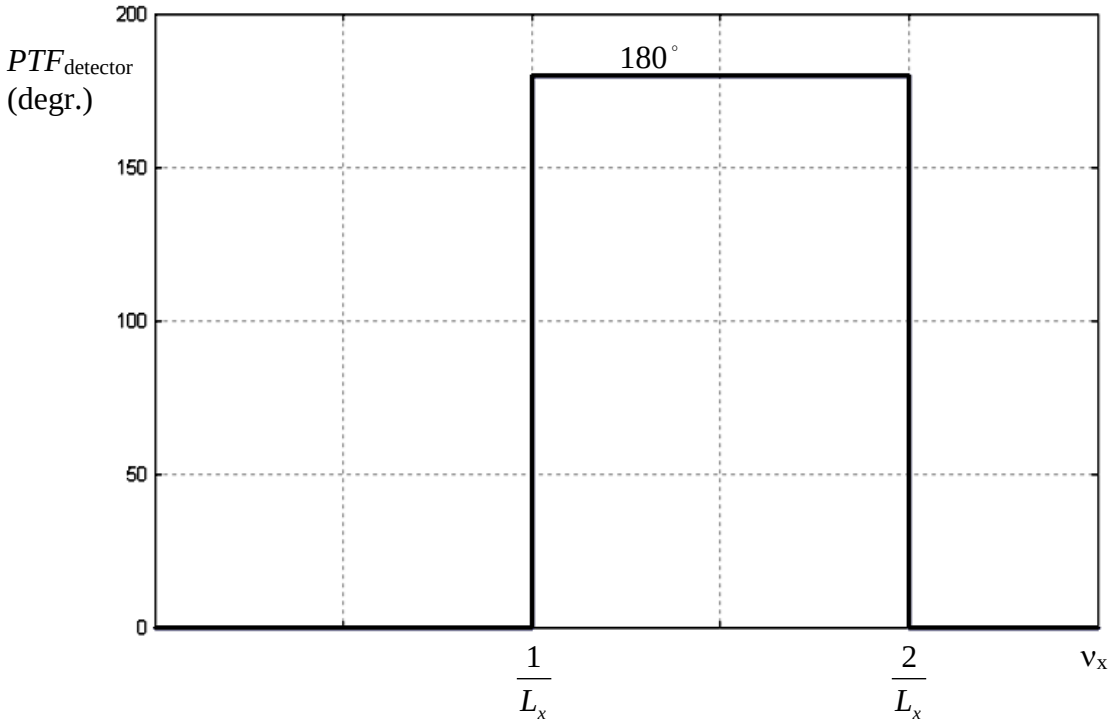
If we consider a pattern in the x-direction only (i.e. $\nu_y = 0$) we obtain:

$OTF_{\text{detector}}(\nu_x) = \frac{\sin(\pi\nu_x L_x)}{\pi\nu_x L_x}$. This function is illustrated on next page.

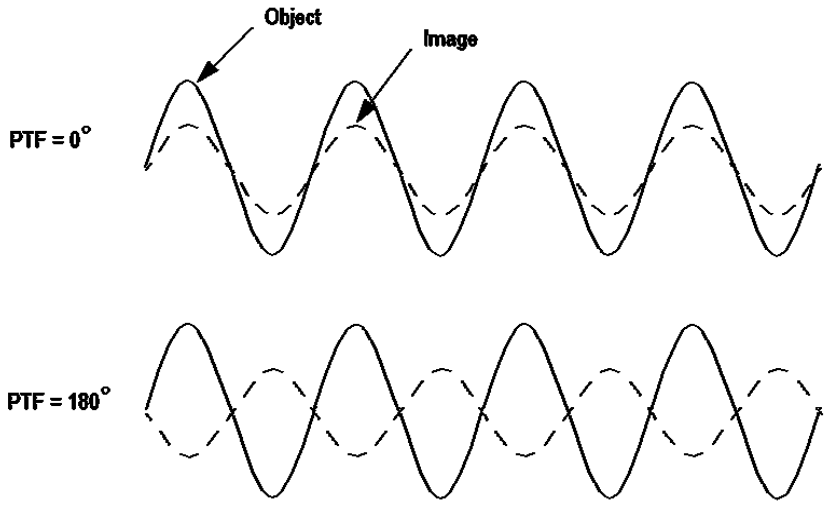


If we separate the *OTF* into *MTF* and *PTF*, we obtain the following figures:

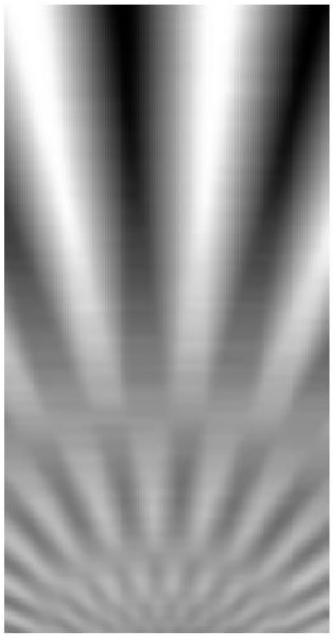




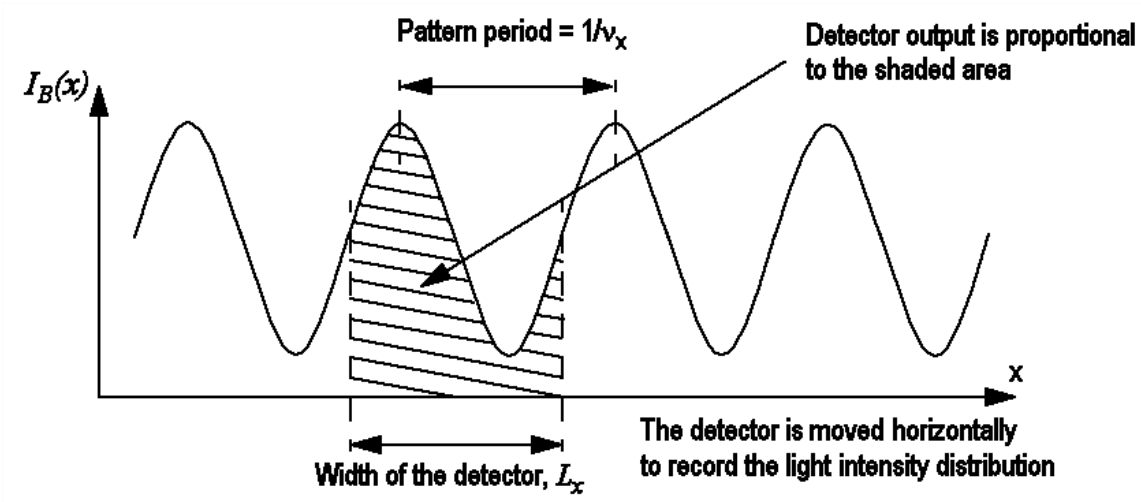
From these *MTF* and *PTF* curves we get some interesting information concerning the influence of the detector on the imaging quality. From the *MTF* curve we see that the general trend is the same as for the optics, namely that the *MTF* values tend to be lower for higher spatial frequencies. Compared with the optics (page 30) we see two differences. First, MTF_{detector} doesn't have a limiting frequency. The detector can, in principle, detect arbitrarily high spatial frequencies, although the modulation will be very low. Also, because of the zero points of the *MTF* curve some frequencies will be lost. Second, MTF_{detector} displays an oscillatory behavior which is not seen in the optics (at least if it is high-quality and properly focused). PTF_{detector} displays an even stranger behavior. When OTF_{detector} changes sign, this will produce an abrupt phase shift of 180° in PTF_{detector}. How will this phase shift affect the imaging properties? Recalling what was said on page 25, we realize that a phase shift of 0° means that when imaging a sine-wave, the object and image functions will be in phase. A phase shift of 180°, on the other hand, means that the image and object functions are out of phase by half a wavelength as illustrated below.



Phase shifts of this type may introduce clearly visible imaging artifacts. This is especially true in cases where the spatial frequency of a line pattern continuously changes over the image, as in the left figure below where the spatial frequency increases towards the bottom. In the right figure we see the effects of $OTF_{detector}$ (convolution with a rectangular function has been performed in the horizontal direction). In addition to a general blurring, the pattern changes repeatedly between positive and negative contrast as one move from top to bottom. This is a result of the phase jumps of the *PTF*.



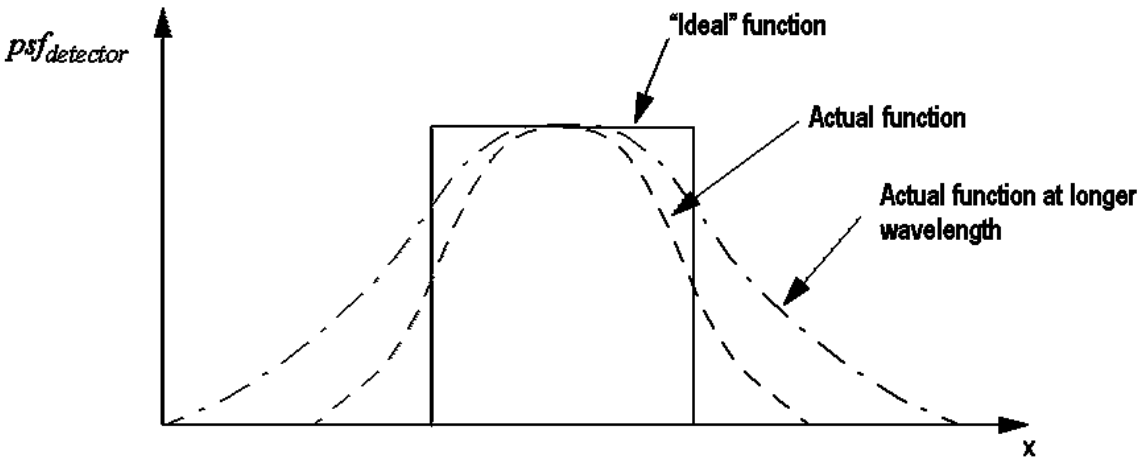
One may now ask why $OTF_{detector}$ displays a number of equally spaced zero points. This is easy to understand if we consider a case where we measure a sinusoidally varying light intensity using a detector width L_x as shown in the illustration below:



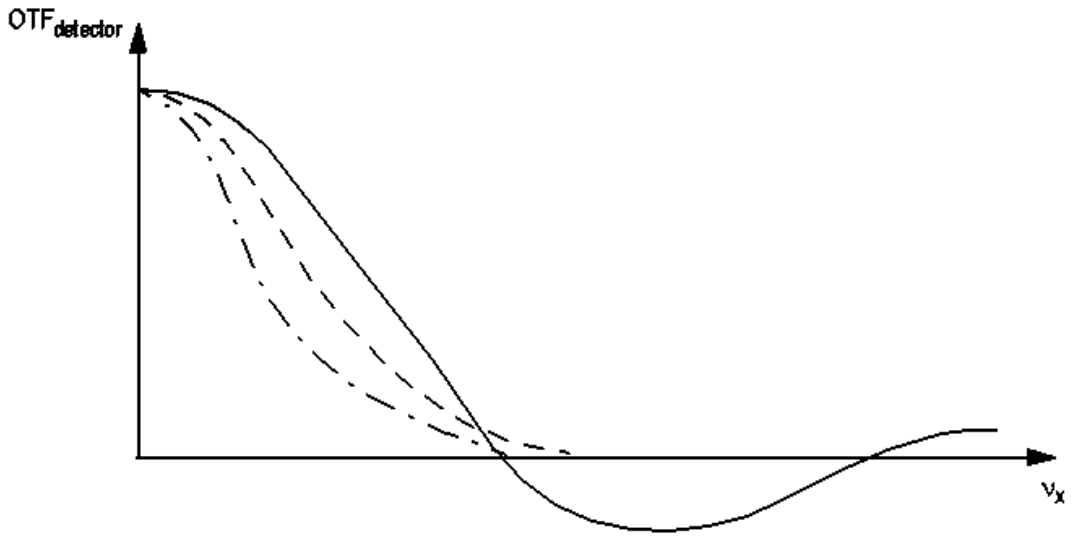
If the width of the detector is equal to the period of the pattern, i.e. if $L_x = \frac{1}{\nu_x}$, the output signal (the mean light intensity over the detector area) will be the same regardless of how the detector

is moved in the horizontal direction. Such a constant output means that the modulation is zero, and therefore $OTF_{detector} = 0$. The same result will be obtained if the width of the detector is equal to several whole periods of the pattern.

In many cases the boundary between the light sensitive and insensitive regions of a detector is not as abrupt as we have assumed so far. This is the case for semiconductor detectors used in linear and area array sensors. Furthermore, the psf may change with wavelength, the trend being that a longer wavelength will produce a broader *psf*, see illustration below. In order to prevent long wavelengths like infrared radiation from deteriorating the image quality, these wavelengths are often removed with an absorption filter.



Changes in $psf_{detector}$ will, of course, affect $OTF_{detector}$. A broader and more rounded function, $psf_{detector}$, means that $OTF_{detector}$ decreases more rapidly at increased spatial frequency, and that it shows only weak or no oscillations at high frequencies. This is shown schematically below, where the three *OTF* curves correspond to the three *psf* curves in the figure above.



In simple terms the relationship between *psf* and *OTF* can be stated as follows:

- The broader the *psf*, the narrower the *OTF* curve (i.e. it will drop more quickly as the spatial frequency increases).
- The more rounded the *psf*, the less oscillations (if any) we will see in the *OTF*.
- A symmetric *psf* means that the *OTF* is real (non-complex). Therefore the only possible values for the *PTF* are 0° or 180° (cf. pages 43 & 44).

A larger detector area means that more light enters the detector, which gives a higher signal level and improved SNR. On the other hand, a large area implies less detail in the recorded image. If the measurement time is not limited, then both good photometry and a high degree of detail can be obtained.

16. The *OTF* for the Whole Imaging Process

We have seen that the optical imaging process can be described by OTF_{optics}, and light detection by $OTF_{detector}$. We now wish to see the total effect of both the optics and the detector. We have the following relationships:

$$I_R = I_B \otimes psf_{\text{detector}} \text{ and } I_B = I_O \otimes psf_{optics}$$

Combining these two expressions we get:

$$I_R = \left(I_O \otimes psf_{optics}\right) \otimes psf_{\text{detector}}$$

The Fourier transform of this expression is:

$$\hat{I}_R = FT\left\{I_O \otimes psf_{optics}\right\} \cdot OTF_{\text{detector}} = \hat{I}_O \cdot OTF_{optics} \cdot OTF_{\text{detector}}$$

We can see that the product $OTF_{optics} \cdot OTF_{\text{detector}}$ is equivalent to the total transfer function from the object to the recorded image. This is the important **multiplication rule** for the *OTF*, which can be extended to cover other factors affecting the image quality, e.g. vibrations. In other words:

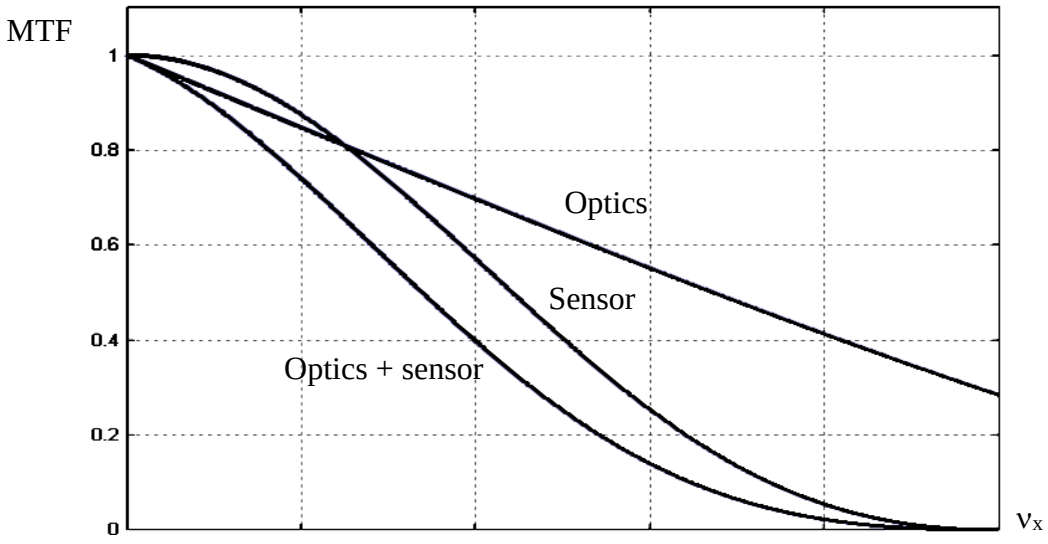
$$OTF_{total} = OTF_{optics} \cdot OTF_{\text{detector}} \cdot OTF_{vibrations} \cdot \ \ldots \ \text{etc.}$$

Separating the modulus and the phase, the multiplication rules gives:

$$MTF_{total} = MTF_{optics} \cdot MTF_{\text{detector}} \cdot MTF_{vibrations} \cdot \ \ldots \ \text{etc.}$$

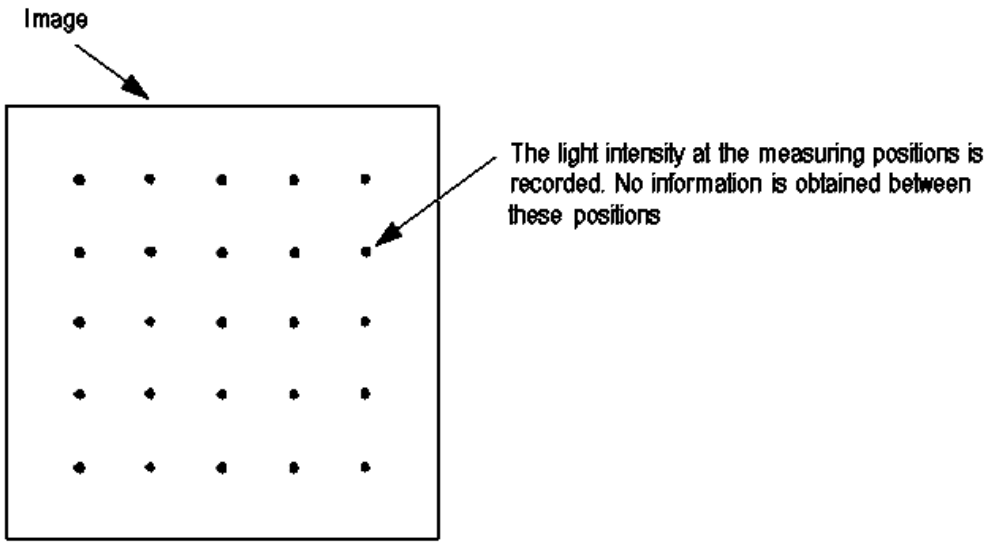
$$PTF_{total} = PTF_{optics} + PTF_{\text{detector}} + PTF_{vibrations} + \ \ldots \ \text{etc.}$$

In the illustration below the *MTF*s for optics and sensor are combined to produce the total *MTF*.



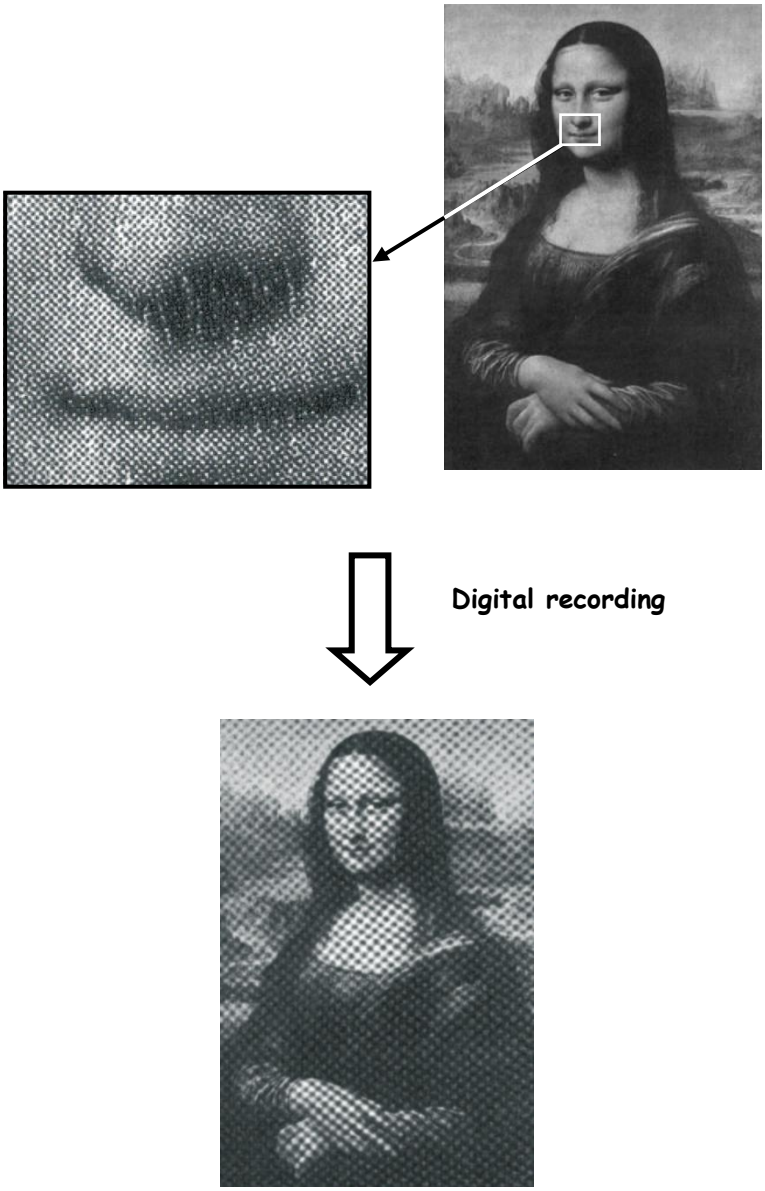
17. Sampling

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.

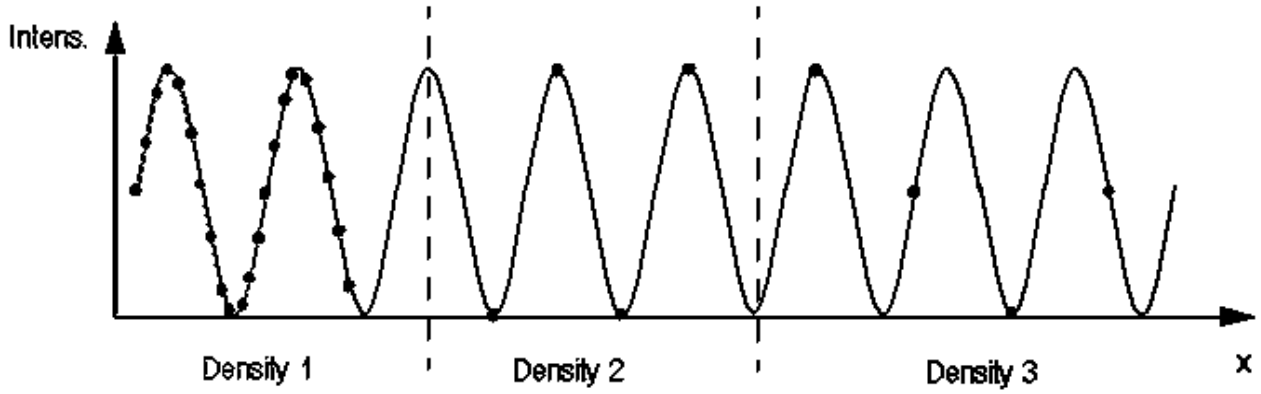


The measurement values, which are normally stored digitally, are called pixels (from "picture cells"). Depending on the application, digital images may consist of anything from just a few pixels up to many million pixels. For comparison it can be mentioned that an ordinary television image consists of 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).

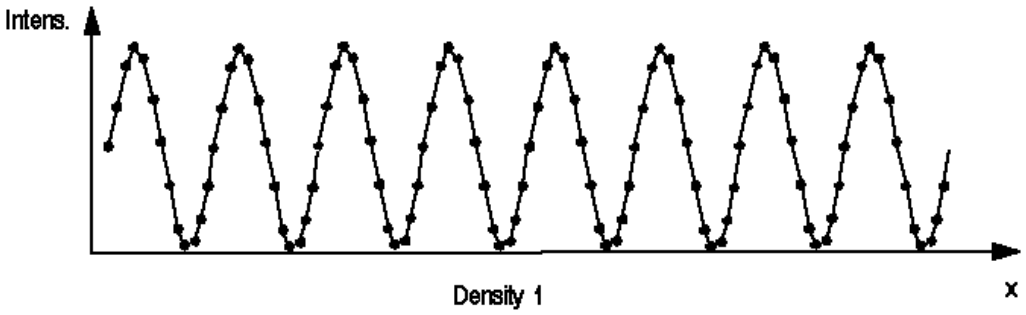
Not only do we get less information if the sampling points are spaced far apart, we may also get imaging artifacts introducing false information in the recorded images. A commonly seen example of this is when printed raster images are scanned in a document scanner, see illustration below. The coarse 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.

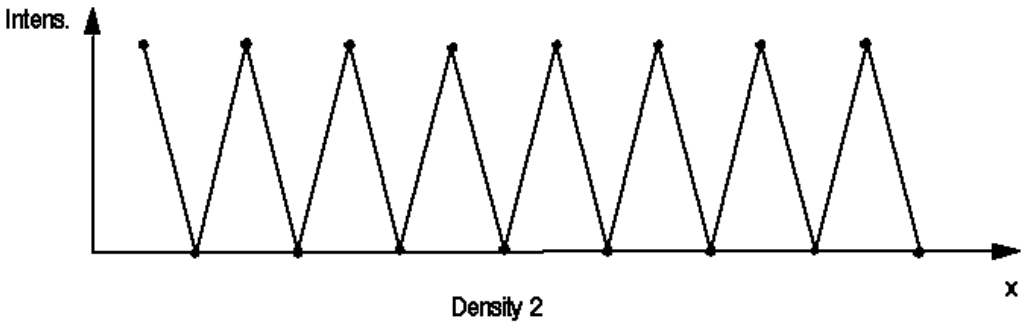


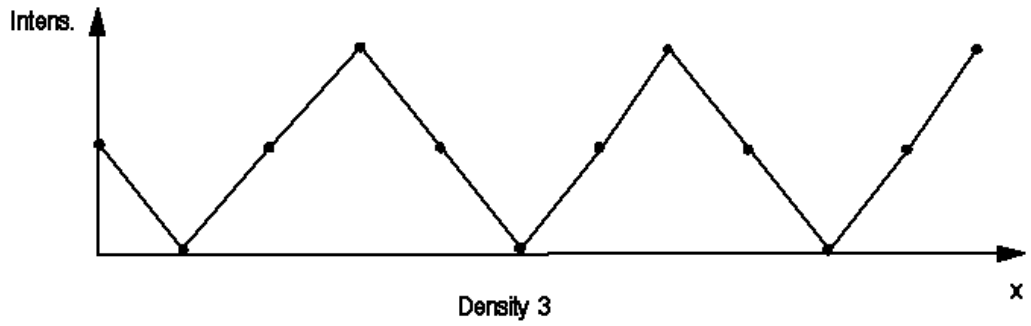
Let us now study sampling in more detail, and express it quantitatively. We will also see how sampling and the *MTF* together affect the image quality. Assume that we have a sinusoidal variation in intensity in the x-direction of the image. Let us consider the results of various sampling densities.



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.)

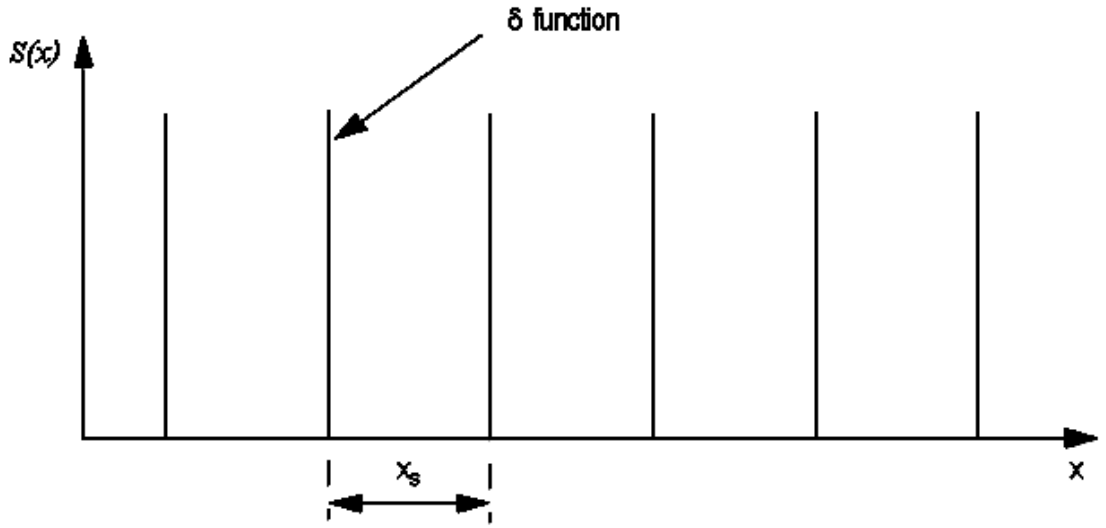
We will now describe the sampling procedure mathematically, and describe how the original function can be reconstructed from the sampled data. For the sake of simplicity, we will restrict ourselves to the consideration of the one-dimensional case, but the definition can easily be generalized for several dimensions.

Assume that we have a recorded image function $I_R(x)$ which we want to sample, and to reconstruct. The question of whether it is possible to recreate the original function **exactly** from the sampled data is of special interest and, if that is the case, what is required. We will here disregard the noise in the signal which in reality, of course, sets a practical limit on how well a function can be recreated.

Let us now assume that we are sampling $I_R(x)$ with a distance x_s between the sampling points (sampling frequency, $\nu_s = \frac{1}{x_s}$). The following sampling function can be defined:

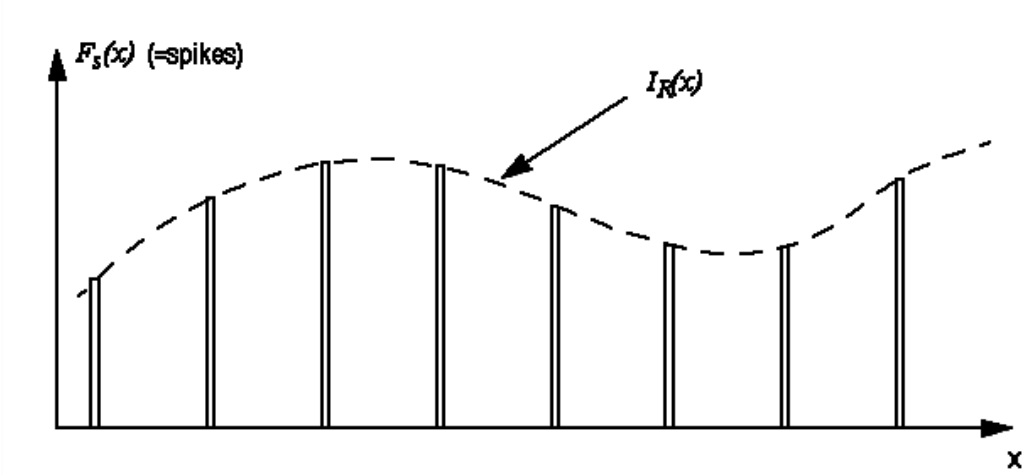
$$S(x) = \sum_{n=-\infty}^{+\infty} \delta(x - nx_s)$$

This function is schematically illustrated on next page.



The sampling process can be represented by the product $I_R(x) \cdot S(x) = F_S(x)$. $F_S(x)$ is called the **sampled function**, and it is this function which is stored, for example in digital form in a computer.

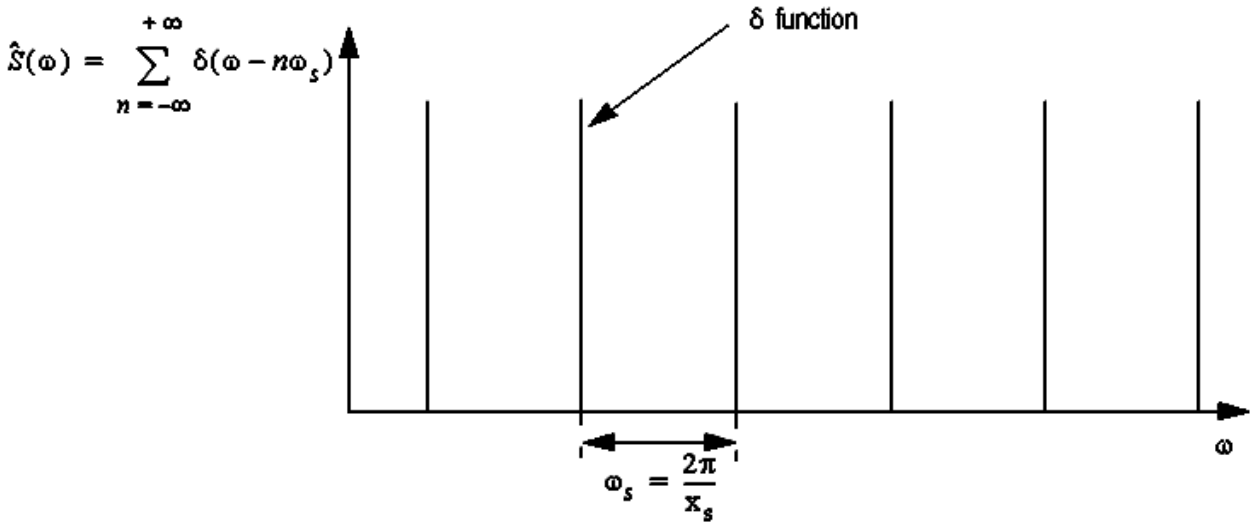
If one reflects on the function $F_S(x)$ as it is described above, it is not physically meaningful and cannot be realized as it consists of a number of infinitesimally narrow and infinitely tall "spikes". In reality, one does not work with δ functions, but with narrow (although not infinitesimally narrow) spikes of finite height, see illustration below. The difference between these two functions can be shown to be unimportant as long as the width of the spikes is considerably less than the smallest structure in $I_R(x)$.



Let us now perform the Fourier transform of the sampled function $F_S(x)$, i.e. consider its spectrum:

$$\hat{F}_S(\omega) = \hat{I}_R(\omega) \otimes \hat{S}(\omega)$$

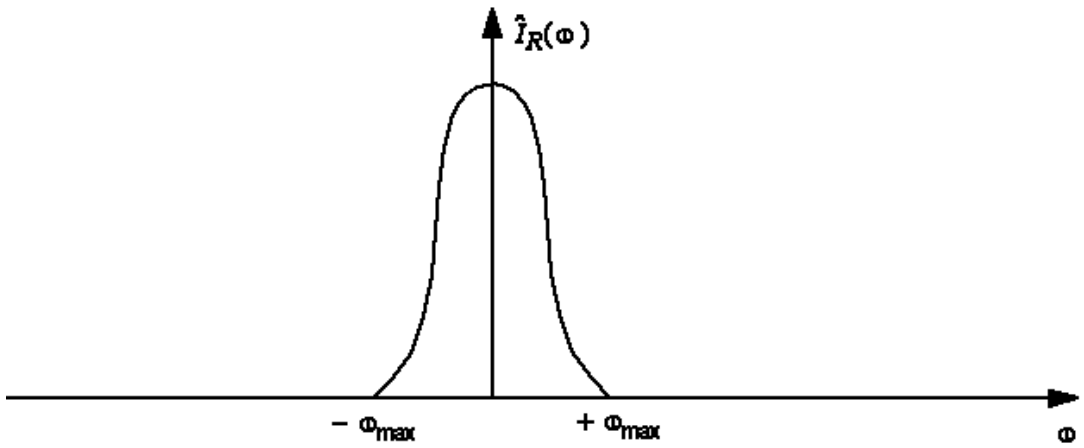
$\hat{S}(\omega)$ is an infinite sum of δ functions, according to the figure below (this can be found in tables of Fourier transforms).



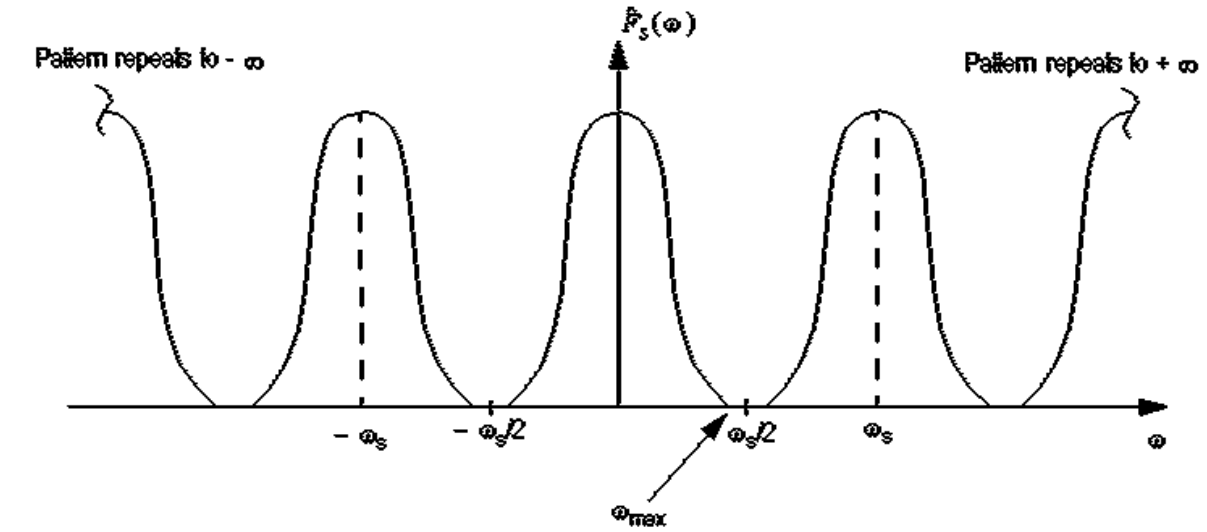
This gives: $\hat{F}_S(\omega) = \int_{-\infty}^{+\infty} \hat{I}_R(u) \cdot \hat{S}(\omega - u) du =$

$$= \int_{-\infty}^{+\infty} \hat{I}_R(u) \cdot \left(\sum_{n=-\infty}^{+\infty} \delta(\omega - n\omega_s - u) \right) du = \sum_{n=-\infty}^{+\infty} \hat{I}_R(\omega - n\omega_s)$$

The Fourier transform of the sampled function $F_s(x)$ is thus infinitely many copies of the Fourier transform of the original function $I_R(x)$. Assume that $\hat{I}_R(\omega)$ has the following appearance:



We assume that the maximum spatial frequency in the original function is $\nu_{max} = \frac{\omega_{max}}{2\pi}$. The Fourier transform of the sampled function, $\hat{F}_s(\omega)$ will then have the following form:



If we multiply $\hat{F}_s(\omega)$ by $rect\left(\frac{\omega}{\omega_s}\right)$ (a function with the value 1 for $-\frac{\omega_s}{2} < \omega < \frac{\omega_s}{2}$ and 0 otherwise), we obtain a function which is identical to $\hat{I}_R(\omega)$, i.e. we mask the frequency-transposed copies of $\hat{I}_R(\omega)$, giving:

$$\hat{I}_R(\omega) = \hat{F}_s(\omega) \cdot rect\left(\frac{\omega}{\omega_s}\right)$$

The Fourier transform of a function contains **all** the information about the function. Thus, the inverse Fourier transform of $\hat{F}_s(\omega) \cdot rect\left(\frac{\omega}{\omega_s}\right)$ will exactly recreate the original function $I_R(x)$

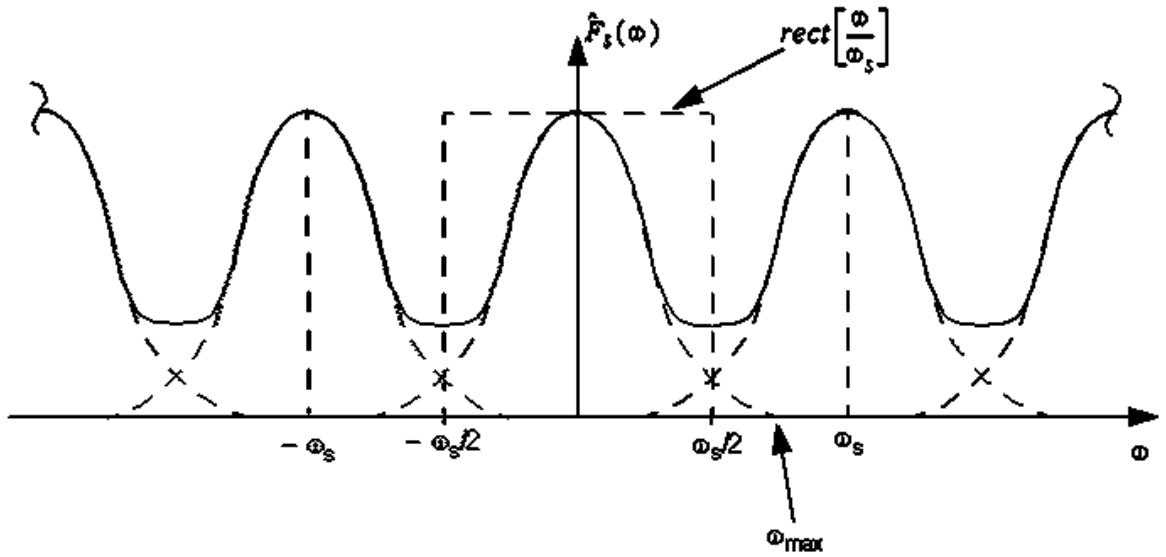
.

It is, however, not necessary to use Fourier transforms to reconstruct the original function from the sampled data. It is possible to instead use the sampled function $F_s(x)$ directly. To see what is needed, let us perform the inverse Fourier transform of the above expression.

$$I_R(x) = FT^{-1}\left\{\hat{I}_R(\omega)\right\} = FT^{-1}\left\{\hat{F}_S(\omega) \cdot rect\left[\frac{\omega}{\omega_S}\right]\right\} = F_S(x) \otimes \left\{\frac{\sin(\omega_S x/2)}{\omega_S x/2}\right\}$$

We have found that the original function can be reconstructed by convolving the sampled function with $\frac{\sin(\omega_S x/2)}{\omega_S x/2}$.

We have now shown that, regardless of whether one is working in Fourier space or in normal space, it is possible (at least in theory) to exactly recreate the original function from the sampled data. There is, however, one important criterion for this. We have assumed that $\omega_{max} < \frac{\omega_S}{2}$ (see figure on previous page). What happens if $\omega_{max} > \frac{\omega_S}{2}$? The Fourier transform $\hat{F}_S(\omega)$ then has the following appearance:



The effect is that angular frequencies over $\frac{\omega_S}{2}$ will also be found under $\frac{\omega_S}{2}$ as the spectra from the different orders overlap. When multiplying by $rect\left[\frac{\omega}{\omega_S}\right]$, the angular frequencies over $\frac{\omega_S}{2}$ will be given as lower frequencies (in the case above, as $\omega_s - \omega$). This phenomenon is called **aliasing**, and we have already seen it illustrated in the figures on pages 49 & 50.

It can be shown that, generally, an angular frequency over the Nyquist limit, $\frac{\omega_S}{2}$, is reproduced as a frequency ω_{alias} which fulfils the following requirements:

$\omega_{alias} = \left| n\omega_S - \omega \right|$, n = 1, 2, 3 ... , and $\omega_{alias} \leq \frac{\omega_S}{2}$

In color area array sensors, like the one shown on page 7, the sampling frequency is different for the green elements and the red/blue ones. Consequently, the Nyquist frequencies will also be different, as well as the aliasing patterns when exceeding these frequencies. We therefore typically get colored fringes in such cases.

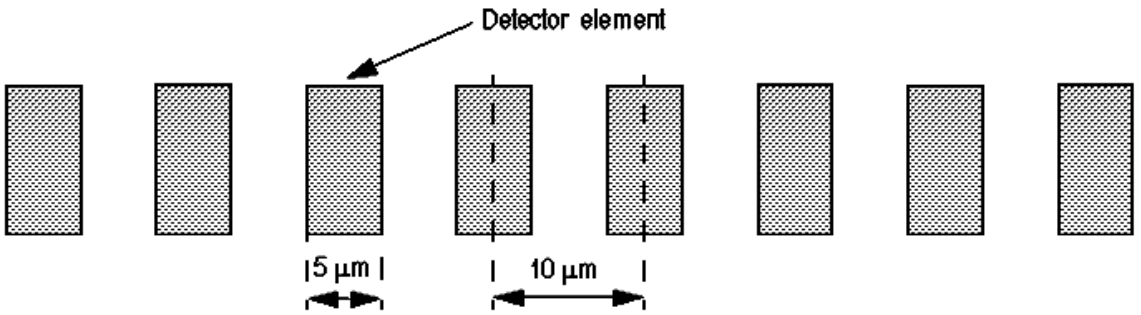
We have now shown mathematically that the statement made on page 51 was correct:

In order to be able to correctly recreate an image with spatial frequencies up to ν_{max}, sampling must be carried out at a frequency of at least $2\nu_{max}$ (that is, two sampling points per period). This is called the sampling theorem. (The theorem was presented by C.E. Shannon in a 1949 publication concerning information theory, not imaging specifically. Similar results had also been published by others during the 1920s and 1930s)

In practice, it is easier to use a somewhat higher sampling frequency than that given by the sampling theorem. This is why, for example, a sampling frequency of 44 kHz is used for music CDs, despite the fact that the highest frequency recorded on a CD is approximately 20 kHz.

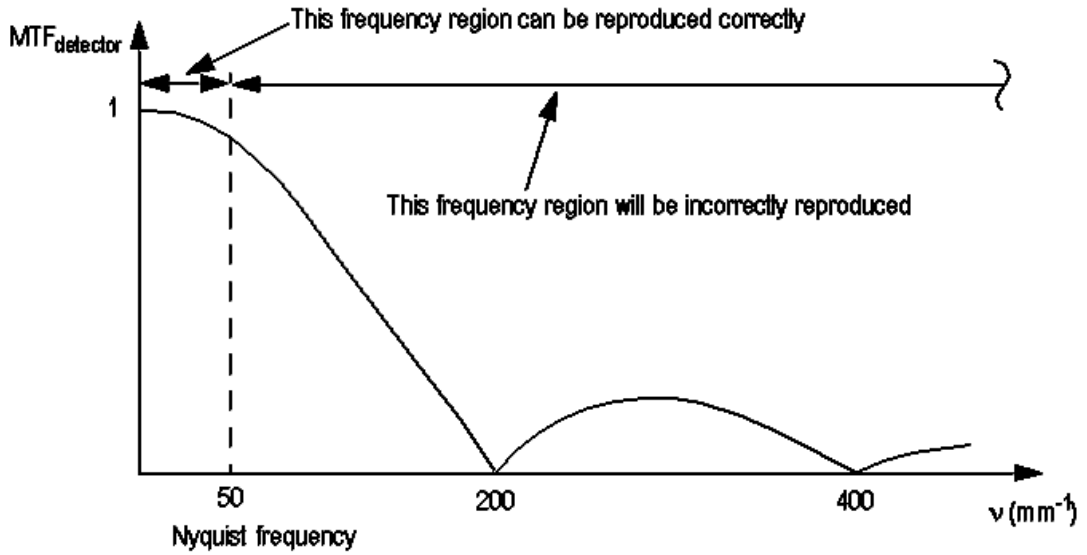
NOTE: Unlike the detector size, sampling does not lead to any loss of contrast in small details in the image. On the other hand, these details will be distorted if the sampling theorem is not fulfilled. This is a tricky problem which can be difficult (or impossible) to detect. The problem is especially difficult with linear and area array sensors where the sampling density is built into the detector, and is far too low.

Example: Assume that we are using a linear sensor with a 5 μm window width, and the center-to-center distance between the windows is 10 μm.



The sampling frequency will be $\nu_S = \frac{1}{10\times10^{-6}}$ m^{-1} = 100 mm^{-1} which gives a Nyquist frequency of $\nu_{max} = \frac{\nu_S}{2} = 50$ mm^{-1}. Let us now consider the *MTF*. Sampling an image function with the linear sensor will give exactly the same result as sampling with a single detector of 5 μm width that is moved horizontally 10 μm between signal read-out. This means that the *MTF* of the linear sensor will be that of a single detector element in the sensor. If we assume a uniform light sensitivity within each detector element, we get the same result as on page 42:

$$MTF_{\text{detector}}(\nu) = \left|\frac{\sin(\pi\nu\cdot5\times10^{-6}}{\pi\nu\cdot5\times10^{-6}}\right|$$

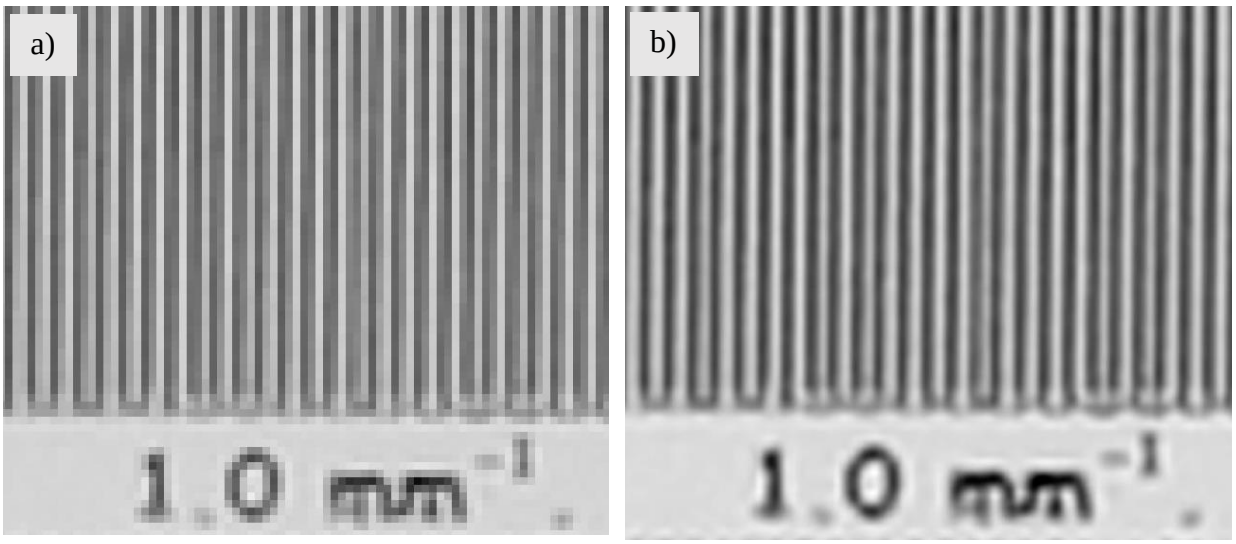


Thus, the detector can reproduce frequencies above the Nyquist frequency with good contrast. These will, however, be reproduced as lower frequencies, i.e. we get ***false*** information. This can be a serious problem in many applications, and one should always keep this in mind when using linear and area array sensors.

Aliasing that has occurred in the sensor cannot be compensated for by subsequent signal processing (information has been irretrievably lost). Therefore some manufacturers introduce controlled optical blurring, using so-called anti-aliasing filters (Appendix 5), so that no (or very little) information above the Nyquist frequency is present in the optical image. However, such methods also degrade image contrast at frequencies well below the Nyquist frequency, giving the images an overall blurred appearance. Since this is undesirable, one sometimes prefers to live with the artifacts produced by aliasing. In general, the aliasing artifacts are not so disturbing for most types of object, and therefore they can often be accepted. An elegant way to virtually eliminate the aliasing problem, is to employ micro-movements of the sensor. Let's take the linear sensor above as an example. Imagine that we make two recordings of the same object, and that the sensor (or optical image) is moved horizontally a distance of 5 μm between the recordings. If we combine data from the two recordings we get only half the normal sampling distance. Therefore the Nyquist frequency is doubled without affecting the *MTF* of the system. By making four recordings with a spacing of 2.5 μm, we can quadruple the Nyquist frequency etc. In this way we can get an arbitrarily high Nyquist frequency. This technique can also be extended to two dimensions. Disadvantages with this technique are that it complicates the imaging system, increases cost, and requires more or less stationary objects. Therefore it has not found widespread use.

Aliasing does not occur only in the sensor, it can also occur in the display screen. Such screens consist of many groups of red, green and blue (RGB) dots. The dot pitch (center-to-center distance between RGB groups) is about 0.25 mm on a computer screen, and three times as large on a TV screen. The dot pitch determines the highest spatial frequency that can be displayed on the screen. If one exceeds this limit aliasing will occur, i.e. a lower frequency will be displayed. This has to be considered by manufacturers of display screens when deciding on the dot pitch and the number of display lines.

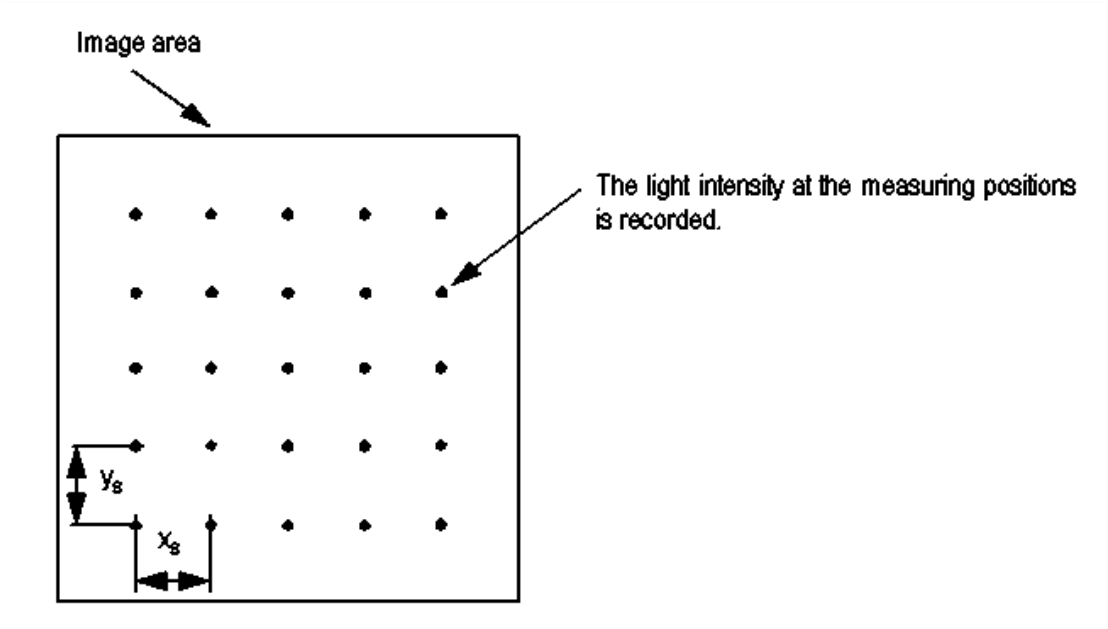
A question that often pops up in connection with digital images is the following: Why does my digital image differ markedly from the object even when the sampling theorem is fulfilled? Such differences are often noticed for regular patterns like stripes, where the pixel values in the image may vary in a seemingly erratic way (although the periodicity of the pattern, measured over many periods, is correct). The reason why the image differs from the object, although all necessary information has been collected, is that no ***reconstruction*** from the sampled data has been made (p. 54). What is displayed on the computer screen is the ***sampled*** function. All the information is there, but we can't see it properly without reconstruction. But in order to view the reconstructed image, we would need to view an analog (not pixellated) image, which of course is impossible on a computer screen. To view something resembling an analog image, we would need a screen with many more pixels than in the sampled image. Therefore, contrary to the case for music-CD, reconstruction from sampled data is usually not done for digital images. In cases where only a smaller part of the image is of interest, this smaller part can be reconstructed and displayed with many more pixels. An example of this is shown in the figure below. On the left we have the original image, representing the sampled function of a vertical line pattern. On the right we see an image where, using interpolation, the original number of pixels has been increased by a factor of six in both the *x* and *y* dimensions. The interpolation method used, bicubic interpolation, is an approximation of the theoretically correct reconstruction method for sampled data (convolution with a sinc-function, see page 54). Note that no new information is obtained through this reconstruction, but the information present in the sampled image is displayed in a better way.



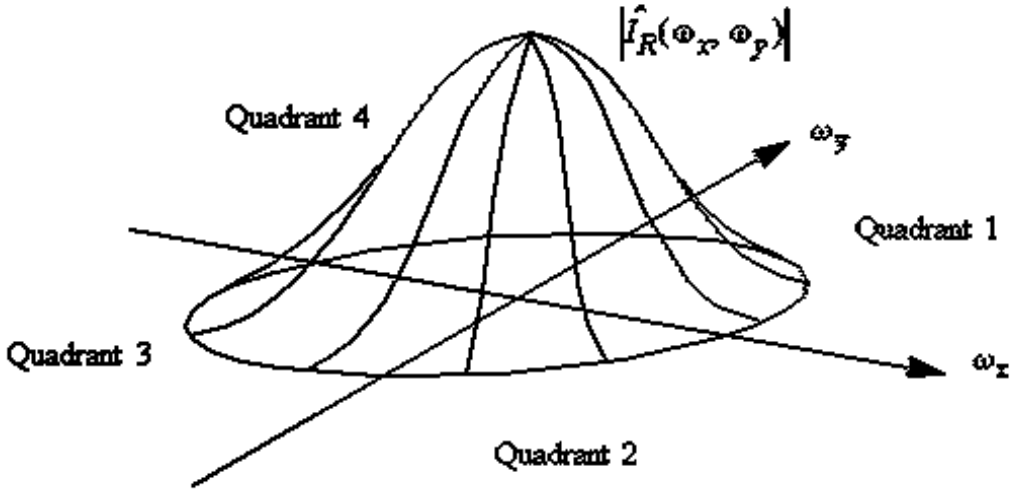
18. Sampling in two dimensions

In the previous chapter a rather comprehensive overview of the topic of sampling was given. For the sake of simplicity only one-dimensional signals were treated. Even with this limitation the results are quite useful for treating several realistic cases, such as a time-varying signal (e.g. audio CD) or measurements carried out along a line in an image (e.g. x- or y-profiles). However, for a full treatment of image sampling it is, of course, necessary to study the general case of two-dimensional sampling. It is rather straightforward to do so by simply extending the treatment in the previous chapter to two-dimensional functions. Basically, most of the results obtained in this way are simple and intuitive, and therefore the mathematical treatment of two-dimensional sampling is omitted in this chapter. Instead the results will be described, and practical applications of the results will be given.

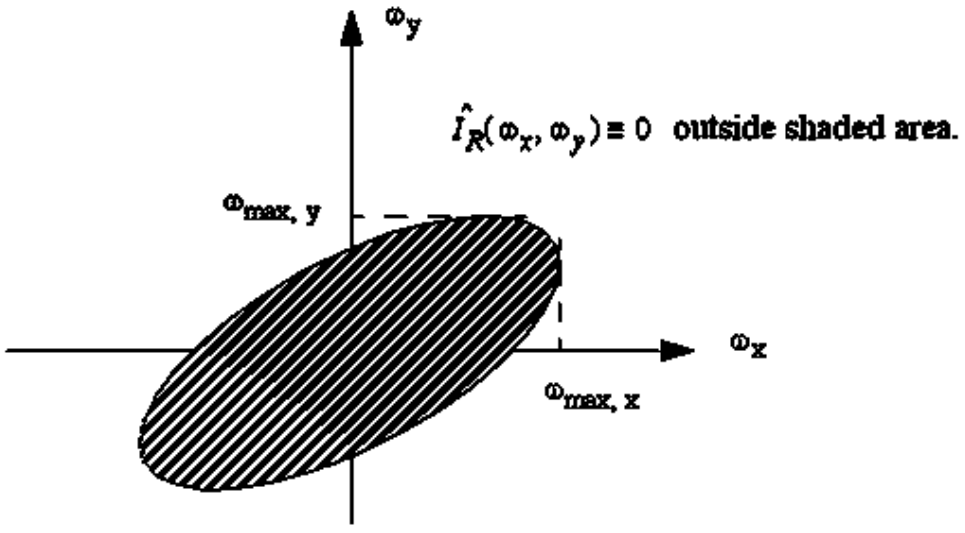
Let us start by looking at a two-dimensional "sampling grid" used for image sampling. Although the sampling points can, in principle, be randomly scattered over the image area, they are usually organized in a periodic pattern with, say, sampling distances x_s and y_s in the x and y directions respectively (often $x_s = y_s$, but this is not necessary as we shall later see).



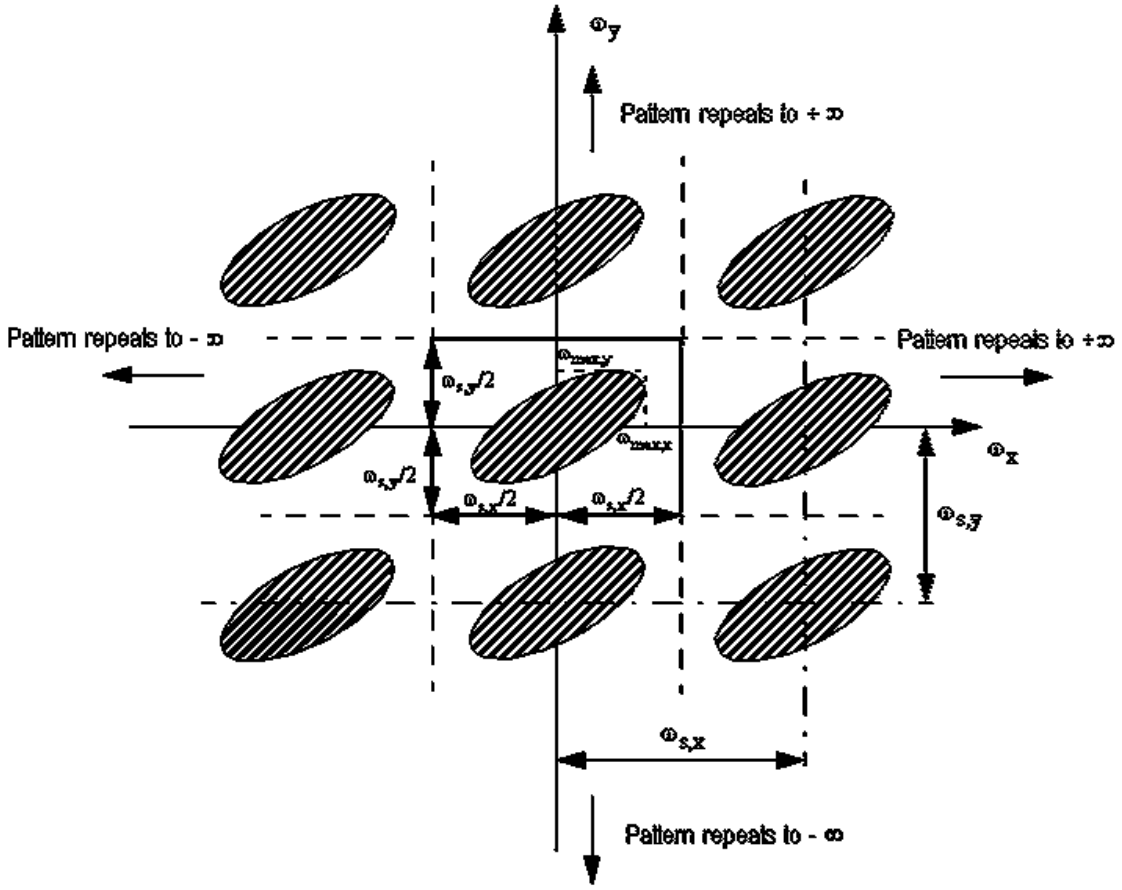
Let us assume that the continuous, recorded image function (i.e. before sampling) is described by $I_R(x,y)$. This function, which corresponds to $I_R(x)$ in the previous chapter, describes the output signal that would be obtained if we could smoothly move (i.e. not in a step-wise fashion) one detector over the entire image area and record the output signal as a function of the (x,y) position of the detector. As described in previous chapters, $I_R(x,y)$, will depend on the luminance distribution of the object being imaged, as well as on the optical transfer functions of both the optics and the detector (and possibly other factors). For the sake of illustration, let us assume that the Fourier transform of $I_R(x,y)$, denoted by $\hat{I}_R(\omega_x,\omega_y)$, has the following appearance:



Since $\hat{I}_R(\omega_x, \omega_y)$ is usually a complex function, the illustration shows only the absolute value, or modulus, $\left|\hat{I}_R(\omega_x, \omega_y)\right|$. If $I_R(x, y)$ is a real function (as is always the case in this course), it follows from the properties of the Fourier transform that $\hat{I}_R(-\omega_x, -\omega_y) = \hat{I}_R^*(\omega_x, \omega_y)$, where * denotes the complex conjugate. This is the reason why $\left|\hat{I}_R(\omega_x, \omega_y)\right|$ displays symmetry in the ω_x, ω_y plane. Thus quadrant 3 is just a mirror image of quadrant 1, and quadrant 4 is a mirror image of quadrant 2. In analogy to the one-dimensional case, we assume that $I_R(x, y)$ contains spatial frequencies only up to certain limits, $\nu_{max,\,x}$ and $\nu_{max,\,y}$, in the x- and y-directions respectively (this is natural since optical systems can only reproduce spatial frequencies up to a certain limit as described in chapters 12 and 13). The corresponding maximum angular frequencies, denoted by $\omega_{max,\,x} = 2\pi\nu_{max,\,x}$ and $\omega_{max,\,y} = 2\pi\nu_{max,\,y}$, are illustrated in the figure below, showing a shaded area in the ω_x, ω_y plane where $\hat{I}_R(\omega_x, \omega_y) \neq 0$. (The area is illustrated as an ellipse for simplicity. In reality it may have any shape consistent with the symmetry conditions for the quadrants described above).



Let us now assume that the continuous function $I_R(x,y)$, characterized above, is sampled with a two-dimensional "sampling grid" with spacings of x_s and y_s. Doing the same mathematical treatment of the sampling process as in the one-dimensional case, but with two-dimensional functions, it is found that the Fourier transform of the sampled function $F_S(x,y)$ is given by an infinite number of copies of the Fourier transform of $I_R(x,y)$, with center-to-center spacings of $\omega_{S,x}=\frac{2\pi}{x_S}$ and $\omega_{S,y}=\frac{2\pi}{y_S}$ in the ω_x and ω_y directions. The illustration below shows the areas in the ω_x,ω_y plane where $\hat{F}_S(\omega_x,\omega_y)\neq 0$.



In analogy to the one-dimensional case, we multiply $\hat{F}_S(\omega_x,\omega_y)$ by a two-dimensional rectangular function, $rect\left(\frac{\omega_x}{\omega_{S,x}},\frac{\omega_y}{\omega_{S,y}}\right)$ (see chapter 15 for a description of this function). By doing so, we retain information only in the central part of the ω_x,ω_y plane, indicated by the solid rectangular box in the figure above. Thus we get rid of all the copies caused by the sampling process, and are left with just the Fourier transform of the continuous function $I_R(x,y)$. By taking the inverse Fourier transform, we can then (in principle) exactly re-create $I_R(x,y)$. Apart from the fact that we are now working in two dimensions, this is exactly the same procedure as we carried out in the one-dimensional case.

It is now straightforward to postulate the conditions that must be fulfilled to avoid aliasing in the two-dimensional sampling case. It is the information in the central rectangular box in the ω_x, ω_y plane that is utilized. Therefore the contents of this box must represent the Fourier transform of $I_R(x, y)$. This will be the case only if none of the "copies" extends into the box. This condition will be fulfilled if $\omega_{\max,x} < \frac{\omega_{S,x}}{2}$ and $\omega_{\max,y} < \frac{\omega_{S,y}}{2}$ are satisfied simultaneously. Or, using spatial frequencies rather than angular frequencies and rearranging, we get:

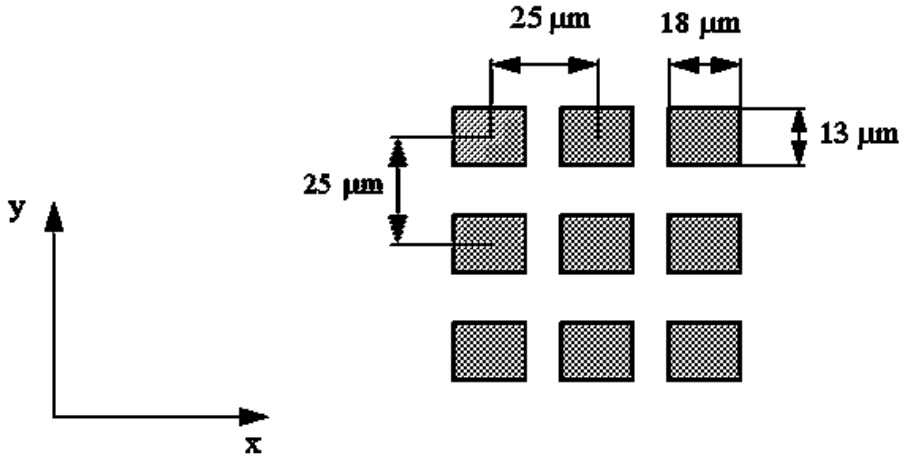
****

$\nu_{s,x} > 2\nu_{max,x}$ and $\nu_{s,y} > 2\nu_{max,y}$

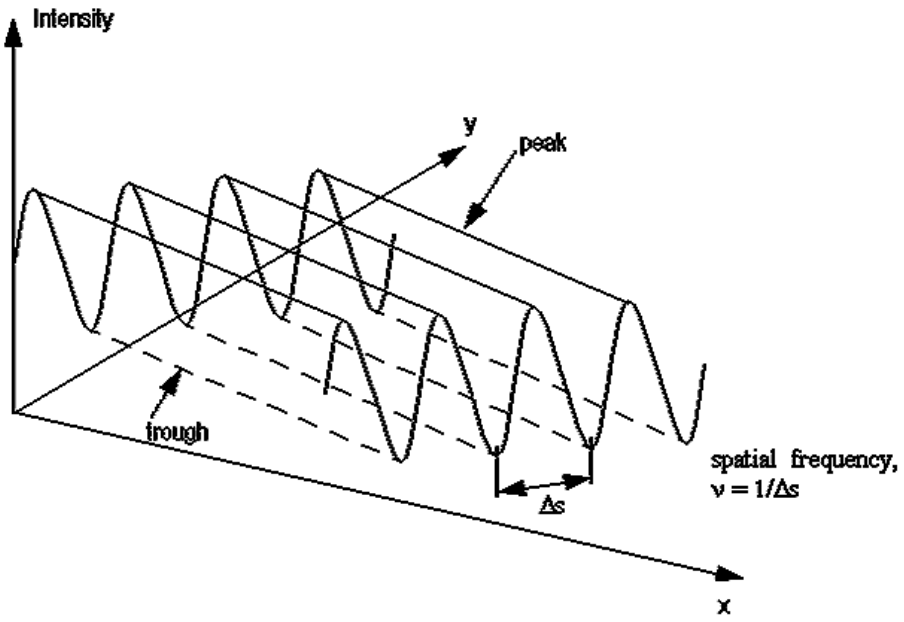
****

This is the sampling theorem in two dimensions. It is a simple and intuitive extension of the one-dimensional case. For the practical implications of this theorem, let us look at a simple example.

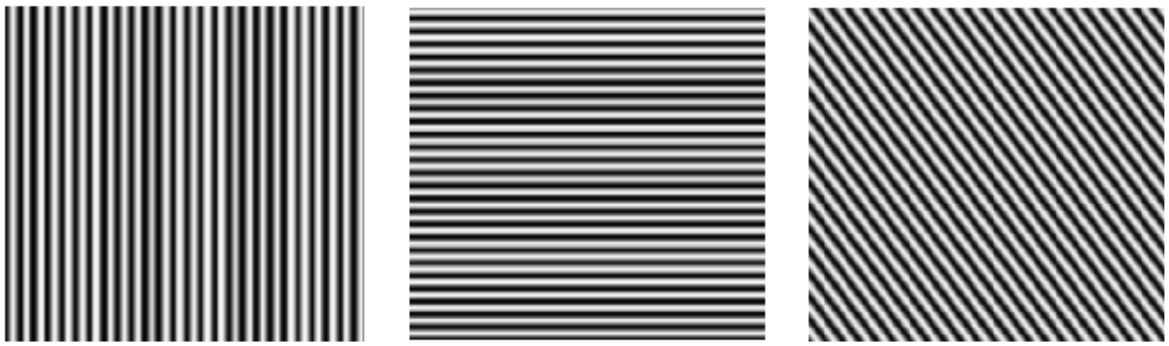
Example: An area array sensor consists of 1024 x 1024 detector elements, each with a size of 13 μm x 18 μm. The center-to-center distance between the elements is 25 μm both in the x and y directions.



A sinusoidal illumination pattern, illustrated in the figure below, is projected onto this sensor.



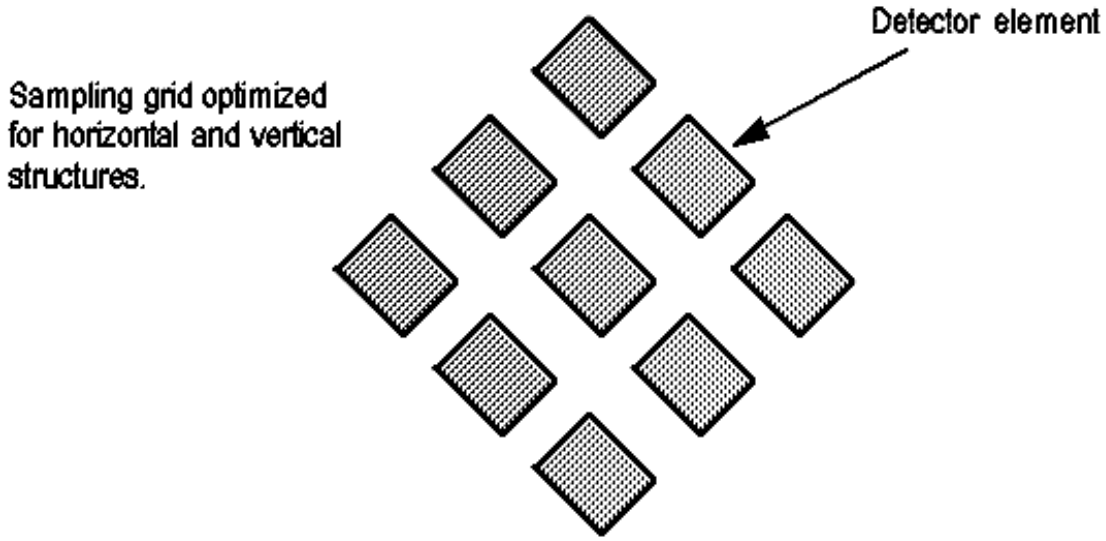
The orientation of the pattern can be varied arbitrarily; three possibilities are shown below.



Which is the highest spatial frequency, ν_{max}, that can possibly be recorded using the sensor?

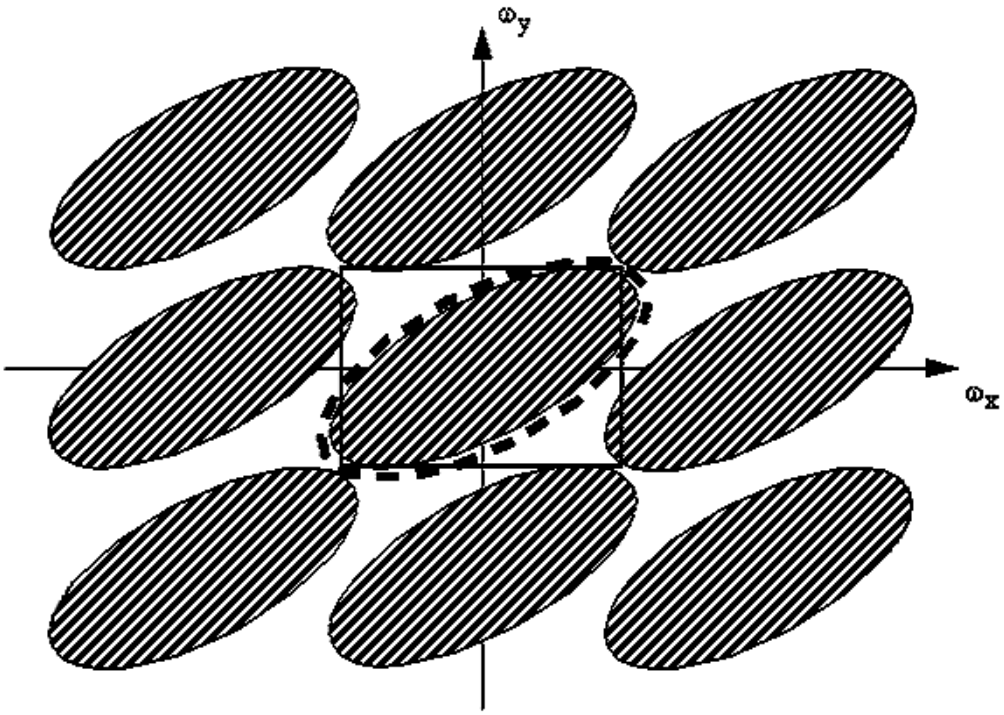
In chapter 13 we have seen that $\nu = \frac{1}{\Delta s} = \sqrt{\nu_x^2 + \nu_y^2}$. Thus $\nu_{max} = \sqrt{\nu_{max,x}^2 + \nu_{max,y}^2}$. The sampling distance is 25 μm in both x and y directions, yielding $\nu_{max,x} < \frac{\nu_{S,x}}{2} = \frac{1}{2 \cdot 25 \times 10^{-6}} = 2.0 \times 10^4$ m^{-1}, and similarly $\nu_{max,y} < 2.0 \times 10^4$ m^{-1}. Using these values, we get $\nu_{max} = 2.8 \times 10^4$ m^{-1}, or 28 mm^{-1}, which is the highest spatial frequency that can be correctly recorded by the sensor. Since $\nu_x = \nu_y$ for this case, the orientation of the pattern is such that it forms an angle of 45° with respect to the x and y axes. For patterns consisting of horizontal or vertical lines, the maximum frequency that can be correctly recorded is only 20 mm^{-1}.

These results are food for thought. In our modern world vertical and horizontal structures tend to be predominant (for example, consider a brick wall). When recording such structures with a camera having an area array sensor, it is not optimal to have a sensor grid with horizontal rows and vertical columns as in the example above. By rotating the sampling grid by 45°, see figure below, higher spatial frequencies in the horizontal and vertical directions can be recorded. It is interesting to note that Fuji has developed a "super CCD" sensor with this orientation of the detector grid.



Document scanners are popular computer accessories. In such scanners the sampling distances x_s and y_s are often different, a common specification being 1200 x 2400 dpi (dots per inch). Such scanners are equipped with a linear sensor, having approximately 10000 elements, that is mechanically scanned in a direction perpendicular to the row of elements, thereby accomplishing two-dimensional sampling. The specification 1200 x 2400 dpi means that the sampling distance in the mechanical scan direction is half of that in the direction of the linear sensor. Since one inch is equal to 25.4 mm, such a scanner has a sampling frequency of 47 mm^{-1} along the sensor and 94 mm^{-1} in the mechanical scan direction. Using the same equations as in the example above, the maximum spatial frequency that can be correctly recorded in a document is found to be 53 mm^{-1} (for a pattern whose lines form an angle of 27° with respect to the linear sensor).

Finally, we will end this chapter by a philosophical note. When looking at the illustration of $\hat{F}_S(\omega_x,\omega_y)$ on page 61, with an infinite number of copies of $\hat{I}_R(\omega_x,\omega_y)$, one may well ask whether it is not possible to allow the copies to extend somewhat into the solid rectangular box in the center, as long as the shaded ellipses don't overlap. In such a case we can still isolate the central ellipse, without interference from the others, see figure on next page. By setting $\hat{F}_S(\omega_x,\omega_y)=0$ outside the dashed region in the ω_x,ω_y plane, and performing the inverse Fourier transform, we can still re-create the continuous function $I_R(x,y)$, although we have <u>not</u> fulfilled the sampling theorem: $\nu_{s,x} > 2\nu_{max,x}$ and $\nu_{s,y} > 2\nu_{max,y}$.



There is, of course, a catch (there always is when you violate the rules of physics) - this method only works if you know something beforehand about the spatial frequency content of the sampled image. In this case you must know that the spatial frequency content of $I_R(x,y)$ is represented by an ellipse of a certain size and orientation in the ω_x, ω_y plane. Such information is usually not available, making it impossible to exclude the copies from the relevant information in the center. In cases where $\hat{F}_S(\omega_x, \omega_y)$ consists of elliptical shapes (or other simple geometries), like in the figures above, you might be able to guess which is which. But when working with real image information, $\hat{F}_S(\omega_x, \omega_y)$ often looks like scattered bird droppings on a pavement, or smudge on a window pane. It is then impossible to isolate the central part and suppress the copies, because you can't tell which is which.

Moral: Stick to the sampling theorem, $\nu_{s,\,x} > 2\nu_{max,\,x}$ and $\nu_{s,\,y} > 2\nu_{max,\,y}$, if you want to keep out of trouble.

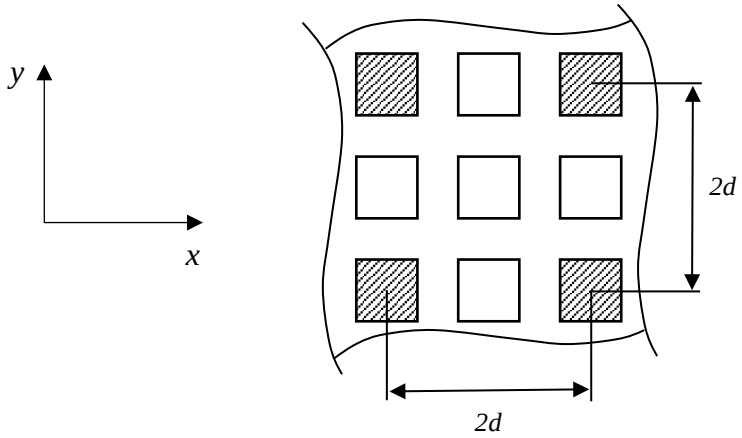
Sampling with an RGB Bayer mosaic pattern

As mentioned in chapter 1, many area array sensors are equipped with an RGB color filter pattern on the pixels so that color images can be recorded. The most common pattern is the so-called Bayer pattern.

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

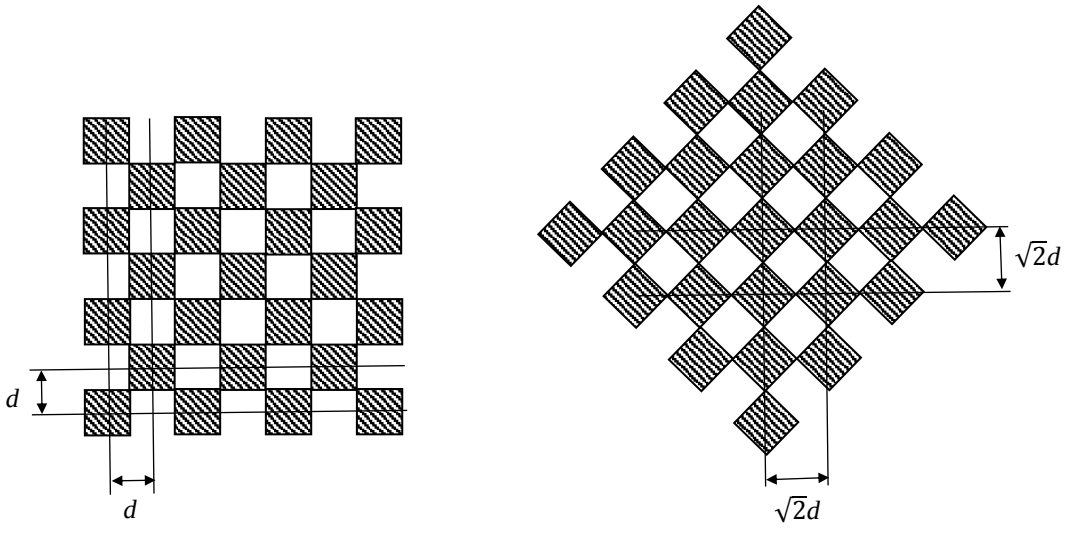
Bayer color mosaic pattern layout

In the Bayer pattern the sampling density is different for the different colors. The green spectral band has the highest sampling density, while the blue and red bands have a lower sampling density. This means that aliasing will occur at different spatial frequencies for different color bands, and furthermore, as we have seen, the pattern orientation is of importance. Since blue and red are sampled with the same density, we get only two different cases that need to be studied, green and red/blue. Let us start with the red/blue case.



The pixel pitch (center-to-center distance between neighboring pixels regardless of color) is denoted by d. The pixel pitch for the red and blue bands will then be $2d$ as shown in the figure. The sampling frequency in the x and y directions will be $\nu_{s,x} = \nu_{s,y} = \frac{1}{2d}$. This means that we can record correctly only frequencies half as high as we could with a black and white sensor with no Bayer pattern.

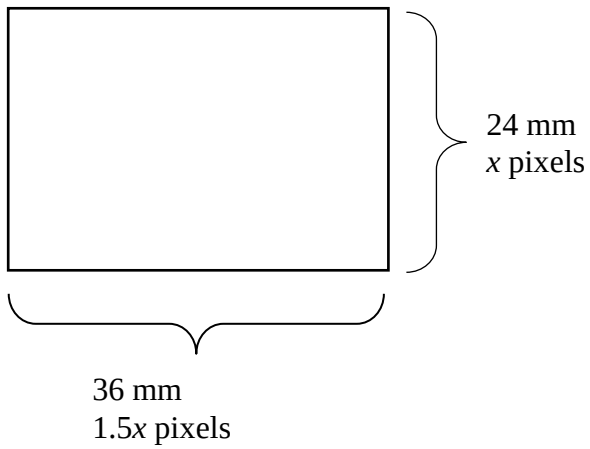
Let us now consider sampling in the green band. It may seem a little confusing that the green pixels don't form regular rows and columns. But if we rotate the sensor by 45 degrees we get the ordinary pattern of rows and columns, but with a pixel pitch of $\sqrt{2}d$.



In the rotated sensor the pixel edges are no longer horizontal or vertical, but that does not influence the sampling at all (it will only influence the directional properties of the sensor MTF). Looking at the rotated sensor, we get the sampling frequencies $\nu_{s,x} = \nu_{s,y} = \frac{1}{\sqrt{2}d}$. The highest spatial frequency that can be recorded correctly (the Nyquist frequency) is then $\frac{1}{2\sqrt{2}d}$ for vertical and horizontal pattern orientation. But, as we have seen previously, for a pattern with 45 degree orientation we can record frequencies that are $\sqrt{2}$ times higher. This means that the maximum frequency is $\frac{1}{2d}$. Since we have already rotated the sensor by 45 degrees, it means that for a non-rotated Bayer sensor matrix, like the one on page 65, we can record vertical and horizontal patterns up to a frequency of $\frac{1}{2d}$, i.e. the same as for a black and white sensor. This may seem a bit surprising, because in a black and white sensor we have twice as many pixels (i.e. sampling points) compared with the number of green pixels in a Bayer sensor (if the total number of pixels is the same). That should influence the sampling somehow. Yes, it does! For line patterns parallell to the rows or coulumns of the sensor, the black and white and Bayer sensors will have the same Nyquist frequency. But for a 45 degree pattern the black and white sensor Nyquist frequency is twice that of the Bayer sensor. One easily gets confused by all these different sampling cases, and therefore the results are summarized in the table below. Also, an example of the practical implications for one specific sensor is given as an example.

Sensor (d = center-to-center distance for pixels)	Nyquist freq. for vertical/horisontal pattern	Nyquist freq. for 45 degree pattern
Black and white	$\frac{1}{2d}$	$\frac{1}{\sqrt{2}d}$
Bayer, red/blue pixels	$\frac{1}{4d}$	$\frac{1}{2\sqrt{2}d}$
Bayer, green pixels	$\frac{1}{2d}$	$\frac{1}{2\sqrt{2}d}$

Example: Sony A7R is a digital camera equipped with a full frame (24 mm x 36 mm) Bayer mosaic sensor. The sensor has 36 megapixels (counting all colors). It has no anti-aliasing filter. Let's first calculate the center-to-center distance for the pixels (the pixel pitch). Let's assume that we have *x* pixels along the 24 mm side of the sensor. We then have 1.5*x* pixels along 36 mm side.



The total number of pixels $= 1.5x \times x = 36 \times 10^6$, which yields x = 4899. Consequently 4899 pixels will occupy a distance of 24 mm, which yields a pixel pitch of d = 4.9 μm. Using this d-value in the table on the previous page, we get

Sensor with d = 4.9 μm	Nyquist freq. for vertical/horizontal pattern (mm^{-1})	Nyquist freq. for 45 degree pattern (mm^{-1})
Red/blue pixels	$\frac{1}{4d} = 51$	$\frac{1}{2\sqrt{2}d} = 72$
Green pixels	$\frac{1}{2d} = 102$	$\frac{1}{2\sqrt{2}d} = 72$

For other pattern orientations, the Nyquist frequency will be between the values given for vertical/horizontal and 45 degree orientation. The fact that aliasing occurs at different frequencies for different colors means that we can expect to get colored aliasing patterns (color moiré) when imaging small grid-like structures.

19. Problems

1) When recording images, one can gain a lot of time by detecting picture elements (pixels) in parallel rather than in series. Consider the following example:
You want to record an image in the form of 500 x 500 pixels. Each pixel must correspond to a light measurement lasting 40 ms (exposure time). What image frequency (images per second) can you get using
a) an area array sensor with 500 x 500 detectors?
b) a linear sensor with 500 detectors?
c) a single detector element?
Assume that the time for data read-out and transmission is negligible.

2) In a semiconductor detector, an energy of at least 1.2 eV is required to create an electron-hole pair. What is the longest light wavelength for which the detector is sensitive?

3) In a photomultiplier tube, the current amplification occurs when an electron collides with a dynode and knocks out more secondary electrons, which in turn accelerate toward the next dynode where they knock out new secondary electrons, and so on. The amplification per dynode can be written as $k \cdot V^{\alpha}$, where k and α are constants and V is the voltage difference between adjacent dynodes.
a) Set up an expression for the total current amplification in the photomultiplier tube if it contains n dynodes.
b) For a certain photomultiplier tube, $k = 0.15$ and $\alpha = 0.7$. Assume that the tube consists of 10 dynodes, plus anode and cathode. The voltage difference between all the dynodes is the same, and we assume that we also have the same voltage difference between the cathode and the first dynode, as well as between the anode and the last dynode. Set up an expression for the total current amplification as a function of the ***total*** voltage, U, over the tube. How much does the current increase when you double U?

4) A photo detector with a quantum conversion efficiency of 10% is exposed to a photon flow whose average is 100 $s^{-1.}$ What is the maximum SNR that can be obtained with a measurement time of a) 1 s, b) 10 s, and c) 100 s?

5) A digital camera is equipped with an area array sensor consisting of a matrix of light sensitive elements (pixels). Each pixel detects photons during the exposure time, and then gives an output signal in the form of a digital value (integer number). The digital value, which is in the interval 0-65535, corresponds to the number of photons that were detected during the exposure time. The camera is used to record an image of a uniformly bright object, resulting (not surprisingly) in a uniformly grey digital image. Superimposed on the image is a certain level of noise, however, resulting in a random variation in the pixel values. Using a computer program, the standard deviation for the pixel values is found to be 0.27% of the mean pixel value. The maximum pixel value was found to be 61796. Explain why these results cannot be correct (there has to be an error somewhere, for example in the computer program).

6) You get a phone call where a person says: "Hi, my name is Peter and I work for a company called Vision-Light. We plan to build a small-size illuminance meter for measuring extremely low levels. The specification for the instrument has been written down by the sales department. In order to check if it sounds reasonable, I would like to get an expert statement from you. I will fax you the specs." A few minutes later you have the following document in your hand:

****
Specification for illuminance meter "Faint-Lite":

Illuminance levels: 1.0×10^{-3} - 1.0×10^{-8} lux
Accuracy (measured as standard deviation): 3%
Detector area: 1.0 x 1.0 mm^2
Number of measurements per second: Minimum one per sec.
****

After thinking for a while (your thoughts include the estimate that 75 lumens correspond to approximately one watt) you call Peter and say ... yes tell me what you say. The answer must include a physical motivation.

7) We are recording light with a photomultiplier tube (PMT), having a quantum conversion efficiency of 30%. Individual photons will give rise to current pulses, and by counting the number of pulses during a well-defined period of time we get an estimate of the photon flux. Assume that we count photons during 1.0 millisecond, and that we perform repeated measurements. We get a standard deviation that is 12% of the mean value (the only source of noise is photon quantum noise).

a) Calculate the mean value (number of detected photons).
b) Calculate the signal-to-noise ratio (SNR).
c) What would the SNR have been if the quantum conversion efficiency had been 100%?
d) What was the real photon flux in the experiments, i.e. how many photons hit the PMT per second?

8) Because of photon noise, light measurements are always encumbered by uncertainty. One does not know if a change in the measured value is due to a change in light level of the measured object or to a random variation caused by photon noise. The only thing we can do is consider probabilities. For photon noise, which is Poisson-distributed, the probability that the measured value deviates from the average by more than $\pm\sigma$ (one standard deviation) is 33%. The probability of deviation of more than $\pm 2\sigma$ and $\pm 3\sigma$ is 5% and 0.3%, respectively. Consider the case where we are recording a digital image (512 x 512 pixels) of an object with perfectly uniform brightness. The exposure time for each pixel is such that, on average, we expect 10 000 photons to be recorded. An 8-bit ADC is used, and adjusted so that the digital value 200 corresponds to 10 000 recorded photons. How many pixels are expected to have values ***outside*** the interval 196 to 204 in a recorded image? (Only photon noise needs to be considered)

9) The faintest stars that can be seen by the naked eye are of the 6:th magnitude. A star of this magnitude produces an irradiated power of approximately 10^{-11} W/m^2 at the earth's surface. What SNR will this produce in the eye, assuming the following:

Pupil diameter: 6mm (eye adapted to darkness)
Quantum efficiency: 3%
λ= 550 nm
The time constant of the eye (which corresponds to the time during which photons are accumulated): 0.2 s

10) In many types of equipment for light measurement (especially at higher light levels), one does not count the photons, but instead simply measures the current from the detector (after amplification). Besides the photon noise, there is in this case also the electronic noise in amplifiers, etc. The latter noise has no coupling at all to the photon noise, and is therefore independent of the light level. The total noise (measured as RMS noise in the output signal) in such cases can be written as $n_{tot} = \sqrt{n_{ph}^2 + n_e^2}$, where n_{tot} = the total noise, n_{ph}= the photon noise and n_e = the electronic noise. By measuring according to these conditions, the average current (after amplification) is measured to be 1.5 mA, and the noise n_{tot} to be 178 μA. If one increases the light level to double, the average current increases to 3.0 mA and n_{tot} to 212 μA.
a) Calculate the SNR for the measurement of the original light level.
b) Calculate the electronic noise n_e.
c) Calculate the SNR one would have obtained without electronic noise in the first measurement.

11) At low light levels it is necessary to measure (i.e. collect photons) for a long period of time to get a good signal-to-noise ratio (SNR). In fluorescence microscopy the situation is further aggravated by specimen bleaching during the measurement, i.e. the light intensity will become weaker and weaker. Assume that the light intensity decays in such a way that it is described by the function $I(t) = I_0 \cdot e^{-\alpha t}$, where $I(t)$ is the rate of detected photons (expected number of photons detected per second) as a function of time t. I_0, the initial rate at the start of the experiment ($t = 0$) is 1.0×10^4 photons per second, and α = 100 s^{-1}. What is the highest SNR we can obtain in such an experiment, regardless of how long the measurement time is prolonged?

12) You have bought a video camera equipped with an area array sensor (size 10.24 mm x 10.24 mm) which has 512x512 detector elements, each with a size of 20 μm x 20 μm. The lens has a focal length of 20 mm and a diaphragm whose f-number can be adjusted between 2 and 16 (f-number = f/D, where f = focal length and D is the diameter of the opening that lets light through the lens). The camera records 50 frames (= images) per second. Estimate the maximum signal-to-noise ratio (at the single pixel level) possible in the images for an object luminance of 1.0 cd/m^2. You are allowed to make the approximation that the light is monochromatic with λ = 550 nm, which means that 1 W of radiated power corresponds to 650 lumens.

13) Although not directly connected with imaging physics, this problem offers interesting insights into information transfer over large distances using low levels of energy (which is interesting in imaging physics).

The Iridium satellite telephone project (financially disastrous!) employs direct communication between cellular phones and a number of satellites in earth orbit. Let's assume that the phone antenna radiates uniformly in all directions, and that the output power is 1 W at a frequency of 30 GHz (microwaves). These phones are to communicate with a satellite at a distance of approximately ten thousand kilometers. It is healthy to wonder if this is really possible using such a low output power. Increasing the output power is not a good idea. Microwaves are efficient for cooking things, and you don't want to cook the brains of the users.
Make a rough estimate of what kind of signal-to-noise ratio you can maximally expect in this type of phone-to-satellite communication, assuming that the satellite receiver is noise-free, and can utilize 100% of the microwave radiation incident on the antenna (approximate area 1 m^2). The frequency content of a telephone signal is such that the maximum time for collecting a signal value (i.e. the "measurement time") is approximately 0.2 ms.

14) On a trip in the space shuttle, the astronauts have brought a really big camera lens with them. The focal length is 2000 mm and the diameter of the front lens is 300 mm. Estimate whether it is physically possible to use this lens for photographing a "Lucia" procession in Stockholm so that the individual candles in the "Lucia crown" will be visible? The altitude of the space shuttle orbit is 350 km.

15) A CD burner uses a solid state laser (wavelength = 780 nm) for burning small dots into a blank disc. Let's assume that a parallel beam of laser light is focused onto the disc by a diffraction-limited lens. What is the minimum theoretical limit for the diameter of the focused spot (Airy spot size) on the disc.

16) The resolution limit of the human eye is approximately one arc minute (= 1/60 degrees). This means that if we look at two equally bright point sources of light, separated by this angular distance, we can just barely discern that there are two sources and not a single one (Rayleigh criterion).

a) Estimate how far away from a color television screen you must be seated in order not to notice the individual red, green and blue color dots. The center-to-center distance between the dots is approximately 0.25 mm.

b) Is the eye nearly diffraction-limited (i.e. are the aberrations small compared with diffraction effects)? Assume that the pupil diameter is 3 mm. (You don't have to take into account that the eye is filled with liquid; you may do the calculation as if it were filled with air)

17) There is strong disagreement between Dr. Hackenbush and his laboratory assistant, Donald, at the astronomical observatory. Donald has mounted on the telescope a video camera, with which he records images and stores them in a computer memory. He then uses an image processing program to "improve" the images. By increasing the contrast in the recorded images, Donald claims that he can distinguish double stars having an angular separation of less than $1.22\lambda/D$, where λ = the light wavelength and D = the lens diameter of the telescope. "Impossible!" roars Dr. Hackenbush who has recently attended a course on Imaging Physics (he didn't pass the examination). "That which is not found in the original optical image cannot be extracted using computer tricks." Who is right?

18) Astronomical telescopes often have very large-diameter mirrors. There are two reasons for this. First, a large mirror diameter means high resolution. Second, it means that many photons are collected, and therefore faint objects can be studied. Assume that we want to study double stars, i.e. stars that are located (on an astronomical scale) close together. We will also assume that they have similar intensities.

a) What is the mathematical relationship between mirror diameter and the smallest angular separation between two stars that can be resolved according to the Rayleigh criterion (i.e. we want to see two separate spots of light, and not a single one). To get a realistic number, let's calculate this separation for a mirror diameter of 5.0 meters and a wavelength of 550 nm. (A mirror behaves in the same way as a lens concerning resolution)

b) You want to measure the photon flux from the stars using a photomultiplier tube and photon counting equipment. For one and the same measurement time, how will the signal-to-noise ratio vary as a function of the mirror diameter? Assume that other sources of noise are negligible compared with photon quantum noise.

c) What is meant by quantum conversion efficiency for a photomultiplier tube, and what is the mathematical relationship between the signal-to-noise ratio and the quantum conversion efficiency in the above measurement?

19)

a) What is meant by point spread function for an imaging system?

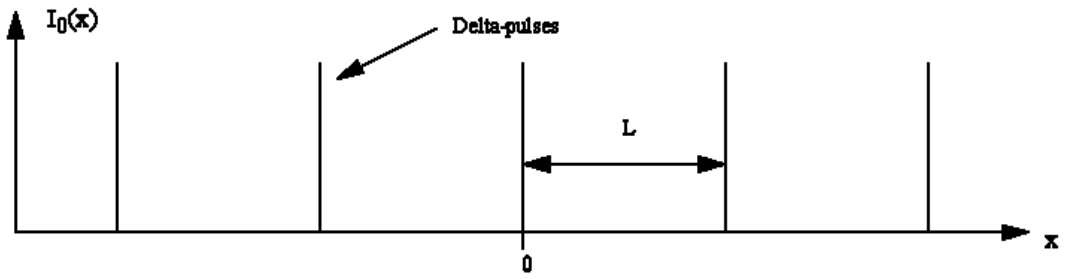
b) How can you calculate the image function if you know the object function and the point spread function?

c) How can you calculate the *MTF* (Modulation Transfer Function) and the *PTF* (Phase Transfer Function) if you know the ***psf*** (point spread function)?

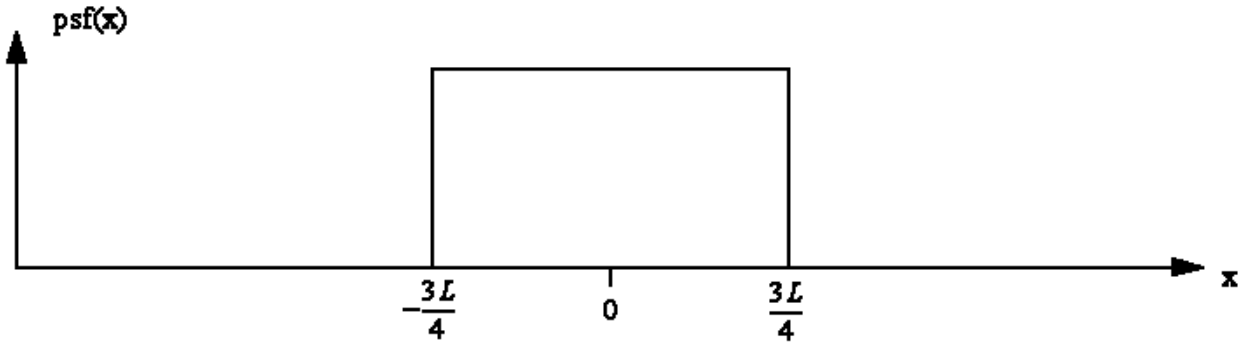
d) What is the physical interpretation of *MTF* and *PTF*, and how should they ideally look in order the get perfect image quality?

20) Calculate and plot the image function, $I_B(x)$, in this one-dimensional case:

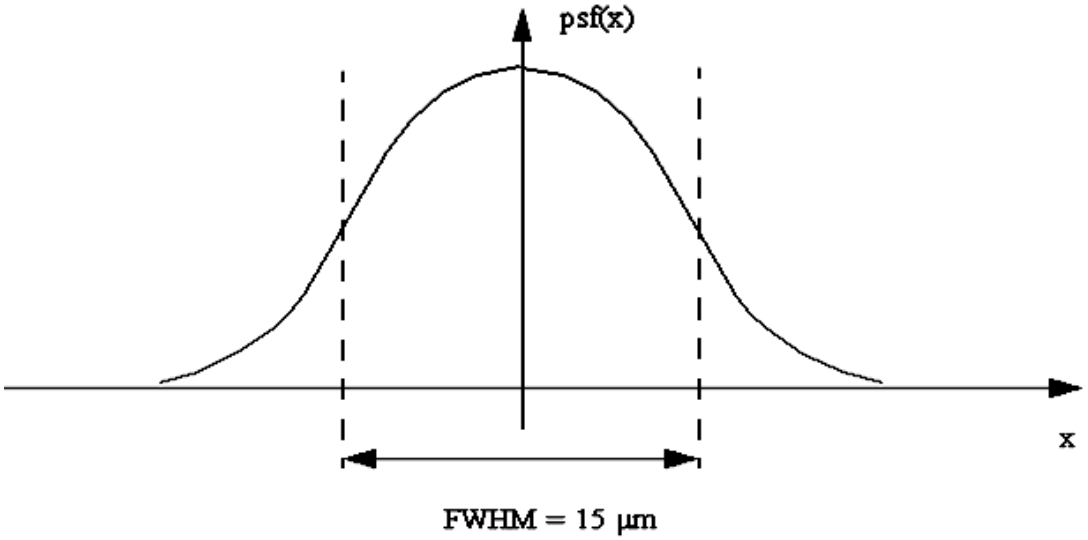
Object function = $I_0(x) = \sum_{n=-\infty}^{+\infty} \delta(x - nL)$, where n = is an integer number, see figure below:



The point spread function of the imaging system is a rectangular function as shown below:



21)



The figure above shows the measured point spread function for a detector. FWHM = Full Width at Half Maximum, i.e. the total width for that part of the curve where the psf-value is > 50% of the max. value. The shape of the curve is approximately Gaussian, i.e. $psf(x) \approx e^{-ax^2}$, where a is a constant. At what spatial frequency can we expect a detector *MTF*-value of 0.10?

22) Assume that we are taking aerial photographs from an altitude of 1000 m with a camera that has an f = 300 mm lens. The exposure time is 1/250 s. The camera points vertically down towards the ground, and the airplane travels at a speed of 750 km/h. Because of the plane's speed, we will get motion blur in the plane's flight direction (we assume that we do not have motion compensation in the camera). Assume that the lens quality is so high that its influence on the picture quality can be neglected

a) Calculate $MTF_{motion\ blur.}$

b) Assume that we photograph a pedestrian crosswalk with stripes oriented perpendicular to the flight direction. Both the black and the white stripes are 50 cm wide. Draw what the exposure distribution in the film will look like along a line in the flight direction.

c) By letting the film move at a constant speed in the flight direction during the exposure time, the motion blur can be avoided. How fast must the film move?

23) To make *MTF* measurements of lenses we have acquired a test pattern consisting of a sinusoidally varying luminance, with a period length of 1.0 mm and close to 100% modulation. We image this test pattern with a lens having a diameter of 50 mm and a focal length of 250 mm. Since we use white light, we assume that the "average wavelength" is 550 nm. The degree of modulation in the image of the pattern is measured for a number of distances between lens and test pattern, and the following results are obtained:

Distance (meters)	Degree of modulation in image
2.0	0.90
7.5	0.65
15	0.36
22	0.16
30	0.04
38	0.00

a) Plot an approximate *MTF* curve from the data given in the table.
b) The modulation measurements were made by moving a detector in the image plane. The width of the detector was 5.0 μm (constant sensitivity over that area). Correct the *MTF* values in a) so that the influence of the detector is eliminated.
c) Would you say that the lens is nearly diffraction-limited (i.e. negligible aberrations)?

24) A lens whose *MTF* can be approximated by $MTF(\nu) = \begin{cases} 1 - \frac{\nu}{100}\ , & \nu \le 100 \\ 0\ , & \nu > 100 \end{cases}$, where ν is the spatial frequency in units of mm^{-1}, is used together with a square detector with side length L. The detector is moved in the image plane to record the light distribution. What is the maximum value for L if the total *MTF* for lens and detector must have a value of at least 0.5 at a spatial frequency of 30 mm^{-1} (let's assume that we are imaging a sine-pattern whose peaks are oriented parallel to two of the detector's edges and perpendicular to the other two)? (Hint: Use trial-and-error to solve the equation obtained)

25) The *MTF* for a lens can be approximated as: $MTF(\nu) = \begin{cases} 1 - \frac{\nu\lambda f}{D}\ , & \nu \le \frac{D}{\lambda f} \\ 0\ , & \nu > \frac{D}{\lambda f} \end{cases}$

where ν = spatial frequency, λ = light wavelength, f = focal length and D = lens diameter. We have assumed that the lens is entirely free from aberrations and that the image distance is equal to the focal length (we are imaging an object that is located far away).
Use this information in order to decide how many light sensitive elements an area array sensor of size 24 mm x 36 mm must have if no false spatial frequencies are to be recorded when using a 50 mm lens with an f-number of 5.6 (f-number = f/D). Assume that the light wavelength is 550 nm.

26) An area array sensor consists of detector elements with a center-to-center distance of 7.0 μm. The sensor is used in a digital camera that is equipped with a lens whose focal length is 18 mm. The lens can reproduce a maximum spatial frequency of 140 mm^{-1} (measured in the image plane). A curtain is to be photographed with this camera. The cloth of the curtain

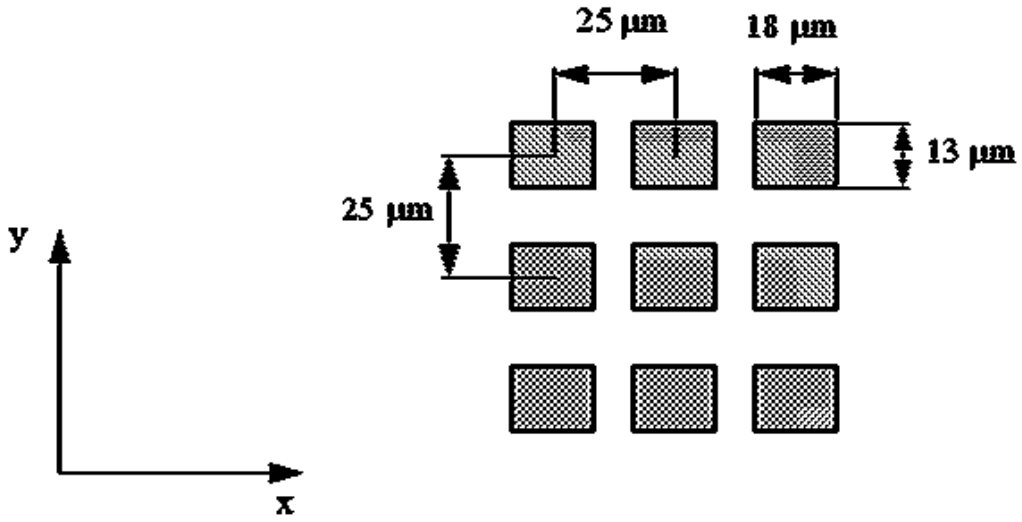
consists of a square mesh (net-like structure) with a 1.0 mm center-to-center distance between the threads. For what distances between camera and curtain is there a risk that we get aliasing in the images?

27) We are sampling an image using a sampling frequency of ν_s = 100 mm^{-1}. What spatial frequencies will be found in the reconstructed image, if the original image contained the frequencies 40, 80 and 130 mm^{-1}?

28) A linear sensor consists of one row of 512 light sensitive detector elements, each with the size 25 μm x 25 μm, and a center-to-center distance of 25μm (i.e. they are located edge to edge). Assume that the light sensitivity is completely uniform over each detector's surface. A periodic pattern is projected so that the intensity will vary sinusoidally along the linear sensor.

a) What is the highest spatial frequency this pattern may have if it is to be recorded with the correct frequency by the sensor?
b) What is the value of the detector's *MTF* at this highest frequency?
c) How should the ideal sensor *MTF* look in order to get as real an image rendition as possible, but avoiding false frequencies. (Assume still 25 μm sampling distance.)
d) How should the sensitivity of a single detector element vary in the linear sensor's lengthwise direction in order to obtain an *MTF* according to exercise c) above? Comment on the result. (Is it realistic?)

29) An area array sensor consists of 1024 x 1024 detector elements, each with a size of 13 μm x 18 μm. The center-to-center distance between the elements is 25 μm both in the x and y directions, see figure.

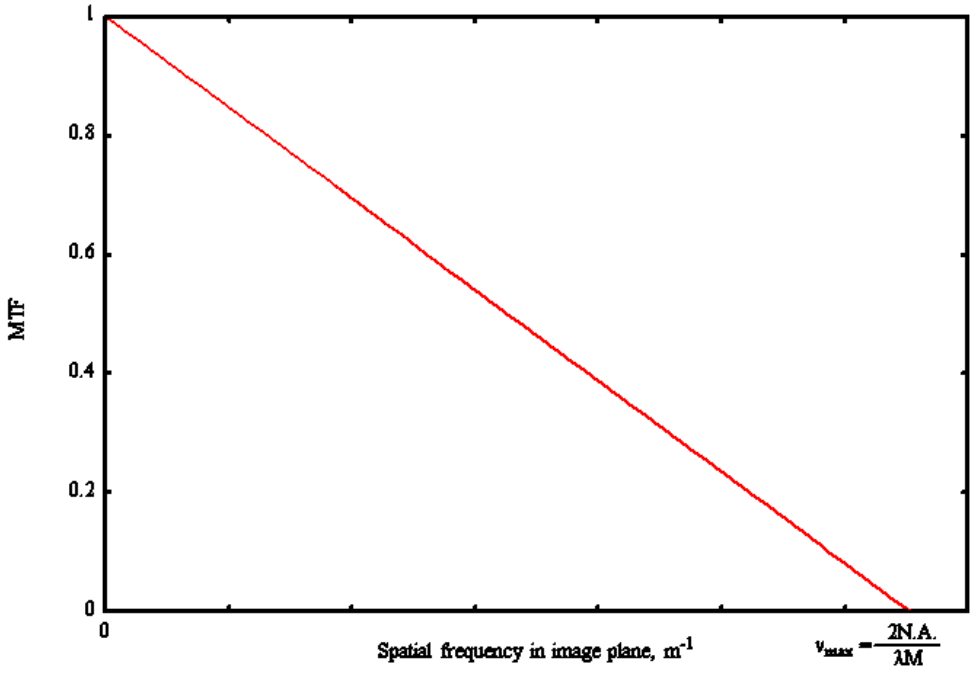


a) Calculate the maximum spatial frequencies that can be correctly recorded in the x and y directions respectively (Nyquist frequency).
b) Calculate $MTF_{detector}$ in the x and y directions at the above frequencies.
c) Assume that we increase the size of the detector elements to 25 μm x 25 μm, which means that they will be located edge-to-edge with no space in between. Will the following properties increase, decrease or remain constant?: 1) The highest spatial frequency that can be recorded correctly (Nyquist frequency), 2) $MTF_{detector}$, 3) The magnitude of the output signal, 4) The signal-to-noise ratio.

30) The linear relationship shown in the figure is a rough estimate of the *MTF* for a microscope objective. N.A. means numerical aperture (a measure of the light collecting ability of the objective) and M is the magnification. To record a microscopic image an area array sensor is placed in the image plane of the microscope objective. The objective magnification M = 100, and the N.A. = 0.9. The individual detector elements in the sensor have a size of 15 μm x 15 μm each, and are located edge-to-edge (no dead space in between). λ = 550 nm.

a) Is the sampling theorem fulfilled, i.e. are the detector elements sufficiently closely spaced to avoid aliasing under all circumstances?

b) Write down an expression for the total *MTF* for the system microscope objective and detector, and sketch what the *MTF* as a function of spatial frequency will look like.



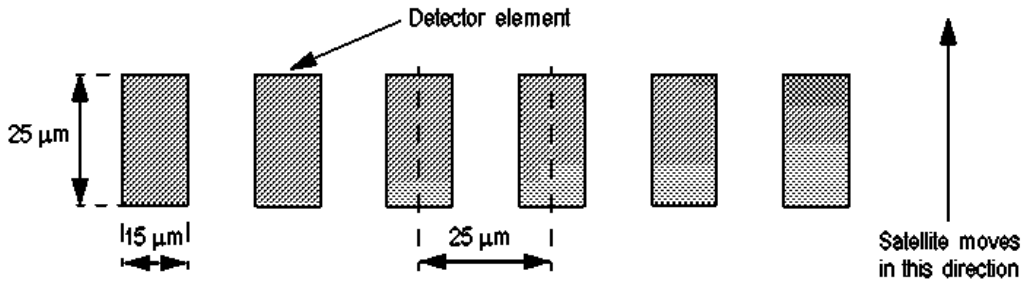
31) A digital camera is equipped with an area array sensor, having a total area of 18 x 14 mm^2. On this area 3600 x 2800 detector elements are uniformly distributed. The camera lens is equipped with a diaphragm which can be used to adjust the amount of light (illuminance) on the sensor. On the diaphragm adjustment ring the following numbers (f-numbers) are printed: 4, 5.6, 8, 11, 16 and 22. These numbers represent the ratio , where f = focal length and D = the diameter of the aperture that lets light into the lens (i.e. effective lens diameter). We will assume that the lens is diffraction limited, which means that we can approximate its *MTF* with $MTF(\nu) = \begin{cases} 1 - \dfrac{\nu \lambda f}{D}, & \nu \le \dfrac{D}{\lambda f} \\ 0, & \nu > \dfrac{D}{\lambda f} \end{cases}$, where ν = spatial frequency. What f-number should we use to avoid aliasing under all circumstances, but still get the best possible image quality? Assume that λ = 550 nm. (You only have to take into account the case where the pattern is orientated in such a way that it is parallel to two of the sensor's edges and perpendicular to the other two.)

32) The video camera in problem **12** is used to make a recording of a politician, who is "properly" dressed in a suit with stripes.

a) What may happen if the stripes of the suit are close together?

b) What is the maximum distance for video-taping the politician if you are to avoid the above phenomenon? The suit has alternating dark and light stripes, each with a width of 1 mm?

33) In this problem we will consider a satellite used for remote sensing. A linear sensor is used, which has 1024 detector elements with a center-to-center distance of 25 μm. The detector elements are rectangular in shape, the width being 15 μm in the direction of the linear sensor and 25 μm in the perpendicular direction, see illustration below.



The light sensitivity of each detector element is uniform within the 15 μm x 25 μm area, and zero outside of this area. The linear sensor is mounted at right angles to the satellite's direction of movement. With this equipment it is possible to record images of the earth by recording the output signals from the 1024 detector elements at regular intervals as the image of the earth moves across the sensor. The altitude of the satellite orbit is 800 km, and the speed relative to the ground is 7500 m/s. The linear detector is located in the image plane of a 2.0 m focal length lens, which is used for imaging the earth's surface. The resolution of the lens is much higher than 25 μm in the image plane.

a) What should the time interval be between successive read-outs of the sensor output, if we are to get the same imaging scale in directions parallel to and perpendicular to the satellite's movement?
b) What is the maximum spatial frequency of a pattern on the ground that can be correctly recorded by this satellite?
c) The linear sensor is an integrating sensor, much like photographic film. This means that it is exposed to light for a certain period of time, and the magnitude of the output signal increases in proportion to the exposure time (up to a certain limit). Two extreme cases are the following:
1) The exposure time is negligible compared with the time interval between successive read-outs.
2) The entire time span between signal read-outs is used for exposure.
Calculate the total system *MTF* for these two cases. Line patterns that are both parallel to and perpendicular to the direction of movement should be considered.

Solutions to problems:

1) **Area array sensor:** All diodes are exposed simultaneously, which means that the entire image will be exposed in 40 ms. It is therefore possible to record an image every 40 ms, which gives 25 images per second (25 Hz).
Linear sensor: One image row (500 pixels) will be exposed every 40 ms. An image consists of 500 rows, and therefore it takes ms, i.e. 20 s, to expose an entire image. In this case an image is recorded every 20 s, which corresponds to 0.05 images per second (0.05 Hz).
Single detector element: An image point (pixel) is exposed every 40 ms. An image consists of 500x500 image points, and therefore takes 10000 s to expose (nearly 3 hours!). This corresponds to 0.0001 images per second (0.0001 Hz).

2) The energy needed to create an electron-hole pair is $h\nu = \frac{hc}{\lambda}$. We therefore get the following equation $\frac{6.63\times10^{-34}\cdot 3.00\times10^{8}}{\lambda} = 1.2\cdot 1.60\cdot 10^{-19}$
which gives a λ of 1.0 μm.

3) a) Since the current is multiplied by $k\cdot V^{\alpha}$ at each dynode, the total amplification becomes $k^{n}\cdot V^{\alpha n}$.
b) The total amplification in the photomultiplier tube becomes $0.15^{10}\cdot V^{0.7\cdot 10} = 5.77\times10^{-9}\cdot V^{7}$. Since we have 12 electrodes (dynodes plus anode and cathode), $V = U/11$. If we insert this into the equation above, we can write the total amplification as $3.0\times10^{-16}\cdot U^{7}$. A doubling of the voltage gives $2^7 = 128$ times higher amplification.

4) $SNR = \sqrt{\overline{N}}$, where $\overline{N}$ = average number of detected photons during the exposure time.
a) $\overline{N} = 0.1\cdot 100\cdot 1.0 = 10 \Rightarrow SNR = \sqrt{10} = 3.2$.
b) $\overline{N} = 0.1\cdot 100\cdot 10 = 100 \Rightarrow SNR = \sqrt{100} = 10$
c) $\overline{N} = 0.1\cdot 100\cdot 100 = 1000 \Rightarrow SNR = \sqrt{1000} = 32$

5) The lowest noise level possible is when only photon quantum noise is present. We then have the following situation: If the average number of detected photons during the measuring interval is $\overline{N}$, the standard deviation $\sigma = \sqrt{\overline{N}}$. The standard deviation divided by the average number will then be $\frac{1}{\sqrt{\overline{N}}}$. In the present situation $\overline{N} < 61796 \Rightarrow$ standard dev. divided by average number > 0.0040, i.e. 0.40%, which is a higher value than was obtained in the measurements. This is unreasonable since 0.40% is the theoretical limit.

6) Let's look at the accuracy requirement, and make a rough estimate. A standard deviation of 3% means that we must have a SNR of 33, which means that we must detect at least 33^2 = 1100 photons. If the quantum conversion efficiency were 100% (which is too optimistic), we would require 1100 incoming photons on the detector each second. 1100 photons with, say, a wavelength of 550 nm represent an energy of $W = 1100 \cdot \frac{hc}{\lambda} = 1100 \cdot \frac{6.63\times10^{-34} \cdot 3.00\times10^{8}}{550\times10^{-9}} = 4.0\times10^{-16}$ J. Since this amount of energy reaches the detector each second, the incident power is 4.0×10^{-16} W. If we assume 75 lumens/watt the incident flux will be $75 \cdot 4.0\times10^{-16} = 3.0\times10^{-14}$ lumens, which is distributed over an area of $\left(1.0\times10^{-3}\right)^2 = 1.0\times10^{-6}$ m^2. The illuminance level will then be 3.0×10^{-8} lux, and getting to lower levels is physically impossible, given the requirements for accuracy.

7)

a) Let $\overline{N}$ denote the mean value of pulses counted during 1.0 ms. Then $\sigma = \sqrt{\overline{N}} = 0.12 \cdot \overline{N} \Rightarrow \overline{N} = \frac{1}{0.12^2} = 69$.

b) $SNR = \sqrt{\overline{N}} = 8.3$.

c) With 100% quantum conversion efficiency, the mean value would have been $\frac{1}{0.30}$ times as large, i.e. 231, which gives *SNR* = 15.

d) 231 photons in 1 ms means 2.3×10^5 photons/s.

8) With an average of 10 000 photons detected per pixel, we will get a standard deviation of $\sqrt{10000} = 100$ or 1% of the average. Thus, σ corresponds to 2 units in the digital image. The interval 196-204 therefore corresponds to $\pm 2\sigma$, which comprises 95% of the values. 5% of the 512 x 512 pixels are therefore expected to fall outside the interval, which means $0.05 \cdot 512^2 \approx$ 13 000 pixels.

9) The power irradiated through the pupil is $10^{-11} \cdot \pi \cdot \left(3\times10^{-3}\right)^2 = 2.83\times10^{-16}$ W. Dividing this by the photon energy, $\frac{hc}{\lambda} = 3.62\times10^{-19}$ J, we find that 783 photons hit the retina per second. With a quantum efficiency of 0.03 and an exposure time of 0.2 s we get, on the average, 4.7 recorded events, which gives a SNR of 2.2 (square root of 4.7)

10)

a) $SNR = \frac{1.5\times10^{-3}}{178\times10^{-6}} = 8.4$

b) Doubling the light level increases the photon noise by a factor of $\sqrt{2}$. The electronic noise remains constant. If we denote the photon noise for an output signal of 1.5 mA by $n_{ph,}$ we get the following equations:

$178\times10^{-6} = \sqrt{n_{ph}^2 + n_e^2}$

$212\times10^{-6}=\sqrt{2n_{ph}^2+n_e^2}$

Solving these equations, we get n_{ph} = 115 μA and *ne* = 136 μA.

c) In the absence of electronic noise, we get $SNR=\frac{1.5\times10^{-3}}{115\times10^{-6}}=13.0$

11) With a measurement time of T seconds the expected number of photons is given by $\overline{N}=\int_0^T I_0e^{-\alpha t}dt$. $T=\infty$ gives $\overline{N}=\frac{I_0}{\alpha}=100$, which gives $SNR=\sqrt{\overline{N}}=10$.

(NOTE: Already after 0.05 seconds, the expected number of photons is 99, so very little is gained by prolonging the measurement time beyond 0.05 s)

12) The highest possible *SNR* one can get is if photon quantum noise is the only source of noise. In this case $SNR=\sqrt{\overline{N}}$, where $\overline{N}$ = the expected number of photons to be detected during the measurement time (= the average value obtained from many repeated measurements). Thus, we must determine the number of photons detected by one detector element. The relationship between the illuminance in the image plane, *E*, and the luminance, *L*, focal length, *f*, and lens diameter, *D*, is given by $E=\frac{L\pi}{4\left(\frac{f}{D}\right)^2}$ lux. If the width of a detector element is denoted by *s*, we get a luminous flux on the element of $E\cdot s^2=\frac{L\pi s^2}{4\left(\frac{f}{D}\right)^2}$ lumens.

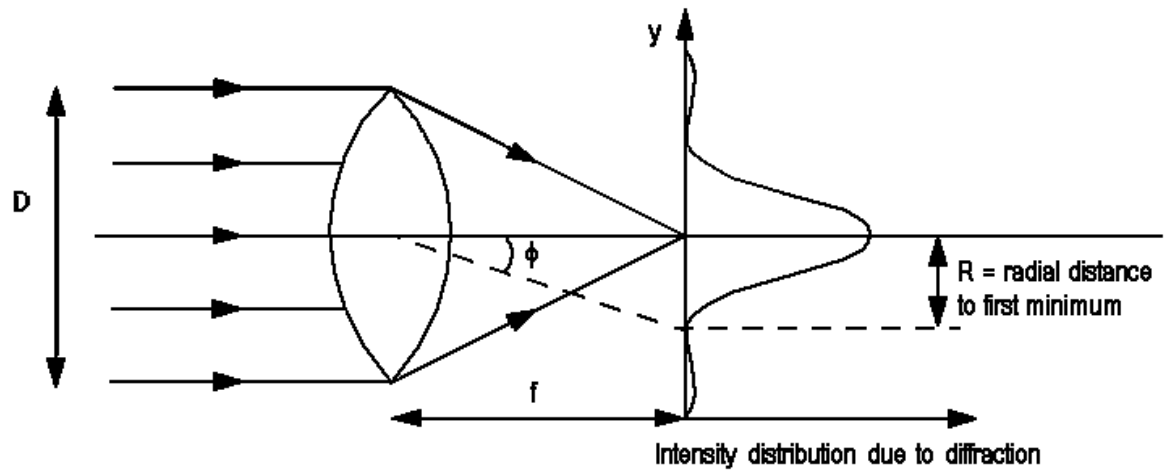
Since we have 650 lumens/watt, the irradiance on one detector element will be $\frac{L\pi s^2}{4\left(\frac{f}{D}\right)^2 650}$ W. Each photon having an energy of $\frac{hc}{\lambda}$, we get a photon flux on the detector element of $\frac{L\pi s^2\lambda}{4\left(\frac{f}{D}\right)^2 650hc}$ photons/s. For an exposure time of 1/50 s, the number of photons will be $\frac{L\pi s^2\lambda}{4\left(\frac{f}{D}\right)^2 650\cdot hc\cdot 50}$, which, with the numbers given in the problem, yields 6700 photons. If all of these are detected (quantum conversion efficiency 100%) we get SNR = 82. (This is, of course, an upper limit. In a real case the quantum conversion efficiency is lower, and the measurement time shorter.)

13) Let P be the transmitter power, and h the altitude of the satellite. The power density at the satellite will then be $\frac{P}{4\pi h^2} = \frac{1.0}{4\pi \cdot 1.0 \times 10^{14}} = 8.0 \times 10^{-16}$ W/m^2. If we assume an antenna area of 1.0 m^2, and a quantum conversion efficiency of 100%, the received power will be = 8.0×10^{-16} W. A measurement time of 0.2 ms will give a received energy of $8.0 \times 10^{-16} \cdot 0.2 \times 10^{-3} = 1.6 \times 10^{-19}$ J. Each microwave photon has the energy $hf = 6.63 \times 10^{-34} \cdot 30 \times 10^{9} = 2.0 \times 10^{-23}$ J. The received energy thus corresponds to $\frac{1.6 \times 10^{-19}}{2.0 \times 10^{-23}} = 8000$ photons. If photon quantum noise is the only source of noise, SNR = $\sqrt{N}$ = 89. This means that the standard deviation in the signal level is approximately 1%, which is quite satisfactory for telephone conversation.

14) Let's make an estimate:

Let d denote the distance between candles, and assume that λ = 550 nm. Lens diameter D = 0.30 m and the altitude $h = 3.5 \times 10^5$ m. If we use the Rayleigh criterion we get: $D \cdot \frac{d}{h} = 1.22\lambda$, which gives $d = \frac{1.22\lambda h}{D} = \frac{1.22 \cdot 550 \times 10^{-9} \cdot 3.5 \times 10^5}{0.30} = 0.78$ m, which is the shortest distance between candles that we can resolve. The candles of the "Lucia crown" are definitely located closer together than this, so they cannot be resolved. (The rest of this solution is optional) The Rayleigh criterion is valid for two point sources, and the Lucia crown consists of several candles. Can this change the conclusion? As we shall see, it can not: Regard the Lucia crown as a periodic luminance variation with period d. A diffraction-limited lens can image a maximum spatial frequency of $\frac{D}{\lambda f}$, which translates to $\frac{D}{\lambda h}$ in object space. The corresponding period is $\frac{\lambda h}{D}$. Therefore, the requirement will be that $d > \frac{\lambda h}{D}$, a result similar to the one given by the Rayleigh criterion, i.e. the astronauts cannot photograph the individual candles.

15)



For a diffraction-limited lens the radius of the Airy spot is given by $R = \frac{0.61\lambda}{\sin\beta}$. Therefore the theoretical lower limit of R is given by $R_{\min} = 0.61\lambda = 0.61 \cdot 780 \times 10^{-9} = 0.48 \times 10^{-6}$ m = 0.48 μm. The minimum spot diameter is 0.95 μm. In reality $\sin\beta < 1$, but it can in practice be as large as 0.9 (an example is microscope objectives). It is therefore realistic to get a spot size (diameter) of about $2 \cdot \frac{0.61 \cdot 780 \times 10^{-9}}{0.9} \approx 1.1$ μm. The possibility to make a cheap and (approximately) diffraction-limited objective with this focusing performance rests on the fact one is imaging on the optical axis with monochromatic light. Thus the only aberration that needs correction is spherical aberration.

16)

a) 1 arc minute is equivalent to $\frac{1}{60} \cdot \frac{\pi}{180} = 2.9 \times 10^{-4}$ rad. If the viewing distance is denoted L we get $\frac{0.25 \times 10^{-3}}{L} = 2.9 \times 10^{-4}$, which gives L = 0.86 m. Thus, the minimum viewing distance is approximately one meter.

b) The angular resolution for a diffraction-limited lens of diameter D is $\frac{1.22\lambda}{D}$. For D = 3 mm and λ = 550 nm, we get an angular resolution of 2.2×10^{-4} rad, which is equivalent to 0.77 arc minutes. This is only slightly better than the performance of the eye, and therefore it is fair to say that the eye is nearly diffraction-limited.

17) Donald is right! The resolution limit according to the Rayleigh criterion, 1.22λ/D, will provide an intensity pattern with a dip of 26% between the two peaks (see page 19 in compendium). By subtracting a constant background level, and increasing the image contrast, an original dip of much less than 26% can be increased so that it becomes 26% or more. Of course, this works only as long as there is a dip at all between the peaks; once the peaks have merged completely, increasing the contrast doesn't work any more (this happens for an angular distance of 1.02λ/D).
One should bear in mind, however, that increasing the image contrast also increases the noise. The method therefore works best for images with a high signal/noise ratio.

18)

a) $\alpha = \frac{1.22\lambda}{D}$. For λ = 550 nm and D = 5.0 m we get α = 1.3×10^{-7} radians = 0.03 arc seconds. In reality turbulence in the atmosphere will reduce the resolution to a considerably lower figure.

b) The number of photons recorded per time unit is directly proportional to the mirror area, i.e. to the diameter squared. SNR is proportional to the square root of the (average) number of detected photons. This means that SNR is directly proportional to the mirror diameter.

c) Quantum conversion efficiency is the percentage of incident photons recorded by the detector. SNR is proportional to the square root of the quantum conversion efficiency.

19) a) The image produced by the system when the object is a point source.

b) The image function is equal to the object function convolved by the point spread function (*psf*).

c) *MTF* is the modulus of the Fourier transform of the *psf*, normalized to unity at zero spatial frequency. *PTF* is the argument of the Fourier transform of the *psf*.

d) Let's assume that we are imaging a pattern with sinusoidally varying intensity, and a degree of modulation of m_1. If our imaging system is linear (as demanded by *OTF* theory) the image will also be a sine pattern, but its modulation will be m_1 multiplied by the *MTF* value. The *MTF* value depends on the spatial frequency of the imaged pattern (one usually refers to the spatial frequency in the image, not the object). In addition to the modulation loss, the image sine pattern will also be phase-shifted. The magnitude of this phase shift is given by *PTF*. For further details, see chapter "Mathematical Representation of the Image Reproduction Process." The ideal *MTF* has a constant value of unity for all spatial frequencies. The ideal *PTF* has a value of zero for all frequencies (it can be shown that if the *PTF* is equal to a constant times the spatial frequency, i.e. direct proportionality, this also produces a perfect image, only a little shifted sideways)

20) The image function is given by the object function convolved by the point spread function:

$$I_B(x) = I_0 \otimes psf = \int_{-\infty}^{+\infty} I_0(t)\, psf\,(x-t)dt\,.$$

Below are two different solutions:

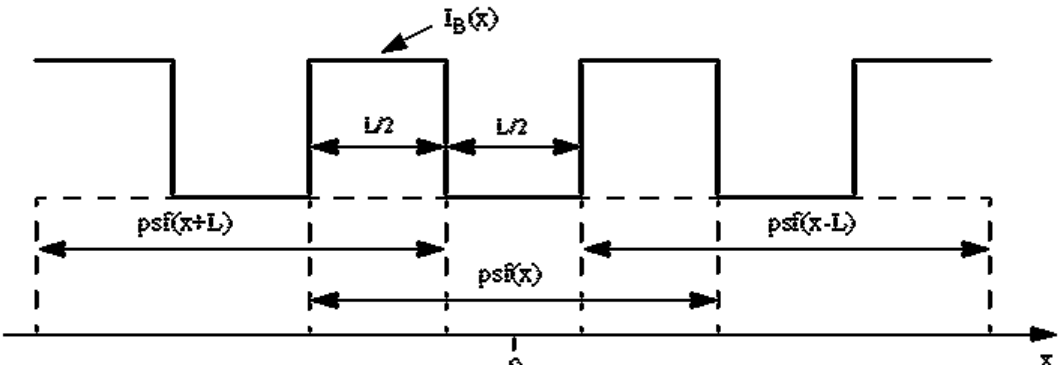
No. 1 (for the mathematically inclined):

$$I_B(x) = \int_{-\infty}^{+\infty} (\ldots + \delta(t-L) + \delta(t) + \delta(t+L) + \ldots) \cdot psf\,(x-t)dt = \ldots + \int_{-\infty}^{+\infty} \delta(t-L) \cdot psf\,(x-t)dt +$$

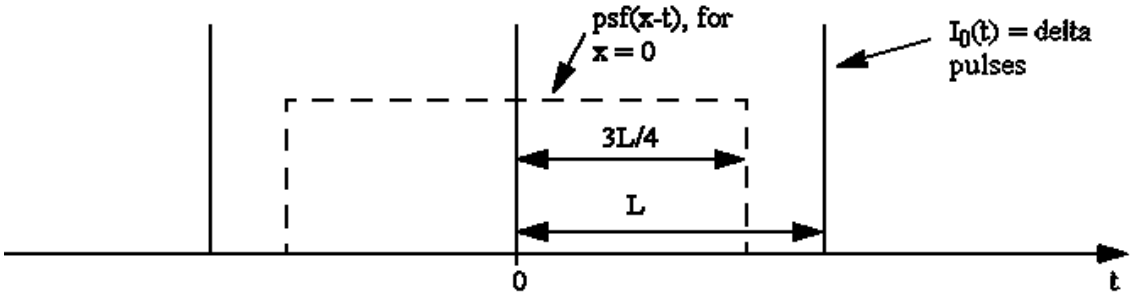
$$+ \int_{-\infty}^{+\infty} \delta(t) \cdot psf\,(x-t)dt + \int_{-\infty}^{+\infty} \delta(t+L) \cdot psf\,(x-t)dt + \ldots = \left\{ \delta(y) \neq 0 \text{ only for } y = 0, \text{ and } \int_{-\infty}^{+\infty} \delta(y)dy = 1 \right\} =$$

$$= \ldots psf\,(x-L) + psf\,(x) + psf\,(x+L) + \ldots$$

We are thus to add an infinite number of displaced copies of the *psf*.



No. 2 (engineering style): We are to multiply $I_0(t)$ by $psf(x-t)$ and integrate over t.



For $x = 0$, as shown in the figure, only the central δ peak is included. If we then integrate over t, the integral will get a value of 1, i.e. $I_B(0) = 1$. As we increase x, the "psf box" will be displaced to the right. As long as x < L/4 nothing changes, i.e. only the central peak is included. But as soon as x > L/4 two δ peaks will be included, and the value of the convolution integral will be 2. By continuing in the same way for increasing (and diminishing) x-values, we get the same result as when using method no. 1.

21) From the given data we can estimate the *MTF*, which is the absolute value of the Fourier transform of the *psf*. From, for example, Beta we get that the FT of the given Gaussian function is $e^{-\frac{\omega^2}{4a}}$ (normalized to 1 for ω = 0, since *MTF* is normalized in this way). The value of the constant a is obtained from the FWHM of the psf: $0.5 = e^{-a\left(7.5\times10^{-6}\right)^2} \Rightarrow a = 1.23\times10^{10}\,\text{m}^{-2}$. We now know the *MTF* function, and can calculate at which spatial frequency its value will be 0.10: $0.10 = e^{-\frac{\omega^2}{4\times1.23\times10^{10}}} \Rightarrow \omega = 3.37\times10^5$. ω is the spatial frequency multiplied by 2π, which gives a spatial frequency of $5.4\times10^4\,\text{m}^{-1}$. Thus the answer is that the *MTF* is down to a value of 0.10 for a spatial frequency of approx. 50 mm^{-1}.

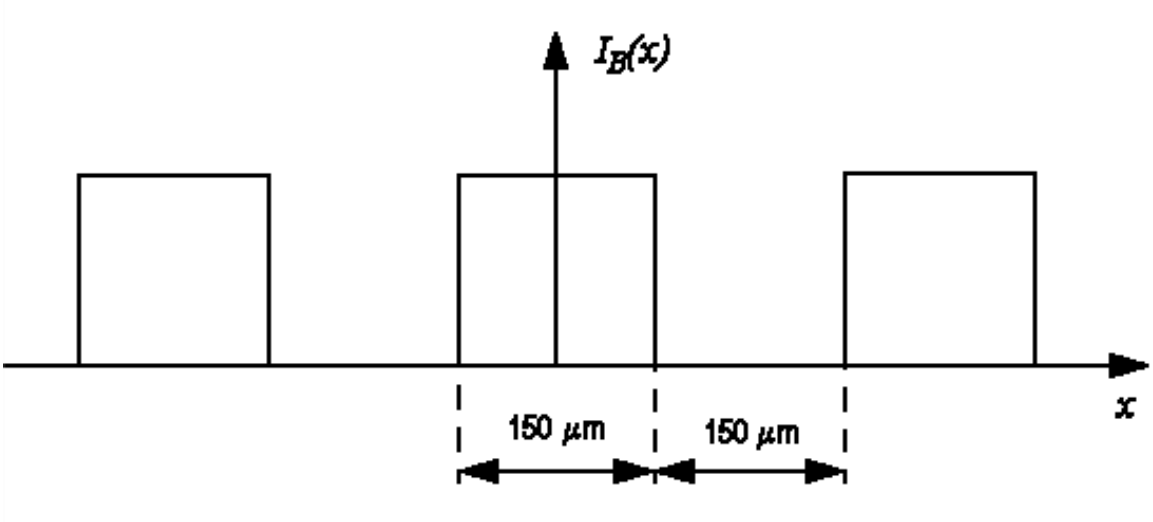
22)

a) We start by calculating $psf_{motion\ blur}$. During the exposure time the airplane has travelled a distance of $s = v_{airplane}\cdot t_{\text{exp}}$. The image in the camera of a point object on the ground is moved a distance of $s' = v_{airplane}\cdot t_{\text{exp}}\cdot\frac{f}{h} = \frac{750}{3.6}\cdot\frac{1}{250}\cdot\frac{0.30}{1000} = 250$ μm. This gives $psf_{\text{motion blur}} = rect\left(\frac{x}{L}\right)$, where L = 250 μm.

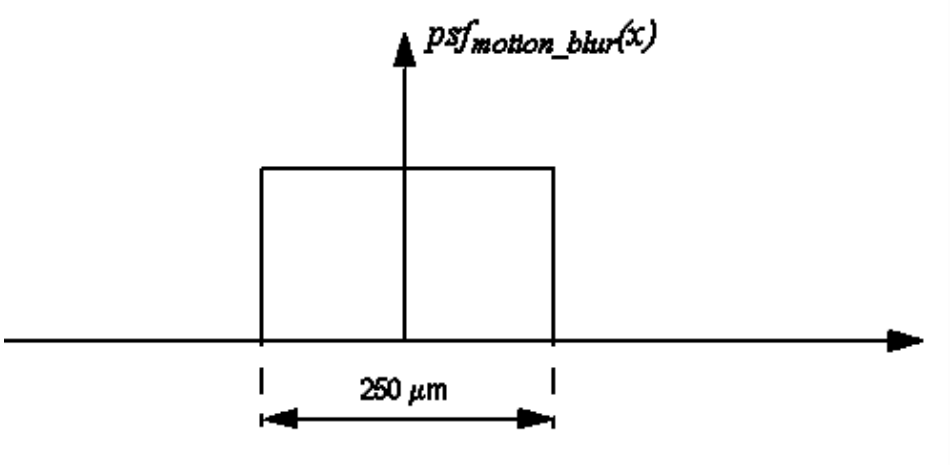
$$MTF_{\text{motion blur}} = FT\left\{rect\left(\frac{x}{L}\right)\right\} = \left|\frac{\sin(\pi\nu L)}{\pi\nu L}\right| = \left|\frac{\sin(\pi\cdot\nu\cdot2.5\times10^{-4})}{\pi\cdot\nu\cdot2.5\times10^{-4}}\right|$$

This function has its first zero at ν = 4.0 mm^{-1} (i.e. the image will look pretty blurred).

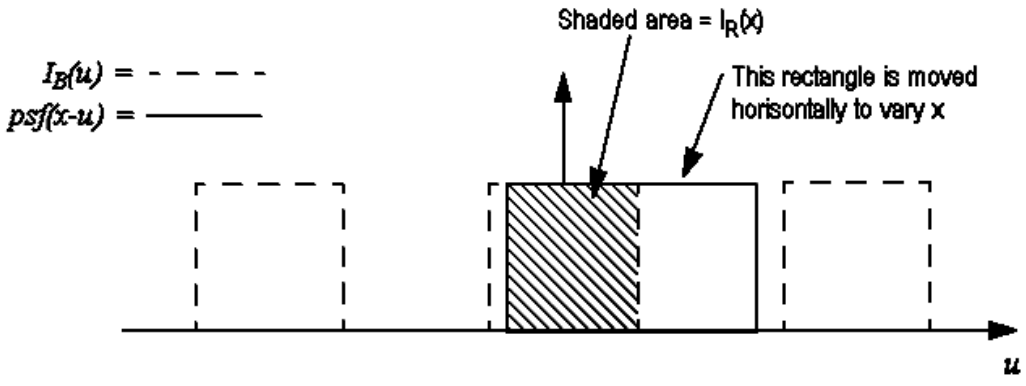
b) The exposure distribution $I_R(x)$ is obtained by the convolution of $I_B(x)$ and $psf_{motion\ blur}(x)$. Since the quality of the objective is high, we assume that I_B is just a demagnified version of the object function. The width of the lines in the image plane will then be $0.50\cdot\frac{f}{h} = 0.50\cdot\frac{0.30}{1000} = 1.5\times10^{-4}$ m = 150 μm. We then get:



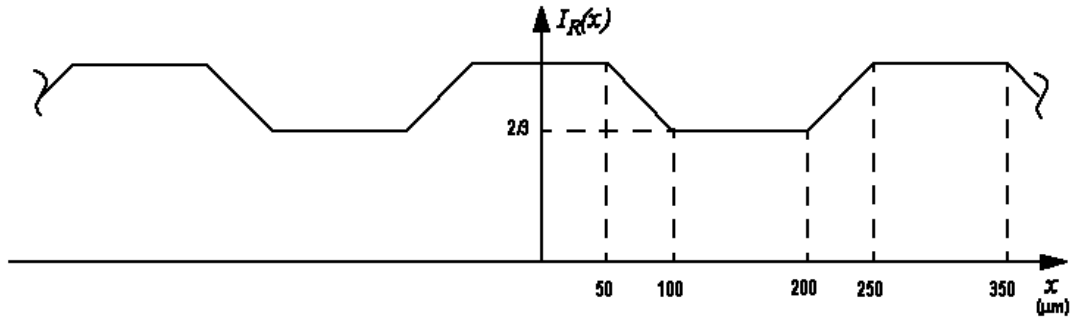
This is to be convolved with the function:



The exposure distribution $I_R(x) = \int_{-\infty}^{+\infty} I_B(u) \cdot psf_{\text{motion blur}}(x-u)du$



By displacing the rectangular function ($psf_{motion\ blur}$), and plotting the size of the shaded area as a function of the displacement, we get the exposure distribution.



c) $v_{film} = v_{airplane} \cdot \dfrac{f}{h} = \dfrac{750}{3.6} \cdot \dfrac{0.30}{1000} = 0.062$ m/s = 62 mm/s.

23)

a) For large object distances a (a fairly good approximation here) the imaging scale is approximately $\dfrac{f}{a}$. This means that spatial frequencies in the image plane will be higher by a factor of $\dfrac{a}{f}$ compared with frequencies in the object (the exact formula is $\dfrac{a}{f} - 1$). For the given values of object distance and focal length, we find that the image modulation has been measured at the following spatial frequencies: 8 (7), 30 (29), 60 (59), 88 (87), 120 (119), 152 (151). All numbers given are in units of mm^{-1}, and the numbers in parenthesis were calculated using the exact formula (not necessary in this problem). Since the object modulation is $\approx 100\,\%$, the modulation values in the table will directly give the *MTF* values. (No plots illustrated here)

b) $MTF_{\text{total}} = MTF_{\text{lens}} \cdot MTF_{\text{detector}}$. The values in the table are MTF_{total}.

$$MTF_{\text{detector}} = \left| \frac{\sin(\pi \nu L)}{\pi \nu L} \right| = \left| \frac{\sin(\pi \nu \cdot 5.0 \times 10^{-6})}{\pi \nu \cdot 5.0 \times 10^{-6}} \right|$$

For the image plane spatial frequencies given in a), we get the following values for MTF_{detector}: 1.00, 0.97, 0.86, 0.72, 0.51, and 0.29. The corresponding values for MTF_{total} should be divided by these values to get MTF_{lens}. This gives the following values for MTF_{lens} at the spatial frequencies given above: 0.90, 0.67, 0.42, 0.22, 0.08, and 0.00.

c) For a diffraction-limited lens, *MTF* drops to zero at the spatial frequency $\dfrac{D}{\lambda f}$, which in this case is $\dfrac{50 \times 10^{-3}}{550 \times 10^{-9} \cdot 0.25} = 360\ mm^{-1}$. From the measurements we see that the *MTF* has dropped to zero before the frequency reaches 150 mm^{-1}. Thus, the lens is far from diffraction-limited.

24)

$MTF_{\text{detector}} = \left|\frac{\sin(\pi\nu L)}{\pi\nu L}\right|$. $MTF_{\text{total}} = MTF_{\text{optics}} \cdot MTF_{\text{detector}}$. For ν = 30 mm^{-1} we get $0.50 = 0.70 \cdot \left|\frac{\sin(\pi\nu L)}{\pi\nu L}\right| \Rightarrow \left|\frac{\sin(\pi\nu L)}{\pi\nu L}\right| = 0.714$. Trial and error tests yield $\pi\nu L$ = 1.37. $L = \frac{1.37}{\pi\nu}$ = (for ν = 30 mm^{-1}) = 15 μm.

25) The highest spatial frequency that can be imaged is $\nu_{\text{limit}} = \frac{D}{\lambda f}$. At a wavelength of 550 nm and an f-number of 5.6 we get $\nu_{\text{limit}} = 3.25 \times 10^5\,\text{m}^{-1}$. The sampling frequency must be twice as high, i.e. $\nu_S = 6.49 \times 10^5\,\text{m}^{-1}$, which gives the density of light-sensitive elements in the area array sensor. For a detector size of 24x36 mm we get 15600 x 23400 elements. (Note: In reality aberrations will limit the highest spatial frequency).

26) The area array sensor has a sampling frequency of $\frac{1}{7.0 \times 10^{-3}} = 143$ mm^{-1}, which means that the Nyquist frequency = 143/2 = 71 mm^{-1} = the highest spatial frequency that can be correctly recorded. For the curtain image not to exceed this limit, the imaging scale, *M*, must satisfy the condition $M > \frac{1}{71}$. We can write $M = \frac{f}{a}$, where *f* = focal length and *a* = distance between camera and curtain. Using this, we get *a* < 1.3 m. <u>But</u> we also avoid aliasing if the lens cannot resolve the curtain pattern. The limiting frequency is 140 mm^{-1}, which gives $M < \frac{1}{140}$, yielding *a* > 2.5 m.

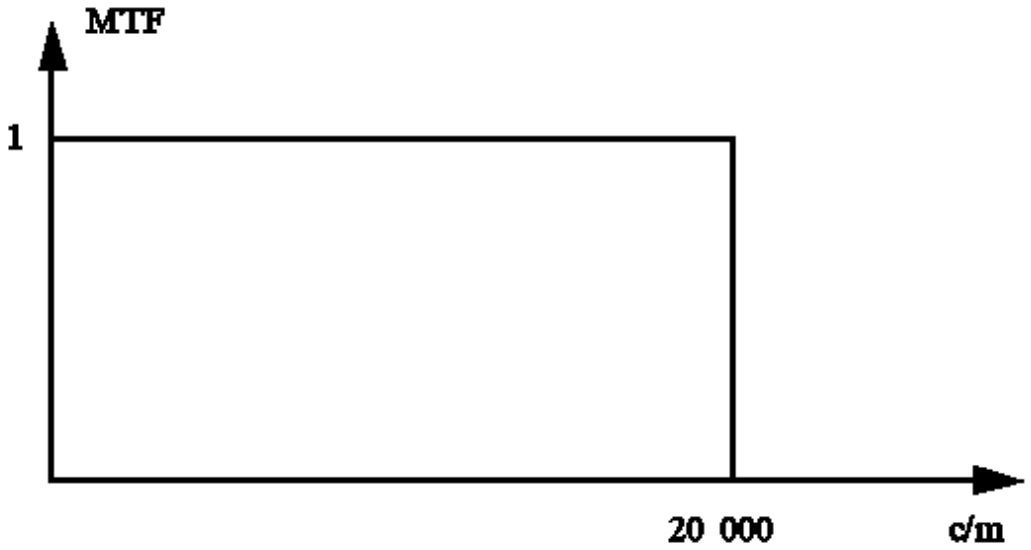
Altogether this means that we can expect to get aliasing for distances between 1.3 and 2.5 m.

27) Plot the original spectrum (peaks at 40, 80 and 130 mm^{-1}), and the corresponding spectra frequency-shifted $\pm\nu_S = \pm 100\,\text{mm}^{-1}$. Where in the region $\pm\frac{\nu_S}{2} = \pm 50\,\text{mm}^{-1}$ do we find frequency peaks? The result is that we find peaks at 20 mm^{-1} (false, should be 80), 30 mm^{-1} (false, should be 130) and 40 mm^{-1} (correct). (We also get peaks at the corresponding negative frequencies, but these correspond to periodic patterns with the same spatial frequencies as those previously mentioned)

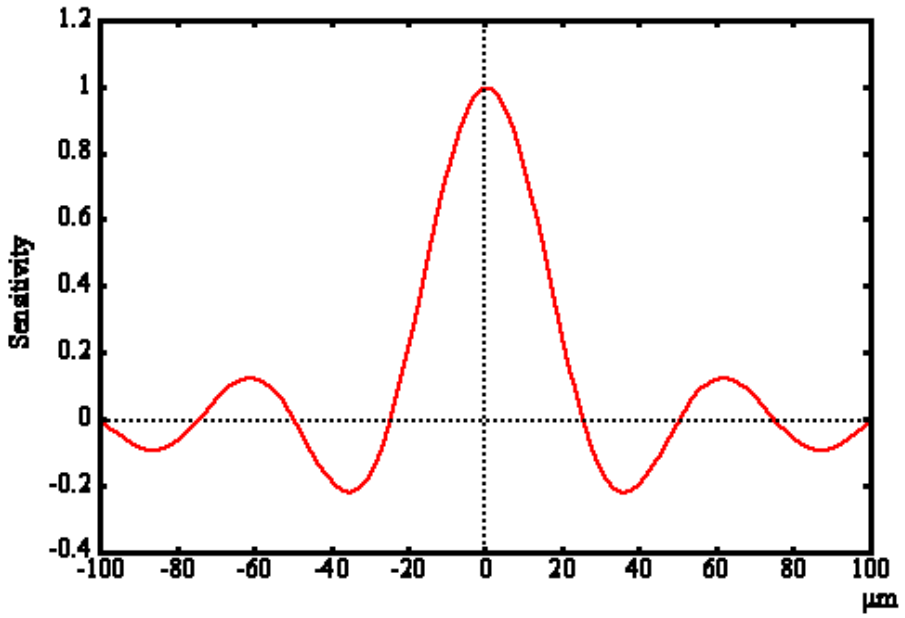
28) a) The center-to-center distance for the detectors is 25 μm, which gives a sampling frequency of $\nu_S = 4.0 \times 10^4\,\text{m}^{-1}$. The highest spatial frequency which can be correctly recorded is then $\nu_{\text{limit}} = 2.0 \times 10^4\,\text{m}^{-1}$.

b) *psf* = *rect*(*x*/*a*), where *a* = 25 μm. *MTF* is the absolute value of the Fourier transform of the *psf*, which gives $MTF(\nu) = \left| \frac{\sin(\pi\nu \cdot 25\times10^{-6})}{\pi\nu \cdot 25\times10^{-6}} \right|$. The *MTF* value at ν_{limit} according to problem a) is 0.64.

c) *MTF* should be such that spatial frequencies up to ν_{limit} are reproduced without loss of modulation (i.e. *MTF* = 1), while frequencies above ν_{limit} are completely suppressed. We then get an *MTF* according to the figure below.



d) The inverse Fourier transform of the *MTF* curve in problem c) yields $psf = \frac{\sin(\pi x \cdot 4\times10^{4})}{\pi x \cdot 4\times10^{4}}$, which should be the sensitivity distribution of the detector. This means that certain parts of the detector have a *negative* sensitivity, see below!



29)

a) $\nu_{\max} = \frac{\nu_S}{2} = \frac{1}{2 \cdot 25 \times 10^{-6}} = 20$ periods/mm. This is true for both x and y directions since the sampling frequency is the same in both directions.

b) $MTF_x = \left| \frac{\sin(18 \times 10^{-6} \cdot \pi \cdot 20 \times 10^3)}{18 \times 10^{-6} \cdot \pi \cdot 20 \times 10^3} \right| = 0.80$

$MTF_y = \left| \frac{\sin(13 \times 10^{-6} \cdot \pi \cdot 20 \times 10^3)}{13 \times 10^{-6} \cdot \pi \cdot 20 \times 10^3} \right| = 0.89$

c) 1) remains constant (the Nyquist frequency is only determined by the center-to-center distance between detector elements)
2) decreases (25 μm instead of 13 and 18 in the sinc-functions in b))
3) increases (more light hits each detector element)
4) increases (more photons means less noise)

30)

a) According to the sampling theorem we have:

$\nu_S \geq 2\nu_{\max} = 2 \cdot \frac{2 \cdot 0.9}{550 \times 10^{-9} \cdot 100} = 6.55 \times 10^4 \text{ m}^{-1}$.

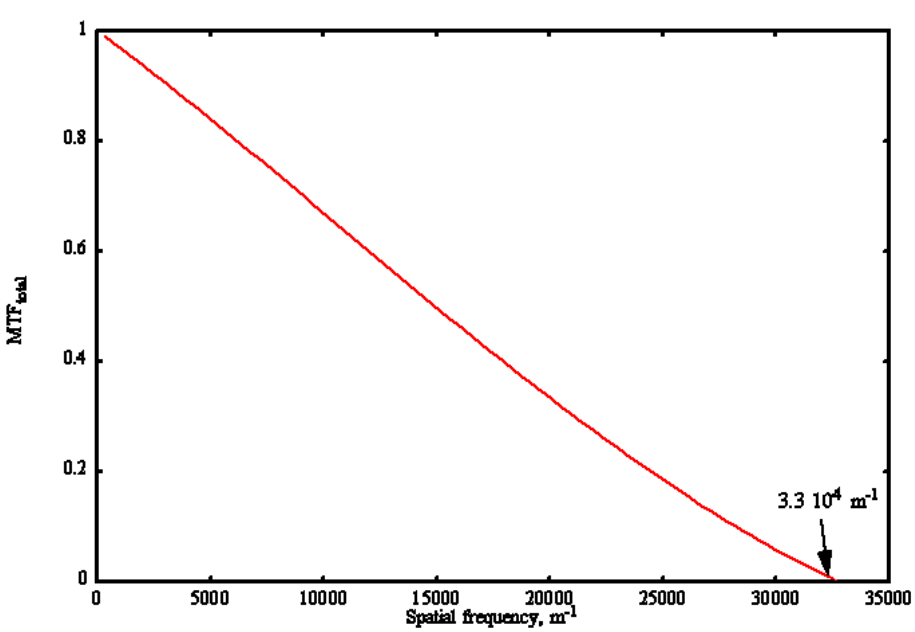
From the detector dimensions we get: $\nu_S = \frac{1}{15 \times 10^{-6}} = 6.67 \times 10^4 \text{ m}^{-1}$.

The sampling theorem is fulfilled.

b) $MTF_{\text{detector}} = \left| \frac{\sin(15 \times 10^{-6} \cdot \pi\nu)}{15 \times 10^{-6} \cdot \pi\nu} \right|$

$MTF_{lens} = 1 - \frac{\lambda M \nu}{2NA}$ up to $\nu = \frac{2NA}{\lambda M}$, above which it is zero.

$MTF_{total} = MTF_{\text{detector}} \cdot MTF_{lens}$, see figure.



31) Sampling frequency $\nu_S = \dfrac{1}{18\times10^{-3}/3600} = 2.0\times10^5\,\text{m}^{-1}$, which gives $\nu_{max} = 1.0\times10^5\,\text{m}^{-1}$.

Set ν_{max} to the maximum spatial frequency that can be reproduced by the lens $\Rightarrow \nu_{max} = \dfrac{D}{\lambda f} \Rightarrow \dfrac{f}{D} = \dfrac{1}{\lambda \nu_{max}} = \dfrac{1}{550\times10^{-9}\cdot 1.0\times10^5} = 18$. Use an f-number approximately half-way between 16 and 22, or use 22 (to be on the safe side).

32)

a) The spatial frequency of the stripes may become so high in the image plane that it exceeds the Nyquist frequency, i.e. we get less than two sampling points per period of the pattern. This will result in incorrect recording of the pattern density (aliasing).

b) The sampling distance in the detector plane is 20 μm, which means that the Nyquist frequency is $\dfrac{1}{2\cdot 20\times10^{-6}} = 25$ mm^{-1}. The spatial frequency of the object pattern is 0.5 mm^{-1}. For an imaging scale of 1:50 we are exactly at the Nyquist frequency. The object distance will then be 50 times the focal length (to be quite accurate the image distance rather than the focal length, but in this case the difference is negligible), which is 1.0 meter. Therefore the maximum distance to the politician is one meter if we are to avoid aliasing.

33)

a) To get equal imaging scales in both directions, the optical image should move 25 μm between read-outs of the sensor, i.e. the same distance as the center-to-center distance between detector elements. The speed with which the optical image moves in the image plane of the lens is given by (satellite speed) x (imaging scale) =

$7500 \cdot \frac{2.0}{8.0 \times 10^5} = 1.88 \times 10^{-2}$ m/s. The time it takes to move the optical image a distance of 25 μm is then given by $t = \frac{25 \times 10^{-6}}{1.88 \times 10^{-2}} = 1.3 \times 10^{-3}$ s.

b) According to the sampling theorem, the sampling frequency should be at least twice the spatial frequency that we want to record. The linear sensor has a sampling frequency of $\frac{1}{25 \times 10^{-6}} = 4.0 \times 10^4$ m^{-1}. The maximum spatial frequency that we can handle in the image is half of this value or 2.0×10^4 m^{-1}. This corresponds to a spatial frequency on the ground of $2.0 \times 10^4 \cdot \frac{2.0}{8.0 \times 10^5} = 5.0 \times 10^{-2}$ m^{-1}, which is equivalent to a period length of 20 meters.

c) For spatial frequencies in a direction parallel to the row of detectors we get a *psf* which is a rectangular function with a width of 15 μm. *MTF* = the modulus of the Fourier transform of this function = $\left| \frac{\sin(\pi\nu \cdot 15 \times 10^{-6})}{\pi\nu \cdot 15 \times 10^{-6}} \right|$. We can assume that the *MTF* of the optics is much better, so its influence can be neglected. Furthermore, we have no motion blur in a direction parallel to the row of detectors, and therefore the above sinc function is the total *MTF*.

For spatial frequencies perpendicular to the row of detectors, we get (in analogy with what was said above) that $MTF_{\text{detector}} = \left| \frac{\sin(\pi\nu \cdot 25 \times 10^{-6})}{\pi\nu \cdot 25 \times 10^{-6}} \right|$. If the exposure time is negligible, the motion blur will also be negligible and $MTF_{\text{total}} = MTF_{\text{detector}}$. If the exposure time is equal to the time between detector read-outs, the image will have moved a distance equal to the detector width, i.e. 25 μm, during the exposure. *psf* for the motion blur will then be a rectangular function with a width of 25 μm, i.e. the same as for the detector. This means that $MTF_{\text{motion blur}} = MTF_{\text{detector}} = \left| \frac{\sin(\pi\nu \cdot 25 \times 10^{-6})}{\pi\nu \cdot 25 \times 10^{-6}} \right|$. $MTF_{\text{total}} = MTF_{\text{detector}} \cdot MTF_{\text{motion blur}} = \left| \frac{\sin(\pi\nu \cdot 25 \times 10^{-6})}{\pi\nu \cdot 25 \times 10^{-6}} \right|^2$.

Appendix 1: Fourier series, Fourier transform and convolution: A brief introduction from a physics viewpoint.

by Kjell Carlsson

The following pages present a brief summary of the mathematics necessary for studies of Imaging Physics. The focus is on physical/technical understanding. As a consequence, the material is presented in a rather simplified way.

Fourier series

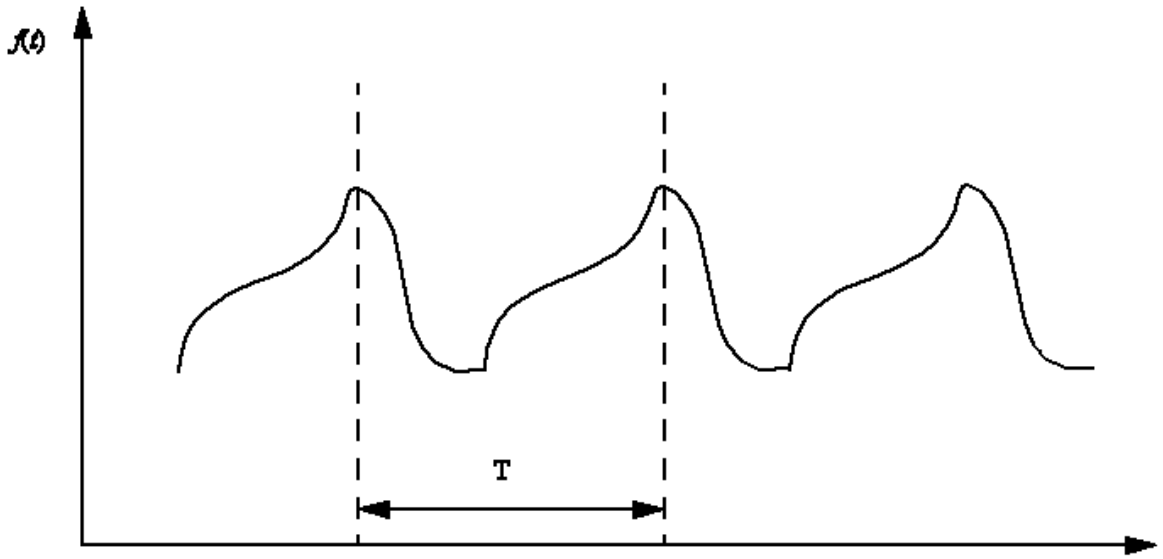


Fig. 1. Periodic function with period (wavelength) T

A periodic function (period = T) can be expressed as a Fourier series:

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty}\left[a_n \cos(n\omega_0 t) + b_n \sin(n\omega_0 t)\right] \quad (1)$$

where $\omega_0 = \frac{2\pi}{T}$, and

$$a_n = \frac{2}{T}\int_{-\frac{T}{2}}^{+\frac{T}{2}} f(t)\cos(n\omega_0 t)dt\,,(n \geq 0) \quad (2)$$

and

$$b_n = \frac{2}{T}\int_{-\frac{T}{2}}^{+\frac{T}{2}} f(t)\sin(n\omega_0 t)dt\,,(n \geq 1) \quad (3)$$

Eq. 1 can be physically interpreted in the following way: The periodic, but non-sinusoidal, function $f(t)$ can be created by adding an infinite number of sine and cosine functions of different frequencies (angular frequencies $n\omega_0$) and different amplitudes (a_n, b_n). The lowest frequency in this sum (apart from the constant $a_0/2$ in eq. 1) is equal to $1/T$, where T is the period of $f(t)$. The second lowest frequency is $2/T$, and so on. $1/T$ is called the fundamental (i.e. lowest) frequency and the higher frequencies (n/T) are called harmonics. In a practical situation, it is usually impossible to add an infinite number of sine and cosine functions, but the more harmonics we add the more closely our sum will approach $f(t)$. An example of this is given in the figures below for a square-wave function.

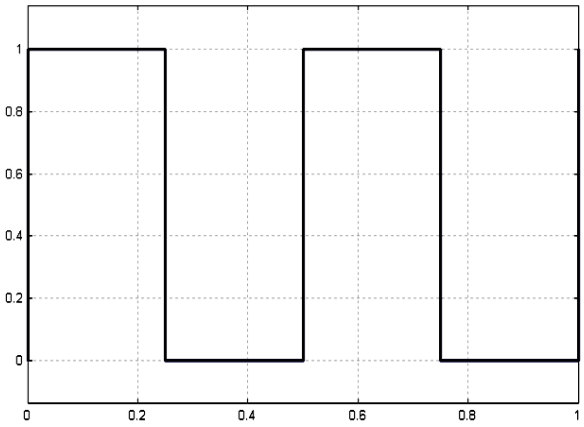


Fig. 2. Square-wave function

Fig. 3. Fundamental frequency only

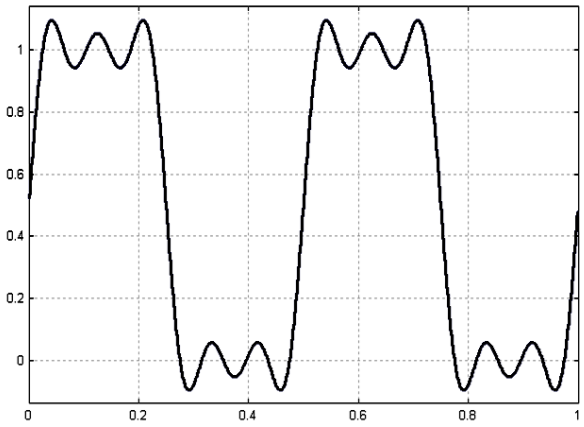


Fig. 4. Fundamental plus two harmonics

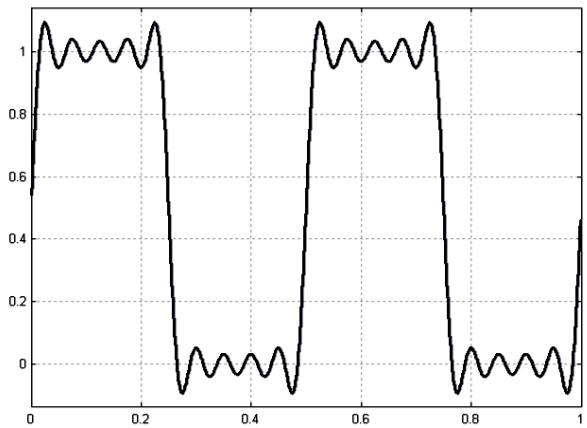


Fig. 5. Fundamental plus four harmonics

Instead of using eq. 1 to express the Fourier series for $f(t)$, we can express it in the following way:

$$f(t) = \sum_{n=-\infty}^{+\infty} c_n e^{in\omega_0 t} \tag{4}$$

where

$$c_n = \frac{1}{T} \int_{-\frac{T}{2}}^{+\frac{T}{2}} f(t) e^{-in\omega_0 t} dt \tag{5}$$

It can be shown that eqs. 4 and 5 express mathematically exactly the same information as eqs. 1-3. Furthermore, one set of equations can be transformed into the other as shown, for example, in "Mathematics Handbook, Beta." Thus, c_n in eq. 5 contains the same amplitude information as a_n and b_n in eqs. 2 & 3, but written in a different form.

We will now make the following changes in order to adapt the nomenclature to a more physical/technical viewpoint, and also in order to prepare for what is coming in the next section, namely Fourier transforms:

1) From now on we will change ω_0 (which is equal to $\frac{2\pi}{T}$) to $\Delta\omega$. This is just a change of "name," but it also reflects the fact that is the angular frequency separation between the harmonics.

2) The angular frequencies of the harmonics are given by $n\omega_0$. We will now introduce a general symbol, ω, for angular frequency. Thus, the different harmonics have ω values of $2\omega_0$, $3\omega_0$, $4\omega_0$, etc. (i.e. $\omega = n\omega_0$).

3) Let us also introduce the quantity $c(\omega)$, defined by $c(\omega) = c(n\omega_0) = \frac{c_n}{\Delta\omega}$, where c_n is the coefficient in eq. 5. As mentioned above, c_n contains amplitude information about the n:th harmonic. Therefore, $c(\omega)$ gives information concerning how much amplitude $f(t)$ has per frequency interval (we will elaborate on this later).

By introducing the above changes 1-3, we can re-write eq. 5 as follows:

$$c(\omega) = \frac{1}{2\pi}\int_{-\frac{T}{2}}^{+\frac{T}{2}} f(t)e^{-i\omega t}dt \qquad (6)$$

Note that eq. 6 contains exactly the same information as eq. 5. The only difference is that we have changed some of the symbols according to points 1-3 above. The stage is now set for Fourier transforms.

Fourier transform

Now, let us consider an arbitrary function $f(t)$ that need not be periodic. One way of doing this is to let $T \to \infty$ in Fig. 1. As a result $\Delta\omega \to 0$, and it will henceforth be denoted $d\omega$. This in turn means that ω will become a continuous variable that can assume any real value. $c(\omega)$, as described by eq. 6, now becomes

$$c(\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} f(t)e^{-i\omega t}dt \qquad (7)$$

Looking at, for example, "Mathematics Handbook, Beta," we see that the Fourier transform of $f(t)$, denoted by $\hat{f}(\omega)$, is given by

$$\hat{f}(\omega) = \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt \tag{8}$$

apart from the constant $\frac{1}{2\pi}$, $c(\omega)$ is identical to $\hat{f}(\omega)$. We can therefore conclude (see point 3 in previous section) that the Fourier transform of $f(t)$ gives information about the amplitude per frequency interval of the sine and cosine oscillations that, when added together, form the function $f(t)$. This is explained in more detail below, as well as in a more "down to earth" manner in section 10, "The Physical Interpretation of a Fourier Transform."

Using the same approach, we can re-write eq. 4 as follows when $T \to \infty$

$$f(t) = \lim_{T \to \infty} \left(\sum_{n=-\infty}^{+\infty} c_n e^{in\omega_0 t} \right) = \lim_{T \to \infty} \left(\sum_{n=-\infty}^{+\infty} c(\omega) e^{i\omega t} \Delta\omega \right) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(\omega) e^{i\omega t} d\omega \tag{9}$$

Comparing again with, for example, "Beta," we see that $f(t)$ is the inverse Fourier transform of $\hat{f}(\omega)$. It is interesting to note that the inverse Fourier transform is mathematically identical to the Fourier transform, apart from the sign in the exponent of the integrand and the factor $1/2\pi$. This means that by calculating the Fourier transform of the Fourier transform of $f(t)$, we get $2\pi f(-t)$. For a symmetric function, this only differs from $f(t)$ by a scaling factor.

Eq. 1 could be interpreted as a superposition of harmonic functions to form an arbitrary periodic function $f(t)$. This must be the case also for eq. 4, which contains the same information as eq. 1, but in a slightly re-written form. In eq. 9 we see that in the limit where $T \to \infty$, eq. 4 becomes the inverse Fourier transform of $\hat{f}(\omega)$. We therefore conclude that the inverse Fourier transform describes how an arbitrary function, $f(t)$, can be described as the superposition of (in the general case) an infinite number of harmonic oscillations (sine/cosine) whose frequencies (in the general case) continuously cover the frequency region from 0 to ∞. In other words, this corresponds to a Fourier series where the frequency difference between the harmonics has become vanishingly small. This should come as no surprise, because we have already seen that the frequency difference between consecutive harmonics is $1/T$, which of course approaches 0 as $T \to \infty$. A simplified explanation for this is also given in section 10 of this compendium.

Using this interpretation of eq. 9, we see that the contribution to $f(t)$ from harmonic functions in the infinitely small frequency interval $\omega \pm \frac{d\omega}{2}$ is given by

$$\hat{f}(\omega) e^{i\omega t} d\omega + \hat{f}(-\omega) e^{-i\omega t} d\omega = \left[\hat{f}(\omega) e^{i\omega t} + \hat{f}(-\omega) e^{-i\omega t} \right] d\omega \tag{10}$$

For simplicity we have omitted the constant $1/2\pi$, which is just a scaling factor. If we limit our discussion to the case of a real function $f(t)$ (which is quite sufficient when dealing with image intensity data), it follows that

$$\hat{f}(-\omega) = \hat{f}^{*}(\omega) \tag{11}$$

where * denotes the complex conjugate. From eq. 11 we can infer that the values of the Fourier transform for negative ω values follow automatically if we know the transform values for positive ω values. All information concerning the Fourier transform can thus be obtained from either the positive or the negative part of the ω axis.

Let us now return to eq. 10, which gives the contribution to $f(t)$ from harmonic functions in the infinitely small frequency interval $\omega \pm \frac{d\omega}{2}$.

$$\begin{aligned}
&\left(\hat{f}(\omega)e^{i\omega t} + \hat{f}(-\omega)e^{-i\omega t}\right)d\omega = \left\{\hat{f}(\omega) = \left|\hat{f}(\omega)\right|e^{i\arg\left(\hat{f}(\omega)\right)}\right\} = \\
&\left(\left|\hat{f}(\omega)\right|e^{i\left(\omega t + \arg\left(\hat{f}(\omega)\right)\right)} + \left|\hat{f}(\omega)\right|e^{-i\left(\omega t + \arg\left(\hat{f}(\omega)\right)\right)}\right)d\omega = \\
&= 2\left|\hat{f}(\omega)\right|\cos(\omega t + \arg(\hat{f}(\omega)))d\omega
\end{aligned} \tag{12}$$

The result is a cosine function with amplitude $2\left|\hat{f}(\omega)\right|d\omega$, and phase angle $\arg(\hat{f}(\omega))$. This gives a very "down-to-earth" physical interpretation of the Fourier transform. The absolute value of the transform, $\left|\hat{f}(\omega)\right|$, tells us, for different frequencies, how much amplitude per frequency interval we have for the function $f(t)$. This is often referred to as the frequency spectrum of $f(t)$. The argument for the transform, $\arg(\hat{f}(\omega))$, is the phase angle for the cosine function with angular frequency ω. This topic is discussed in "popular" form in section 10 of this compendium.

Convolution

Let f and g be functions of one variable. We then define the convolution of these functions, $f \otimes g(t)$, by

$$f \otimes g(t) = \int_{-\infty}^{+\infty} f(t-\tau)g(\tau)d\tau \tag{13}$$

The forming of an image by a lens, for example, can be described as a convolution which is illustrated in other parts of this compendium. Different symbols can be used to denote convolution. In these pages we use $\otimes$, which is also used in other parts of this compendium. In many mathematical tables, however, the symbol * is used.

A very important and useful relationship between convolution and Fourier transform is the following

$$FT\{f(t)\cdot g(t)\}=\frac{1}{2\pi}\,\hat{f}\otimes\hat{g}(\omega) \tag{14}$$

$$FT\{f\otimes g(t)\}=\hat{f}(\omega)\cdot\hat{g}(\omega) \tag{15}$$

where *FT* denotes the Fourier transform of the function within brackets. These relationships are used extensively in imaging physics.

The delta function

The delta function, denoted by δ, is an important concept in connection with Fourier transforms (strictly speaking, it is not a mathematical "function," but we will ignore this in these pages). $\delta(t)$ is a function whose value is zero everywhere except for $t = 0$. At $t = 0$, on the other hand, its value is infinite. Thus, the function consists of only a single, infinitely high "spike" at $t = 0$. This sounds like a rather bizarre behavior, but the function is well-behaved in the sense that it has an area of unity, that is

$$\int_{-a}^{+a}\delta(t)dt=1 \tag{16}$$

where a is an arbitrary real number > 0. It is, of course, impossible to plot $\delta(t)$, but a symbolic representation is shown in Fig. 6.

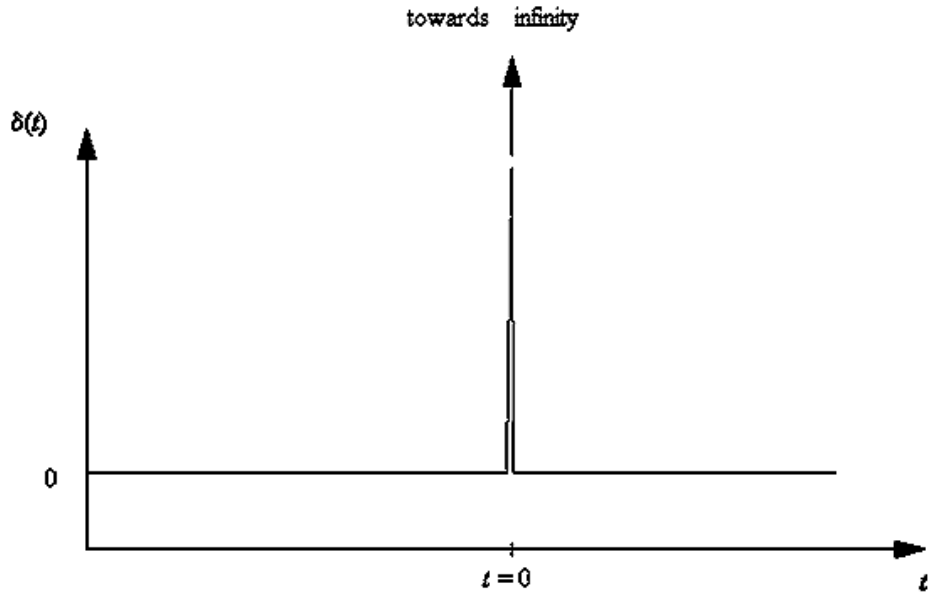


Fig. 6. Symbolic representation of a δ pulse.

In physics $\delta(t)$ is often used as a mathematical model for point-like objects, for example a point source of light (such as a star seen in the sky). It is healthy to realize, however, that it is just a

model. A star is a pretty big lump of matter, but when viewed from a distance of a few light years in a telescope the result is very close to what it would be for a point source. Thus, the delta function is often used as an *approximation* of the real light source, and it is usually much simpler to treat mathematically. To exemplify this simplicity, we can note that the Fourier transform of $\delta(t)$ is the constant 1, and the Fourier transform of the constant 1 is $2\pi\delta(\omega)$. Can't be much simpler, can it?

Fourier transform, convolution and δ function in two dimensions

The mathematical expressions for Fourier transform and convolution (eqs. 8 & 13) can easily be extended to two (or more) dimensions. The same is true for the δ function. Thus, the two-dimensional Fourier transform of a function of two variables, $f(x,y)$, is given by

$$\hat{f}(\omega_x,\omega_y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)e^{-i(\omega_x x+\omega_y y)}dxdy \tag{17}$$

Two-dimensional convolution is given by

$$f \otimes g(x,y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x-u,y-v)g(u,v)dudv \tag{18}$$

The physical interpretation of eqs. 17 & 18 is discussed in other sections of this compendium. It will also be discussed during student laboratory exercises and lectures.

The two-dimensional δ function is denoted by $\delta(x,y)$. It is a function whose value is zero everywhere except for $x = 0$ and $y = 0$, where its value is infinite. By integrating $\delta(x,y)$ in two dimensions we get

$$\int_{-a}^{+a}\int_{-a}^{+a}\delta(x,y)dxdy = 1 \tag{19}$$

where a is an arbitrary real number > 0. A symbolic representation of $\delta(x,y)$ is shown in Fig. 7.

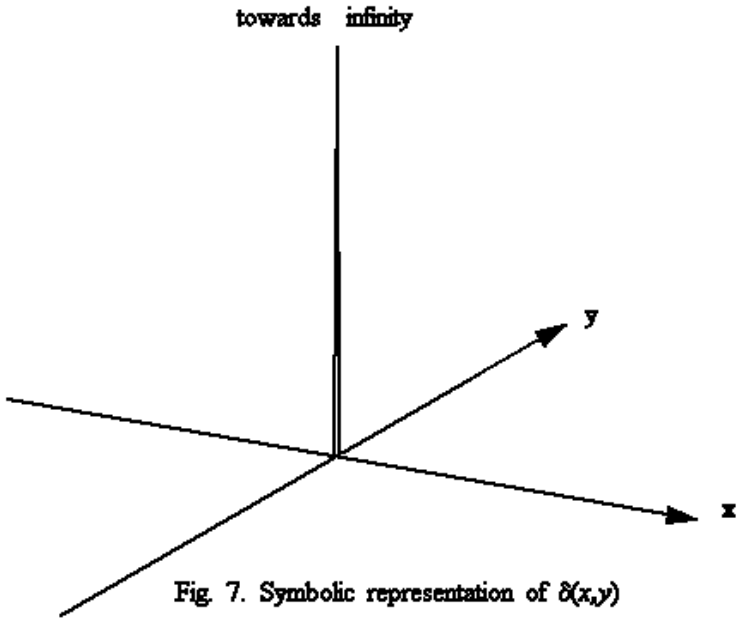


Fig. 7. Symbolic representation of $\delta(x,y)$

Examples of calculation of Fourier transforms and convolutions

The aim of these pages is not to make the student fluent in solving Fourier transforms and convolutions, but rather to define these concepts and to give a physical interpretation. Nevertheless, a few examples are given below to illustrate how such problems can be solved. Note that many mathematical handbooks, for example "Beta," include tables of Fourier transforms for many functions.

Problem 1. Calculate the Fourier transform of $f(t) = \cos(\omega_0 t)$.

Solution and comments.

$$\hat{f}(\omega) = \int_{-\infty}^{+\infty} \cos(\omega_0 t) e^{-i\omega t} dt = \int_{-\infty}^{+\infty} \frac{(e^{i\omega_0 t} + e^{-i\omega_0 t})}{2} e^{-i\omega t} dt =$$

$$= \frac{1}{2}\int_{-\infty}^{+\infty} e^{-i(\omega-\omega_0)t} dt + \frac{1}{2}\int_{-\infty}^{+\infty} e^{-i(\omega+\omega_0)t} dt = \{\text{the FT of 1 is } 2\pi\delta(\omega)\} =$$

$$= \frac{1}{2}\cdot 2\pi\cdot\delta(\omega-\omega_0) + \frac{1}{2}\cdot 2\pi\cdot\delta(\omega+\omega_0) = \pi[\delta(\omega-\omega_0) + \delta(\omega+\omega_0)]$$

$\hat{f}(\omega)$ is a sum of two delta functions, whose spikes are located at $\omega = \omega_0$ and $\omega = -\omega_0$ respectively. This agrees with our previous finding that, for a real function, the positive and negative ω-axes of the Fourier transform contain the same information. It is therefore sufficient to study the positive axis, along which we have a single delta function centered on $\omega = \omega_0$. This means that $f(t)$ contains only one single frequency, ω_0, which is natural considering that $f(t) = \cos(\omega_0 t)$. The fact that $\hat{f}(\omega)$ has an infinite value at $\omega = \omega_0$ is natural, since we have seen that

$\hat{f}(\omega)$ represents how much amplitude we have *per frequency interval*. Since we have only a single frequency, $\hat{f}(\omega)$ should become infinite. The total amplitude is given by integrating over the delta spike, which gives a finite value. Finally it can be mentioned that *f*(*t*) in this case is a pure AC signal (i.e. its average value over time is zero). As a result its FT does not display any delta function at the origin ω = 0. It can easily be shown that by adding a constant to *f*(*t*), a delta function will appear at ω = 0 in the FT. It is true for all functions *f*(*t*) that the area of the spike at ω = 0 represents the average value of the function over *t*.

Problem 2. Calculate the Fourier transform of a function *f*(*t*) according to the figure below

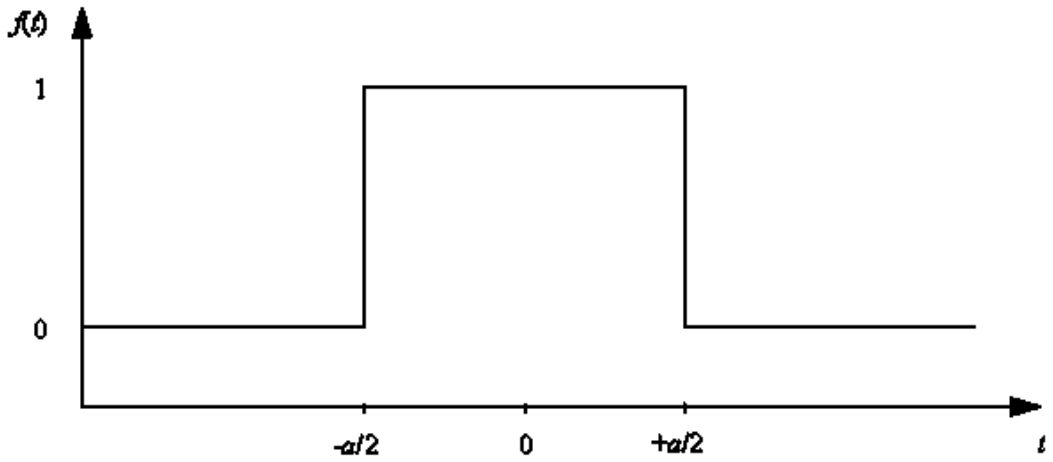


Fig. 8.

This so called rectangular (rect) function plays a very important role in imaging physics (for example when describing the influence of the detector size on image quality). A rectangular function of width *a* (as in the figure above) is often denoted by $rect\left(\frac{t}{a}\right)$.

Solution

$$\hat{f}(\omega)=\int_{-\infty}^{+\infty} f(t)e^{-i\omega t}dt=\int_{-\frac{a}{2}}^{+\frac{a}{2}} e^{-i\omega t}dt=\left[-\frac{e^{-i\omega t}}{i\omega}\right]_{-\frac{a}{2}}^{+\frac{a}{2}}=\frac{e^{\frac{i\omega a}{2}}-e^{-\frac{i\omega a}{2}}}{i\omega}=\frac{2}{\omega}\sin\left(\frac{\omega a}{2}\right)=a\ \mathrm{sinc}\left(\frac{\omega a}{2}\right)$$

where the sinc function is defined by $\mathrm{sinc}(x)=\frac{\sin(x)}{x}$. The sinc function also plays an important role in imaging physics.

Problem 3. Calculate the convolution $f\otimes g(t)$, where $f(\tau)=rect\left(\frac{\tau}{a}\right)$ and $g(\tau)=rect\left(\frac{\tau}{b}\right)$. Assume that $a > b$.

Solution Experience has shown that the convolution integral is a rather difficult concept for students to grasp. We will therefore present a rather detailed solution with many illustrations.

$f \otimes g(t) = \int_{-\infty}^{+\infty} f(t-\tau)g(\tau)d\tau$. The functions $f(\tau)$, $g(\tau)$ and $f(t-\tau)$ are illustrated in the figures below.

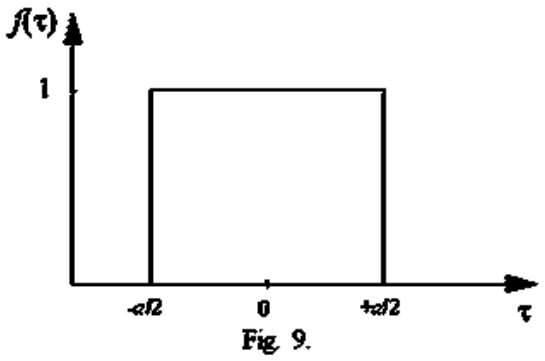


Fig. 9.

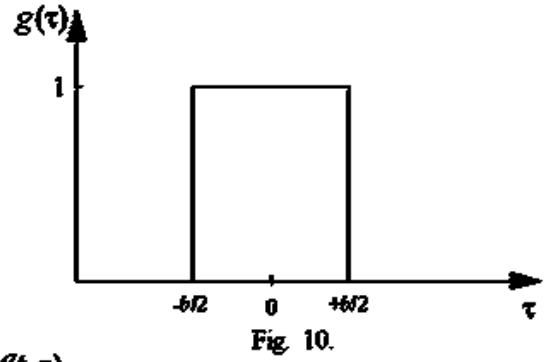


Fig. 10.

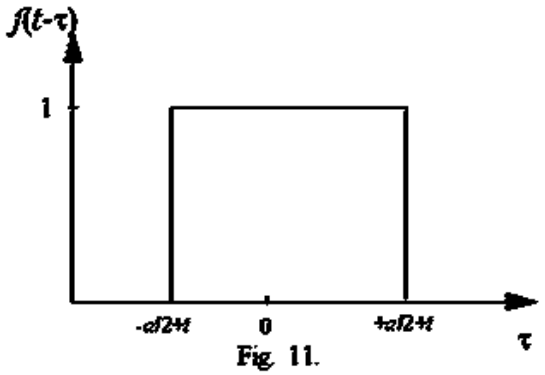


Fig. 11.

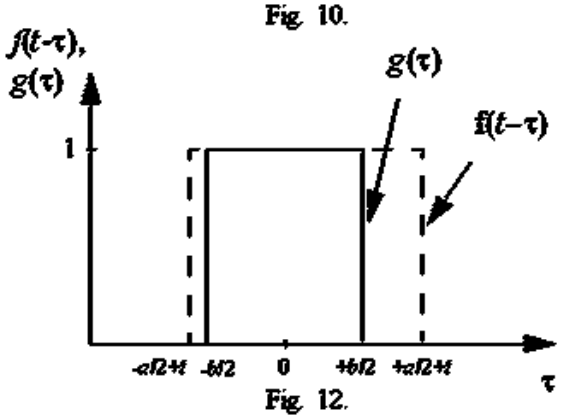


Fig. 12.

Note: $f(\tau)$ is a symmetric function (i.e. $f(-\tau) = f(\tau)$) centered on $\tau = 0$. This means that $f(t-\tau)$ is symmetric around the point $\tau = t$ (i.e. where $t - \tau = 0$).

Let's first assume that $t \geq 0$. From Fig. 12 we conclude that as long as $-\frac{a}{2} + t < -\frac{b}{2}$, and thus $t < \frac{a-b}{2}$, the convolution can be written $\int_{-\infty}^{+\infty} f(t-\tau)g(\tau)d\tau = \int_{-\frac{b}{2}}^{+\frac{b}{2}} 1 \cdot d\tau = b$.

For the interval $\frac{a-b}{2} \leq t \leq \frac{a+b}{2}$ the convolution becomes $\int_{-\frac{a}{2}+t}^{+\frac{b}{2}} 1 \cdot d\tau = \frac{a+b}{2} - t$.

For $t > \frac{a+b}{2}$ $f(t-\tau)$ and $g(\tau)$ do not overlap at all, and therefore the convolution integral is 0.

Because of symmetry in $f(\tau)$ and $g(\tau)$ it follows that $f \otimes g(-t) = f \otimes g(t)$. Below is a plot of $f \otimes g(t)$.

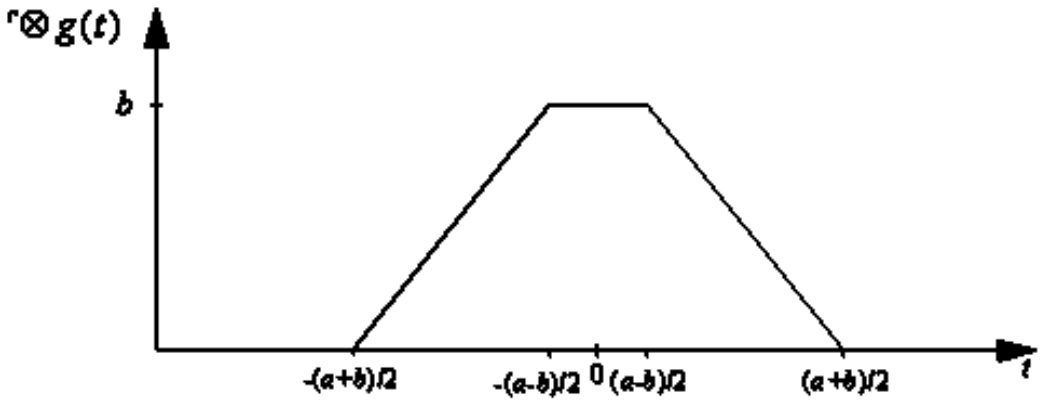


Fig. 13.

Problem 4. Calculate $f \otimes g(t)$, where $f(\tau) = 1$ for $-1 \le \tau \le +1$ and 0 elsewhere.
$g(\tau) = 1 - |\tau|$ for $-1 \le \tau \le +1$ and 0 elsewhere

Solution

Below is a plot of $f(t-\tau)$ and $g(\tau)$.

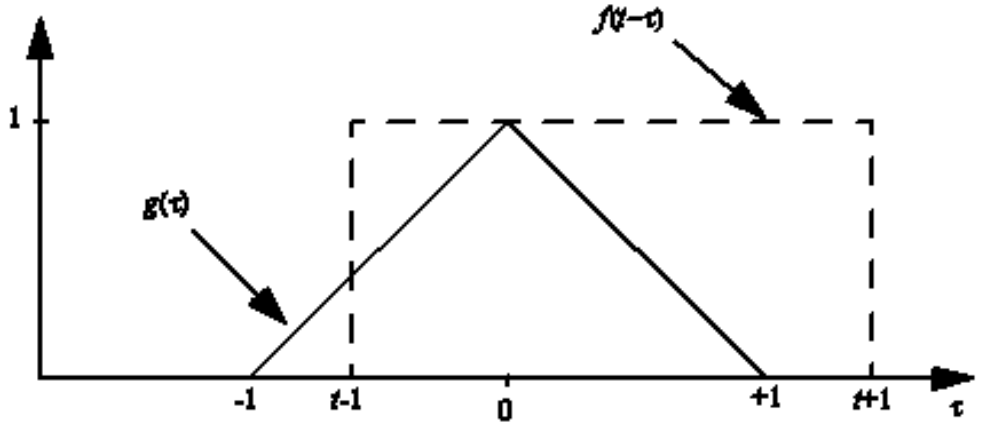


Fig. 14.

For $0 \le t \le 1$ the convolution integral becomes

$$\int_{t-1}^{0} (1+\tau)d\tau + \int_{0}^{1} (1-\tau)d\tau = 1 - \frac{t^2}{2}$$

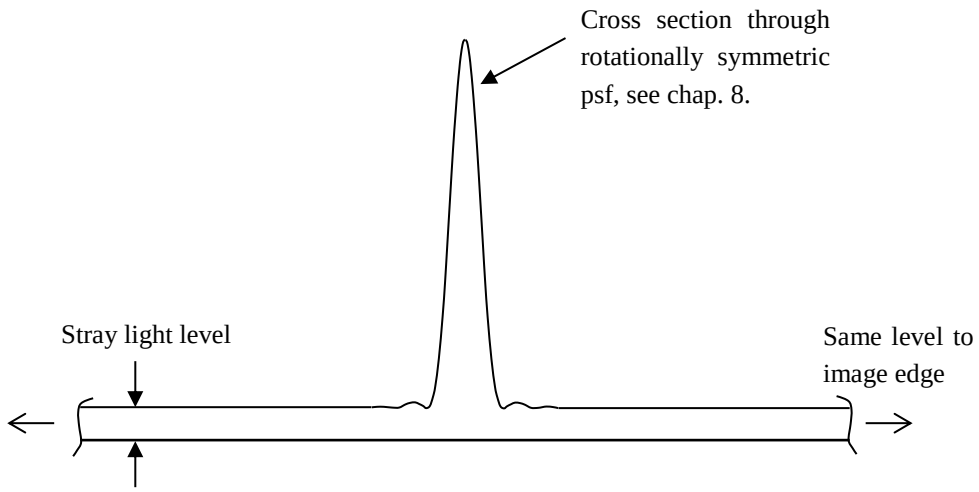
For $1 < t \le 2$ the convolution integral becomes

$$\int_{t-1}^{1} (1-\tau)d\tau = 2 - 2t + \frac{t^2}{2}$$

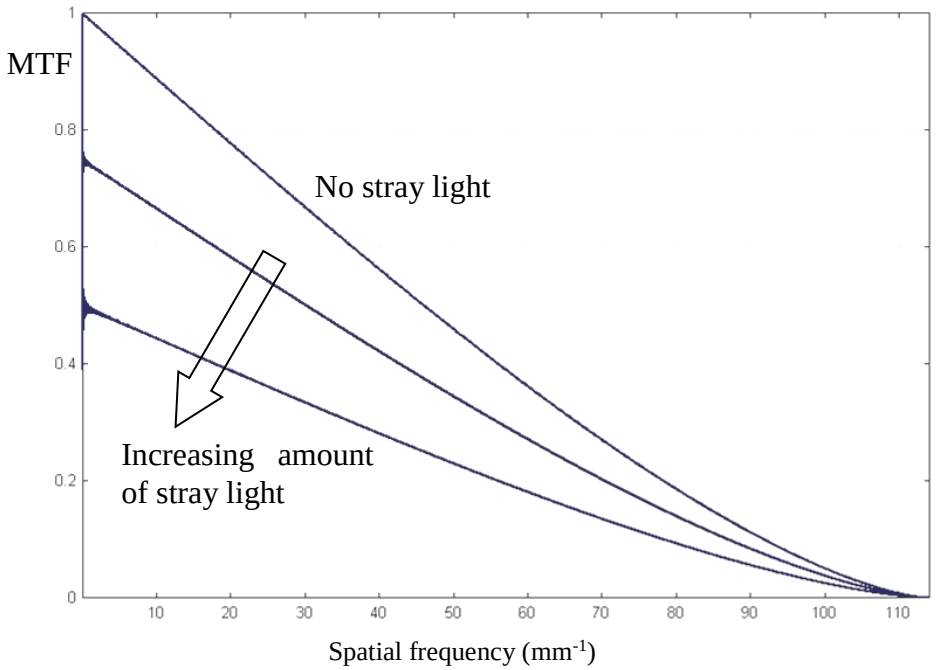
For $t > 2$ the convolution integral becomes 0, because the functions $f(t-\tau)$ and $g(\tau)$ do not overlap. For symmetry reasons $f \otimes g(-t) = f \otimes g(t)$.

Appendix 2: Influence of stray light on *psf* and *MTF*

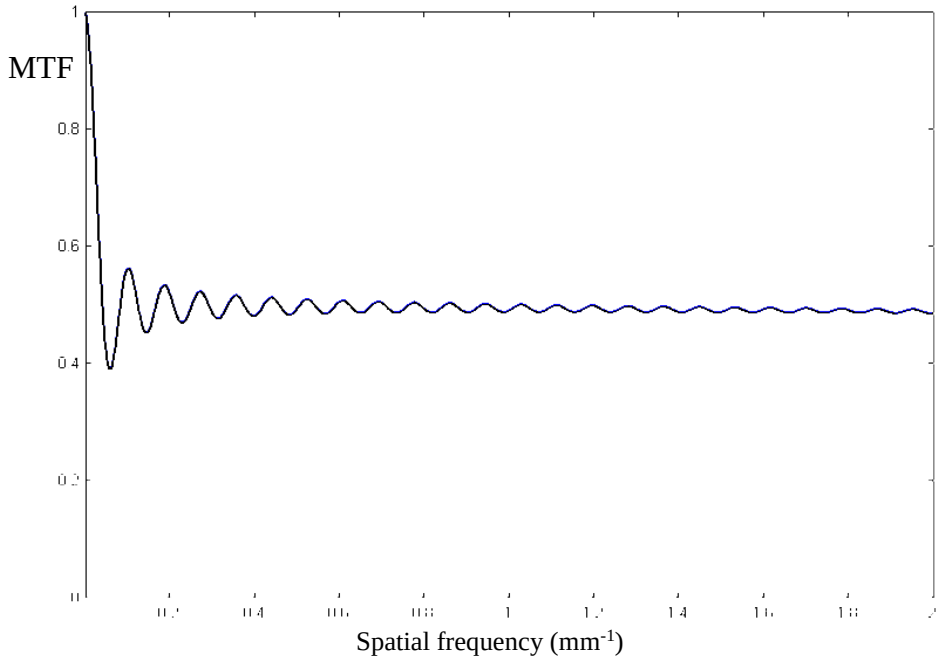
The *psf* of an imaging system will be influenced not only by diffraction and aberrations, but also by stray light. The name stray light indicates that the light somehow turns up in the wrong place. Instead of forming a sharp and nice image, it is spread out more or less (usually less) uniformly over the image area. One source of stray light is reflection and scattering of light from lens surfaces. Such stray light is present to some extent even if high-quality antireflective coatings are used. Stray light can also be produced by oblique incoming light hitting the inside of the objective barrel or the camera housing, and is then reflected more or less diffusively onto the sensor. Depending on how the stray light is produced the effects on image quality can be quite different, and analysis of stray light is usually complicated. In this appendix we will only look at two simple cases that illustrate what kind of effects on the imaging properties we can expect in the presence of stray light. Let us start with a case where stray light produces a uniform illumination over the whole sensor area. This means we just add a constant to the intensity distribution we would get without stray light. Assuming diffraction-limited optics we then get a *psf* according to the figure below.



The psf in this case will be the sum of a diffraction-limited *psf* and a rectangular function extending over the entire sensor area. To obtain the *MTF*, we take the 2-dimensional Fourier transform of the *psf*. The result is a sum of an ordinary diffraction-limited *MTF*, illustrated in chapter 13, and a 2-dimensional sinc function (a function of the type $\frac{\sin x}{x}$, compare chapter 15 where such a function is described). To investigate the practical implication of this, we can consider a case where we have a nearly diffraction-limited photographic lens used at F-number 16 (F-number is the focal length divided by the lens diameter). The sensor size is assumed to be 24 mm x 24 mm. In the figures on next page we can see how the *MTF* curve changes as the level of stray light increases.



All *MTF* curves have a value of 1 at spatial frequency 0, but as the spatial frequency increases the *MTF* will drop very quickly to a lower value in the presence of stray light. Below is a detail of the lower curve showing this behavior (note the difference in horizontal scale).

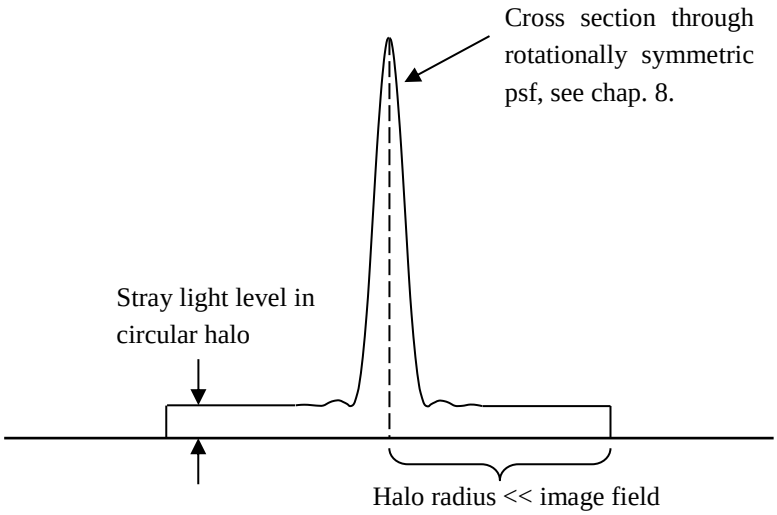


The period length of the oscillations in the *MTF* curve is inversely proportional to the total sensor width, in this case 24 mm. Thus, if the image area is increased, the oscillations will die out more quickly. The oscillations can therefore be regarded as a truncation effect caused by the finite size of the image field recorded. In the limit as the image field tends toward infinity

the oscillations will be infinitely compressed. We then get a single point of the *MTF* curve with a value of 1 at frequency zero, and then an abrupt step down to a lower level.

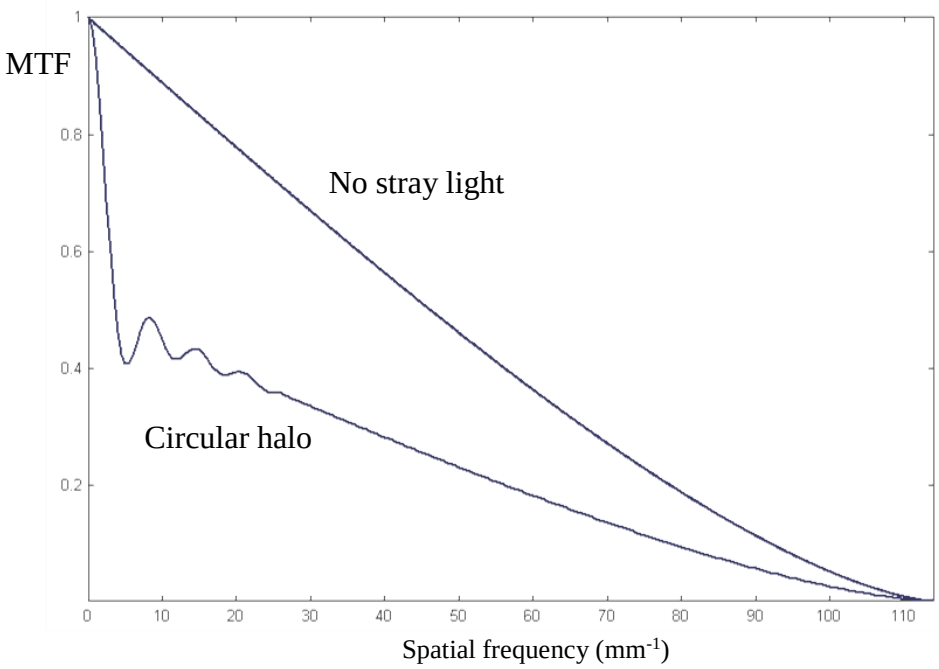
As seen from the *MTF* curves, stray light can have a strongly detrimental effect on *MTF*. But we can easily compensate for this if image recording is done electronically (and not by photographic film). By subtracting a constant value, corresponding to the stray light level, from the recorded images, we can restore the *MTF* to the "No stray light" case. But we have to pay a price in terms of increased noise and reduced bit-depth (gray scale resolution). So it's always better to reduce stray light at the source, rather than by post-processing.

Let's now look at a more realistic case of stray light. We will assume that the stray light contribution is in the form of a circular halo around a diffraction-limited *psf*. This is illustrated in the figure below. This kind of stray light may be produced by scattering from optical coatings on lens elements or filters. The outer edge of the halo can be rather sharp, but not as sharp as in the figure - we are looking at a simplified case. The diameter of the halo depends on the scattering properties of the surfaces, and in the calculations we will assume a halo radius that is 15 times larger than the Airy radius of the optics. This means that the stray light is spread over a much smaller area than in the previous example.

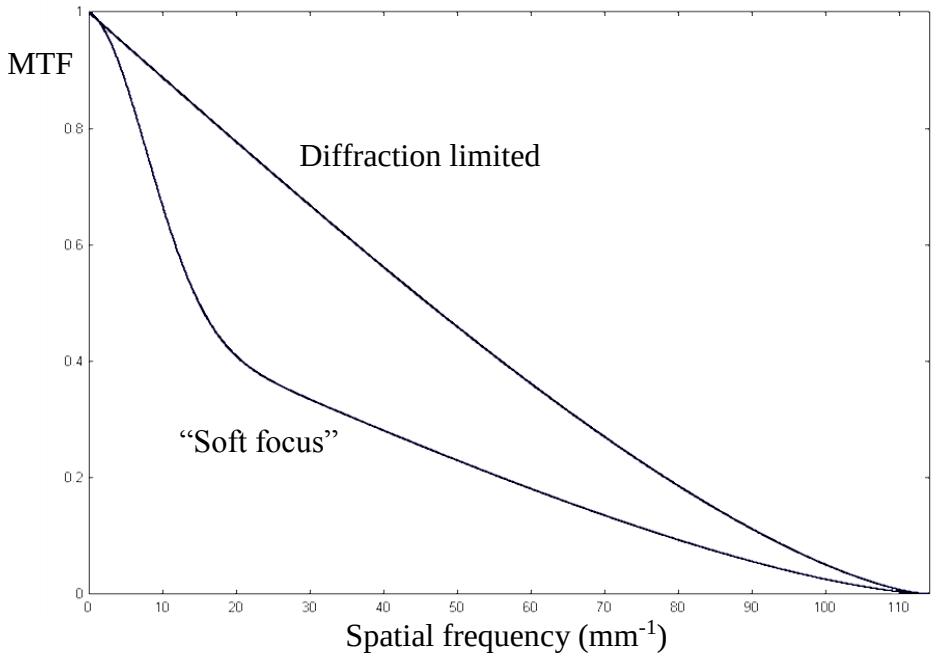


Taking the 2-dimensional Fourier transform of the *psf* above, we get the *MTF* curve shown on next page. Again we assume a diffraction-limited lens with F-number 16. Only one stray light level is illustrated, but as in the previous example a higher stray light level results in a larger initial drop in *MTF*. The initial drop in *MTF* is not as rapid in this case as when the stray light covered the entire sensor area - the wider the halo the more rapid the drop will be.

Compensating for stray light in this second case with a limited halo is not as simple as in the first case, where the stray light had a constant level over the entire image plane. Different spatial frequencies are now affected to different degrees by the stray light. If one makes a full deconvolution processing (Appendix 6) the effects of both stray and diffraction/aberrations can in principle be compensated for up to a certain frequency limit. This is, however, a complicated process that requires an accurate measurement of the total *psf*. Also, as mentioned in Appendix 6, noise will be amplified in the process. So, again the conclusion is that stray light should be eliminated at the source as far as possible.

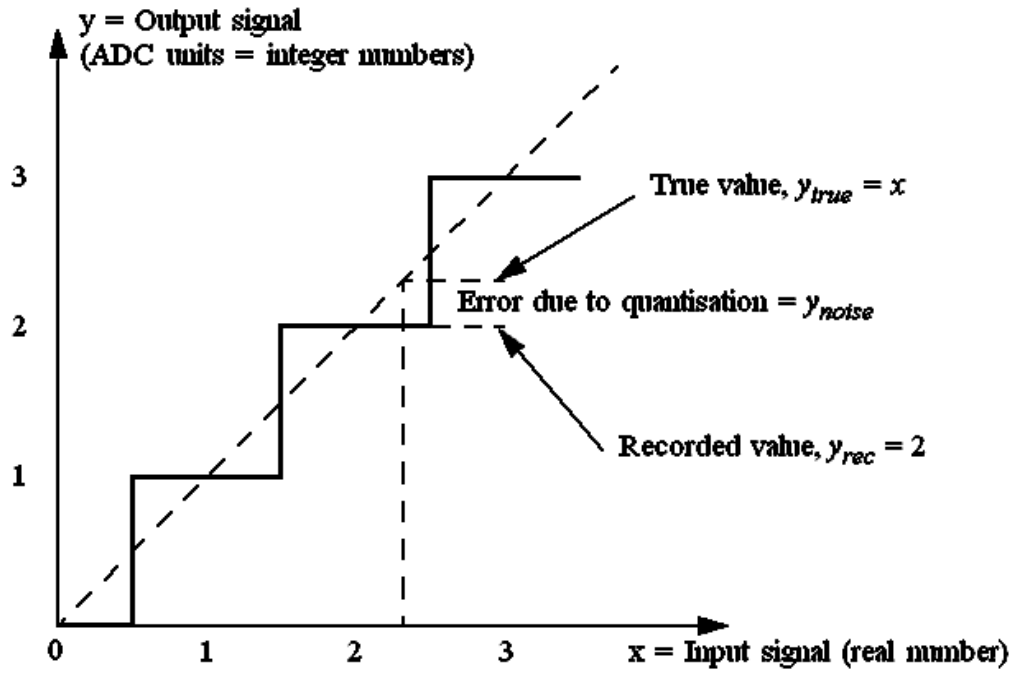


Finally it can be mentioned that *psf*s displaying a halo can, in some cases, be desirable. This is the case in portrait photography, because in lends a romantic touch to the images and smoothes out wrinkles (soft focus effect). But the halo edge should not be as sharp as shown on previous page, resulting in an oscillating *MTF* curve. By making the halo edge smoother, we can obtain an *MTF* as shown in the figure below. *Psf*s with such halos can be obtained through a suitable optical design, or added later by using a diffusing filter in front of the lens.



Appendix 3: Quantization Noise and the Number of Bits in the ADC

Noise is a random variation from the true signal value. When a signal is digitized, the true value is rounded off so that it can be described as, for example, an integer in the range 0-255 (8 bits). Therefore, digitization will introduce noise. The digitization process is depicted schematically in the figure below.



The digitization process means that, for example, input signals in the range 1.5 to 2.5 will all give the same output value, namely 2. If we assume that all signal values between 1.5 and 2.5 occur with equal probability, we can calculate the RMS quantization noise as:

$$n_{ADC} = \sqrt{\frac{1}{2.5-1.5}\cdot\int_{1.5}^{2.5} y_{noise}^2 dx} = \sqrt{\int_{1.5}^{2.5}(y_{rec}-y_{true})^2 dx} = \sqrt{\int_{1.5}^{2.5}(2-x)^2 dx} = \frac{1}{\sqrt{12}}$$ ADC units.

The same result would, of course, be obtained for signals within any $\pm\frac{1}{2}$ range around an integer number on the x axis, and therefore the RMS noise will be the same regardless of the signal level.

The quantization noise is added to other sources of noise, and, assuming that there is no correlation between the noise sources, we get: $n_{tot} = \sqrt{n_1^2 + n_2^2 + n_3^2 + \ldots}$, where n_1 etc. are the RMS noise values for the different sources (photon quantum noise, amplifier noise, ADC noise etc.).

An important question in a digital imaging system is how many bits we need in the ADC. We want the quantization noise to be so low that it does not affect the image quality in a negative

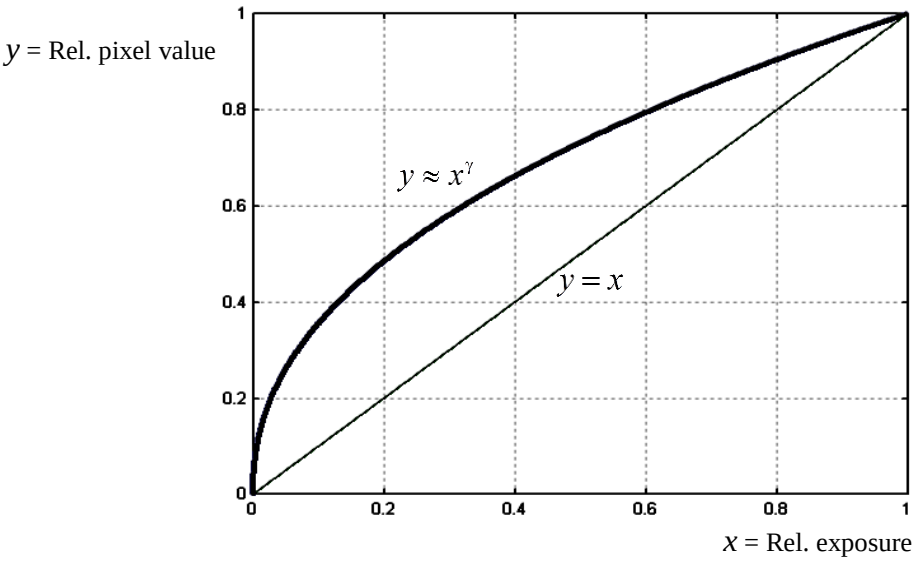
way. The requirement is that we must be able to handle the highest light values from the sensor, and at the same time be able to handle subtle differences in gray value in the shadows of a scene. Let's start with the shadows. In the darkest shadows the light level is virtually nil. This means we can forget photon noise, and the noise level is mainly determined by dark current noise and amplifier noise (we don't include quantization noise at this stage). We will call this level "dark noise," denoted by n_{dark}. Let's assume that the RMS value of n_{dark} corresponds to one ADC unit. In the presence of quantization noise, we then get a total dark noise n_{tot} (in ADC units) given by

$n_{tot} = \sqrt{n_{dark}^2 + n_{ADC}^2} = \sqrt{1 + \frac{1}{12}} = 1.04$ ADC units, i.e. the level of the dark noise has risen by 4% because of the added quantization noise. As a consequence the dynamic range has been reduced by the same percentage. This seems acceptable, and therefore a simple practical rule is to let the sensor dark noise (including amplifier noise) correspond to one ADC unit. This means that the maximum output signal simply corresponds to the dynamic range of the sensor. So, if the dynamic range is 4000 (typical for high-quality digital cameras) the maximum output signal will be 4000, corresponding to approximately 2^{12}. Consequently a 12-bit ADC would be appropriate in this case. In reality an ADC is not perfect. It will add some electronic noise, and the quantization steps may not be perfectly equal. Therefore many high-quality digital cameras employ 14-bit ADCs.

Cameras that employ gamma correction (see Appendix 4) before analog-to-digital conversion, can use fewer bits in the ADC without sacrificing dynamic range because of quantization noise.

Appendix 4: Gamma correction

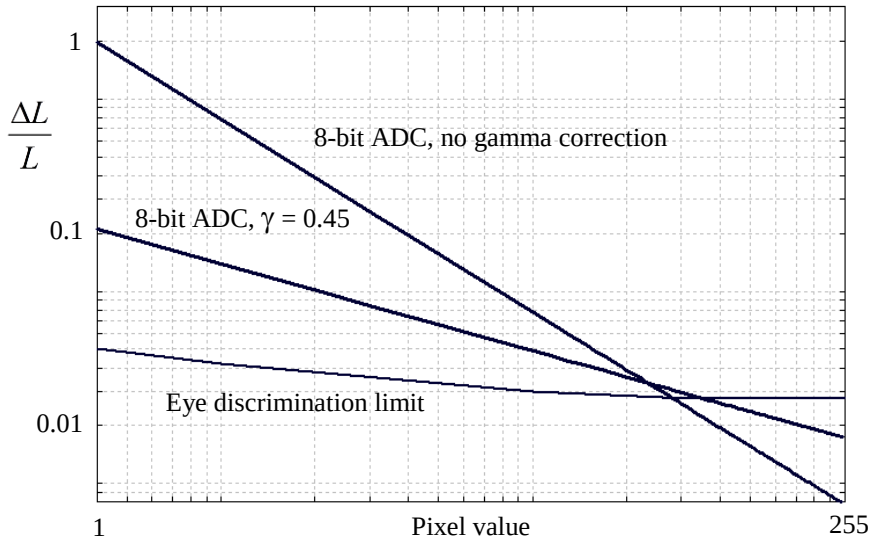
Electronic sensors used in digital cameras (CCD and CMOS sensors) have a linear response, i.e. the signal output increases linearly with light exposure. This is the desired behavior in most technical and scientific applications where images are used to extract quantitative photometric information. In cases where the images are intended to be looked at, but not used for quantitative evaluation, the situation is a bit different. In such cases we need to consider the non-linear behavior of the human eye. Therefore camera images produced in standard output formats, for example jpeg and tiff, are processed in such a way that the relationship between exposure and pixel value is non-linear. This processing is called gamma correction, and is illustrated by the thick black line in the figure below. The value of γ is often approximately 0.45.



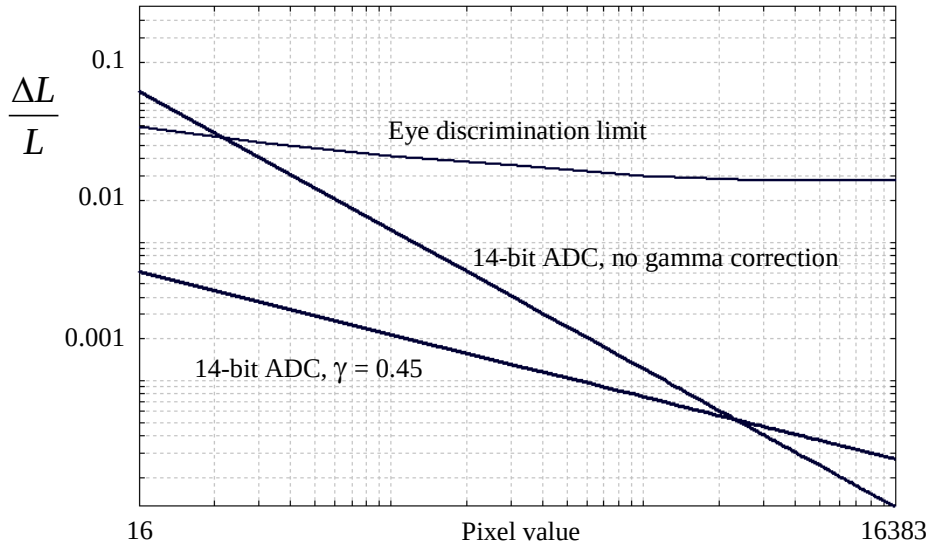
The effect of gamma correction is that the quantization of gray levels performed by the ADC will be more closely spaced in dark regions than in bright regions. This is desirable because the eye is more sensitive to small gray level variations in dark image regions. The eye's ability to discriminate between different luminance levels can be very roughly described by the equation $\frac{\Delta L}{L} \approx 0.01$. This means that small luminance variations ΔL can be detected at low luminance levels. When using gamma correction, it can be shown that a change of one unit in the pixel value corresponds to $\frac{\Delta L}{L} = \frac{1}{\gamma \cdot 2^N \cdot x^\gamma}$, where N is the number of bits in the ADC, and x is the relative exposure (x = 1 gives the maximum pixel value 2^N)♣. On next page graphs of $\frac{\Delta L}{L}$ are shown for different N-values with and without gamma correction. Also displayed is a more accurate curve for the eye discrimination capability than the equation $\frac{\Delta L}{L} \approx 0.01$ used above.

♣ In reality the maximum pixel value will be 2^N-1, but we neglect this small discrepancy because its effect on the result will be negligible.

The graphs below display what one ADC unit corresponds to in terms of relative luminance change, $\frac{\Delta L}{L}$. Also shown is the discrimination threshold for the eye. In regions where $\frac{\Delta L}{L}$ is higher than the eye threshold, there is a risk that the quantization steps of the image grayscale will be visible. The steps will be more clearly visible if $\frac{\Delta L}{L}$ is much larger than the eye threshold, which is the case in dark regions when no gamma correction is used. The common image file format JPEG only allows 8-bit pixel data, so in this case gamma correction is necessary to avoid visible grey steps. Note that the scales in the diagram are logarithmic, and therefore the effect of gamma correction is larger than it appears. At the lowest pixel levels $\frac{\Delta L}{L}$ is an order of magnitude lower for the gamma-corrected case.



Below is the corresponding diagram when using a 14-bit ADC. This number of bits is often used for RAW file format in digital cameras, and then no gamma correction is made (and as can be seen in the diagram it is hardly needed). The total range of pixel values displayed is 1000:1, which corresponds approximately to what is available in output media (display screens etc.).



Gamma correction also has implications concerning the number of bits that are necessary to utilize the dynamic range of the sensor. This was investigated for the linear case (i.e. $\gamma = 1$) in Appendix 3. We will now see how $\gamma \neq 1$ will change the situation.

Let us denote the analog, linear output signal from the sensor by *x*. Furthermore, we normalize the *x*-value so that $x = 1$ corresponds to the largest output signal (saturation). It can then be shown that one ADC unit corresponds to $\Delta x \approx \frac{x^{1-\gamma}}{\gamma \cdot 2^N}$, where *N* is the number of bits in the ADC. Let's take a practical example to see what the implication of this is.

Example: 8-bit ADC, $\gamma = 0.45$.

Let's assume that the dark level (average value) of the sensor is 5. This means $x = \frac{5}{256} = 0.02$. Inserted into the above equation this gives that one ADC unit corresponds to $\Delta x = 1.0 \times 10^{-3}$. If, as in Appendix 3, we let the dark noise level correspond to one ADC unit, we get that the maximum dynamic range of the sensor that can be utilized is $\frac{x_{\max}}{\Delta x} = \frac{1}{1.0 \times 10^{-3}} = 1000$ (assuming we don't want to lose more than 4% of the sensor's dynamic range due to quantization noise). Without gamma correction the corresponding maximum dynamic range would be approximately 250 (cf. Appendix 3). Using gamma correction, it is thus possible to better utilize a high dynamic range sensor with fewer bits. The result is, however, somewhat dependent on the dark level, which we have assumed to be 5. By reducing the dark level to 2, we can utilize sensors with a dynamic range of ≈ 1700. If we instead increase it to 10, the corresponding dynamic range will be ≈ 700. It is therefore an advantage to keep the dark level low. But one must be careful not to reduce the dark level too much, because then there is a risk that we get negative "clipping", i.e. pixel values of zero. Zero pixel values mean we have lost information, even if it occurs only in the negative noise "spikes", because it will perturb the average value recorded. One should also keep in mind that the dark level of a sensor will change somewhat with the ambient temperature (unless the sensor is temperature stabilized).

The case of 8-bit ADC with gamma correction is of great interest in digital photography, because most amateur cameras produce output files in JPEG format which allows only 8 bits of data. With typical parameter values, as shown in the example above, we can thus expect that this format will be quite sufficient for sensors with a dynamic range of the order of 1000. If, on the other hand, we have a high-quality sensor with a dynamic range of 4000, this number will be reduced to about 2600 by an 8-bit ADC with a gamma of 0.45. Therefore high-quality cameras have another file format in addition to JPEG. This so-called RAW format usually supports 12- or 14-bit pixel values. With so many bits gamma correction in not needed (Appendix 3), and therefore these RAW files usually use $\gamma = 1$.

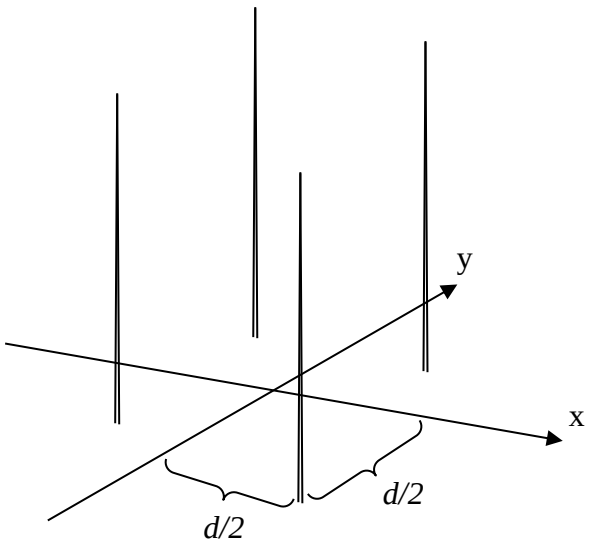
Finally it should be mentioned that most digital cameras don't use the simple gamma correction $y = x^{\gamma}$ that we have assumed in this section. Instead different gamma values are used depending on the exposure level. This of course influences the results, which should therefore only be regarded as rough approximations.

Appendix 5: Anti-aliasing filters

To avoid aliasing problems due to under-sampling (see chapter 17), anti-aliasing filters are used in many digital cameras and video cameras. They are often placed immediately in front of the electronic sensor. A common filter design is to use two birefringent crystals with a quarterwave plate in between. The effect of such a filter is to produce double images both in the vertical and horizontal directions. So instead of a single image, we add together four images with small displacements horizontally and vertically. These displacements, *d*, are approximately the same as the center-to-center distance between pixels, i.e. typically in the range 3-8 μm in digital cameras.

The effect of an anti-aliasing filter is that it blurs the image, so that spatial frequencies above the Nyquist frequency are attenuated. This blurring effect can be described by a *psf* and *MTF* for the filter, as seen below. It is not obvious that *MTF* theory is applicable, because this requires that the four split-up images are added incoherently. Analysis shows, however, that due to the birefringence there is a difference in optical pathlength between the split up images. This difference in pathlength is, under normal circumstances, larger than the coherence length of the light. Therefore, we will assume that the total illuminance level on the sensor is given by the sum of the illuminance levels in the four displaced images.

The *psf* of the filter is given by the four delta functions symbolically represented below.



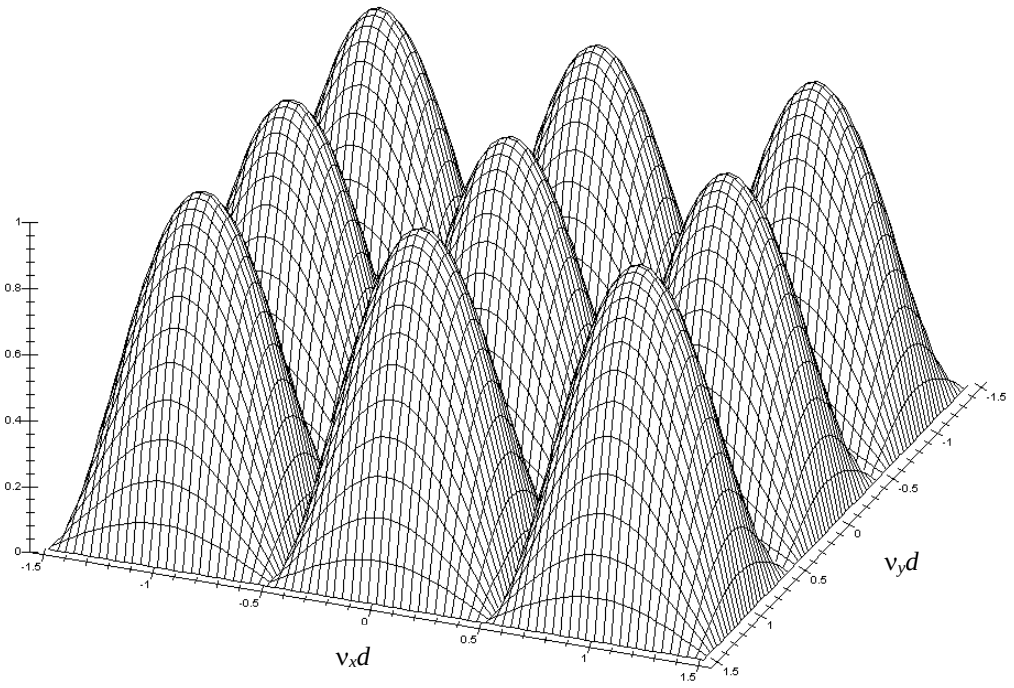
Mathematically the *psf* can be described by

$$psf(x,y) = \delta(x-\frac{d}{2}, y-\frac{d}{2}) + \delta(x-\frac{d}{2}, y+\frac{d}{2}) + \delta(x+\frac{d}{2}, y-\frac{d}{2}) + \delta(x+\frac{d}{2}, y+\frac{d}{2})$$

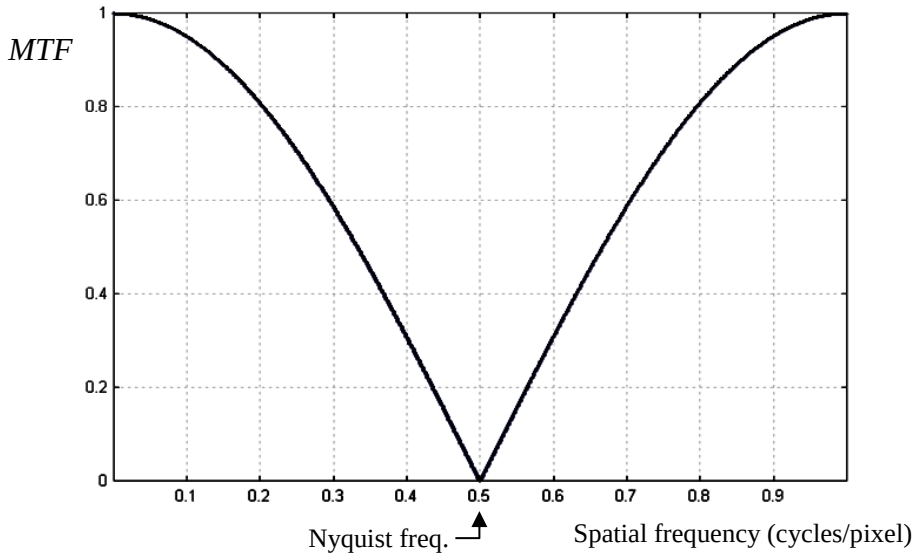
Taking the 2D Fourier transform of this, and normalizing to 1 at the origin, we get

$$OTF(\nu_x, \nu_y) = \cos(\pi\nu_x d)\cdot\cos(\pi\nu_y d)\,.$$

$MTF = |OTF| = |\cos(\pi\nu_x d) \cdot \cos(\pi\nu_y d)|$ is illustrated in the figure below.

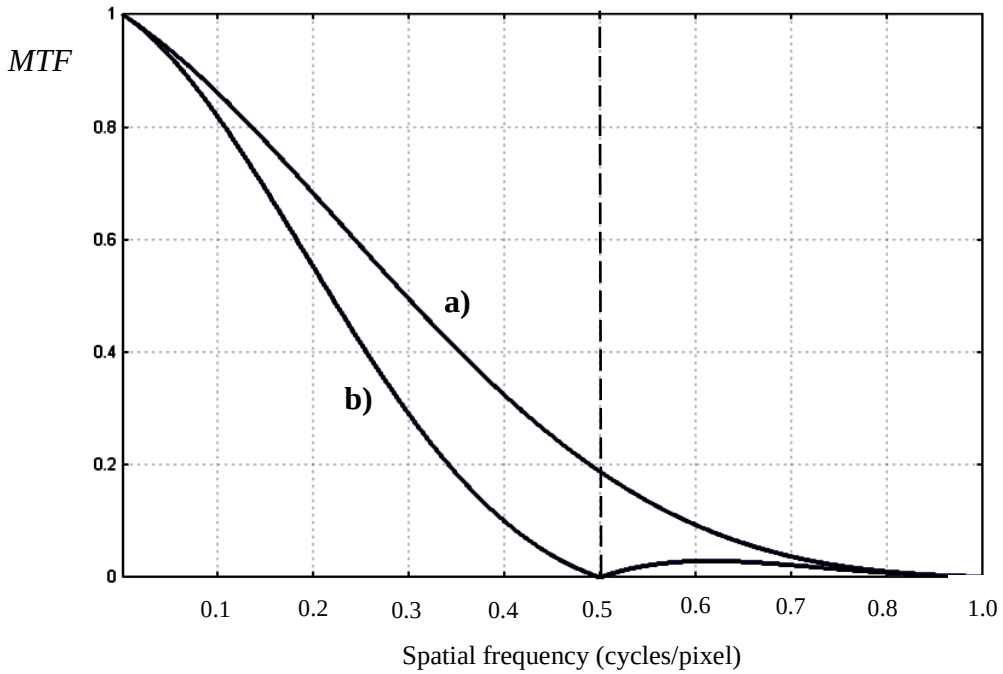


In the one-dimensional case ($\nu_y = 0$) we get $MTF(\nu_x) = |\cos(\pi\nu_x d)|$, which is illustrated below for a *d*-value equal to the center-to-center distance between pixels.

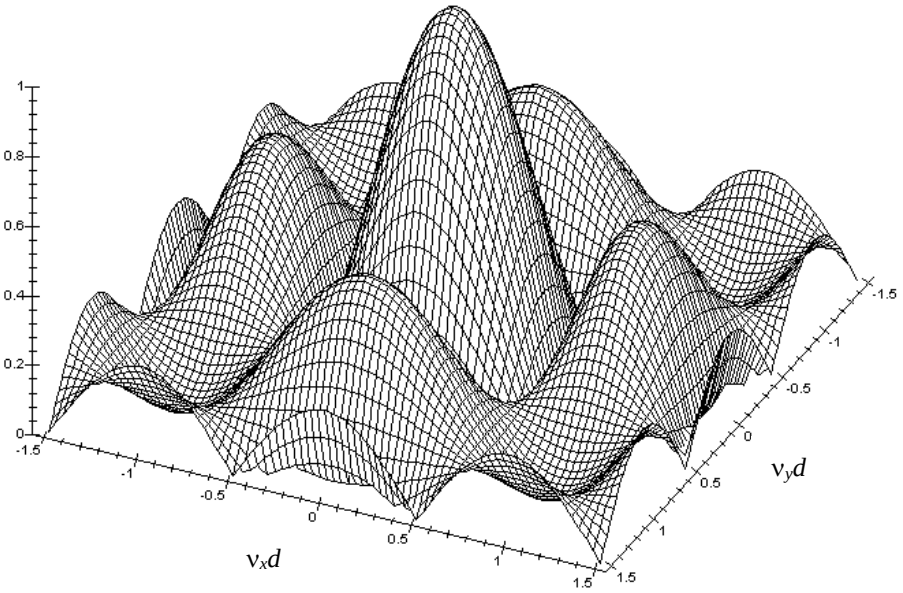


The spatial frequency is given in units of cycles/pixel. A spatial frequency of 1 therefore corresponds to a line pattern with a period length equal to the center-to-center distance between pixels. This way of expressing frequencies is very common in connection with digital photography. It is convenient, because it means that the Nyquist limit is always located at frequency 0.5. Looking at the curve, we can see that the *MTF* of the filter is far from ideal. Ideally it should eliminate all spatial frequencies above the Nyquist limit, and let all frequencies below pass without attenuation. This means that the *MTF* should be 1 up to 0.5 cycles/pixel, and 0 above that frequency. The high *MTF* values at frequencies close to one cycle/pixel may look alarming, because they can potentially result in very coarse, high-contrast moiré effects in

the image. In practice they are usually not so problematic, because at these high frequencies the total *MTF* is usually rather low because of limitations in optics and sensor. To illustrate this, typical *MTF* curves are shown below for a digital camera without (a) and with (b) anti-aliasing filter. The filter strongly reduces aliasing effects, but it doesn't eliminate them. From the curves we can also see that we have to pay a rather high price for this aliasing reduction. With an anti-aliasing filter we get considerably lower *MTF* values also well below the Nyquist limit. As a result the images will look slightly unsharp and soft.



It is possible to design anti-aliasing filters with more than two birefringent crystals. In this way one can get better attenuation of frequencies above the Nyquist limit. This is illustrated in the figure below (using 4 birefringent crystals), which can be compared to the figure on previous page showing a 2-filter design. Usually a 2-crystal design works well enough in practice.



Appendix 6: Deconvolution

With cheap and powerful computers available, it is today possible to a certain extent to compensate for the influence of the point spread function (*psf*) on image data. As we have seen, this influence can be described mathematically as

$$I_B = I_O \otimes psf$$

where I_B and I_O represent the image and object functions respectively, and $\otimes$ denotes convolution. Taking the Fourier transform of this equation, and rearranging, we obtain

$$\hat{I}_O = \frac{\hat{I}_B}{\hat{psf}}$$

where the symbol ^ denotes the Fourier transform. This means that if we have measured the *psf* of our imaging system, and recorded an image function I_B, we can take their Fourier transforms and insert into the equation above. This will give us the Fourier transform of the true object function I_O. By taking the inverse transform, we then get I_O. This procedure is called deconvolution, and it can be carried out in one, two and higher dimensions. There is a catch, of course, otherwise we could get infinitely high resolution by using this procedure. The catch is that both $\hat{I}_B$ and $\hat{psf}$ become zero at high spatial frequencies, so that we end up dividing zero by zero. However, up to the frequency where $\hat{psf}$ becomes zero (i.e. the limiting frequency of the *MTF*), it should in principle be possible to restore the amplitudes perfectly. What this means in practice, is that the contrast of small specimen details can be improved so that they can be seen more clearly. A practical problem with this restoration is that it also amplifies high-frequency noise. Therefore deconvolution methods often employ some technique for noise suppression (but there is always a trade-off between resolution and noise).

Appendix 7: English-Swedish Dictionary

Följande lilla ordlista är verkligen i mikro-format. Avsikten har inte varit att sammanställa något slags komplett ordlista, utan att ta med dels de speciella ord och uttryck som har en koppling till ämnesområdet "Imaging Physics," vilket närmast blir "Bildfysik" på svenska, dels i övrigt sådant som bedömts viktigt för förståelsen av texten.

aberration avvikelse, avbildningsfel
accessible tillgänglig
actual verklig
ADC level ADC-nivå. Exempelvis en 8-bitars ADC har $2^8 = 256$ diskreta nivåer
adjacent intilliggande
aggravate förvärra
aliasing (används ofta även i svenskan) vikning (innebär att frekvenser över Nyquistfrekvensen "viks" tillbaka och återges som lägre frekvenser)
amplifier förstärkare
angular vinkel-
anode anod (pos. elektrod)
application tillämpning
arbitrary godtycklig
arc minute bågminut (1/60 grad)
arc second bågsekund (1/60 bågminut)
area array sensor matrisdetektor
array ett antal element som är regelbundet ordnade (i rad eller fyrkant etc)
artifact (i detta sammanhang) felaktighet
avalanche lavin
average value medelvärde
barely nätt och jämnt
bias (systematisk) avvikelse
binary star dubbelstjärna
bleach bleka
blur sudda, suddighet
boost förhöja
boundary gräns
brightness ljushet (inte fysikaliskt väldefinierat)
cathode katod (neg. elektrod)
cellular phone mobiltelefon
cf. jmf.
characteristic egenskap
charge laddning
circuit krets
compatible kompatibel, passa ihop med
comprehensive uttömmande
concept begrepp
consecutive på varandra följande
contribution bidrag
conversion omvandling
convolution faltning

coolant köldmedel
correlation korrelation, samband
current ström
dark current mörkerström
decay avklinga, avta
degree of modulation modulationsgrad
denote beteckna
depict avbilda
deplete utarma
deviate avvika
diaphragm bländare
diffraction-limited diffraktionsbegränsad
digitise (även digitize) digitalisera
discern urskilja
discrete diskret (till skillnad från kontinuerlig)
displacement förskjutning
distribution fördelning (t.ex. i "light distribution", "probability distribution")
down to earth jordnära
drum trumma
dynamic range kallas ofta samma på svenska. Egentligen "dynamiskt område"
dynode dynod (elektrod mellan anod och katod)
elaborate utveckla, förtydliga
elementary charge elementarladdning
encumbered behäftad
envelope hölje
event händelse
exceed överskrida
expose exponera
expression uttryck
finite ändlig (motsats: infinite)
fluorescence fluorescens
flux flöde
f-number bländartal
focal length brännvidd
frame rate bildfrekvens
fundamental (tone) grundton
gain förstärkning
grainy grynig, kornig
hardware hårdvara, dvs. elektronik, mekanik mm. Motsats: software (mjukvara)
harmonic överton
i.e. dvs.
illuminance belysning
illustrious lysande (i bildlig betydelse, har inget med fysik att göra)
imaging physics bildfysik
imaging scale avbildningsskala
impinge stöta på, kollidera
incident infallande
infer sluta sig till

infinite oändlig (motsats: finite)
integer (number) heltal
interpret tolka
intrinsic inneboende
irradiate instråla
irretrievably oåterkalleligen
layman lekman
lattice gitter (regelbundet mönster)
lens lins, objektiv (= system med flera linser)
limiting frequency gränsfrekvens
linear sensor raddetektor
luminance luminans
luminous flux ljusflöde
luminous intensity ljusstyrka
magnification förstoring
magnitude storlek, styrka
mask maskera (bort)
mean value medelvärde
measure mäta, mått
merge sammansmälta
micron mikrometer
mobile rörlig
modulus absolutbelopp
negligible försumbar
nil noll, inget
nitrogen kväve
noise brus
numerical aperture numerisk apertur (mått på ljusinsamlande förmåga)
omit utesluta
optical transfer function optisk överföringsfunktion
optional extra, utöver vad som krävs
origin origo
peak topp
pedestrian fotgängare
pedestrian crosswalk övergångsställe
perpendicular vinkelrät
phase transfer function fasöverföringsfunktion
photomultiplier fotomultiplikator
point source punktkälla
point-spread-function punktspridningsfunktion
power (fysikaliska storheten) effekt
power density effekttäthet
probability sannolikhet
process behandla, (fotografiskt) framkalla
prolong förlänga
property egenskap
pun skämt
pure AC signal en signal vars medelvärde är noll

quadruple fyrfaldiga
quality measure kvalitetsmått
quantization noise kvantiseringsbrus
quantize kvantisera
quantum conversion efficiency kvantverkningsgrad
quote citera
radiate utstråla
random slumpvis
range område, intervall
rate hastighet
real reell
reception mottagning
recreate, re-create återskapa
rect(angular) function rektangulärfunktion
rendition återgivning
resolution upplösning
retina näthinna
reverse-biased backspänd
root-mean-square (RMS) effektivvärde
sampling (används ofta i svenskan också) provtagning
saturation mättnad
scaling factor skalfaktor
semiconductor halvledare
sensor begreppen sensor och detektor i samband med digital bildregistrering tenderar att skilja sig något i engelskt och svenskt språkbruk. På engelska innebär sensor en samling (rad, matris) av enskilda detektorelement. På svenska tenderar man att istället för sensor kalla hela raden eller matrisen för detektor.
shift förskjuta
shutter slutare
shutter speed slutartid
signal-to-noise ratio signal/brus förhållande
sinc (function) funktion av typen sin(x)/x
software mjukvara, dvs. datorprogram. Motsats till hardware (hårdvara)
space shuttle rymdfärja
spatial frequency ortsfrekvens
speed (i samband med objektiv) ljusstyrka
spike tagg, spets
standard deviation standardavvikelse
storage lagring
stripe rand
subtend uppta (i fråga om vinkel)
subtle subtil, hårfin
superimposed överlagrad
suppress undertrycka
thermal termisk
transfer function överföringsfunktion
transition övergång
transmitter sändare

trial-and-error (används också ofta i svenskan) prova sig fram
trough vågdal (kan även betyda tråg, t.ex. matho för svin o.dyl.)
unequivocal entydig
uniform jämn
unit enhet
unity 1 (100%)
utilize utnyttja
video-tape video-banda
viewing distance betraktningsavstånd
viewpoint synpunkt

Appendix 8: Formulas

Physical constants:

Speed of light in vacuum	c_0 = 299 792 458 m/s (exactly)
Planck's constant	h = 6.626×10^{-34} Js
Elementary charge	e = 1.602×10^{-19} As

Diffraction: $D\sin\phi = 1.22\,\lambda$, where D = lens diameter and ϕ = diffraction angle.

Lens formula: $\frac{1}{a}+\frac{1}{b}=\frac{1}{f}$

Stefan Boltzmann's law: $M_e = \sigma T^4$ ($\sigma = 5.6705\times10^{-8}$ Wm^{-2} $K^{-4)}$

Wien displacement law: $\lambda_{max}\ T = const.$ ($const. = 2.8978\times10^{-3}\,m\cdot K$)

Luminous intensity: $S = \frac{d\Phi}{d\Omega}$ (cd = lumen/steradian)

Luminance: $L = \frac{d^2\Phi}{dA\,d\Omega\cos\vartheta}$ (L is constant, i.e. independent of ϑ, for a diffuse source/reflector)

Total flux from diffuse source: $\Phi = \pi LA$

Illuminance: $E = \frac{d\Phi}{dA}$ (lux = lumen/$m^{2)}$

Illuminance in image plane: $E = \frac{L\pi}{4}\left(\frac{D}{f}\right)^2$, L = luminance, D = lens diameter and f = focal length.

Root-mean-square (RMS) noise: $i_{noise} = \lim_{T\to\infty}\sqrt{\frac{1}{T}\int_0^T (i-i_{average})^2\,dt}$ (analog signal i)

Standard deviation: $s = \sqrt{\frac{\sum(i_k - i_{average})^2}{n-1}}$ (digital signal $i_{k)}$

(In the following equations RMS means standard deviation in cases where the signal is digital rather than analog)

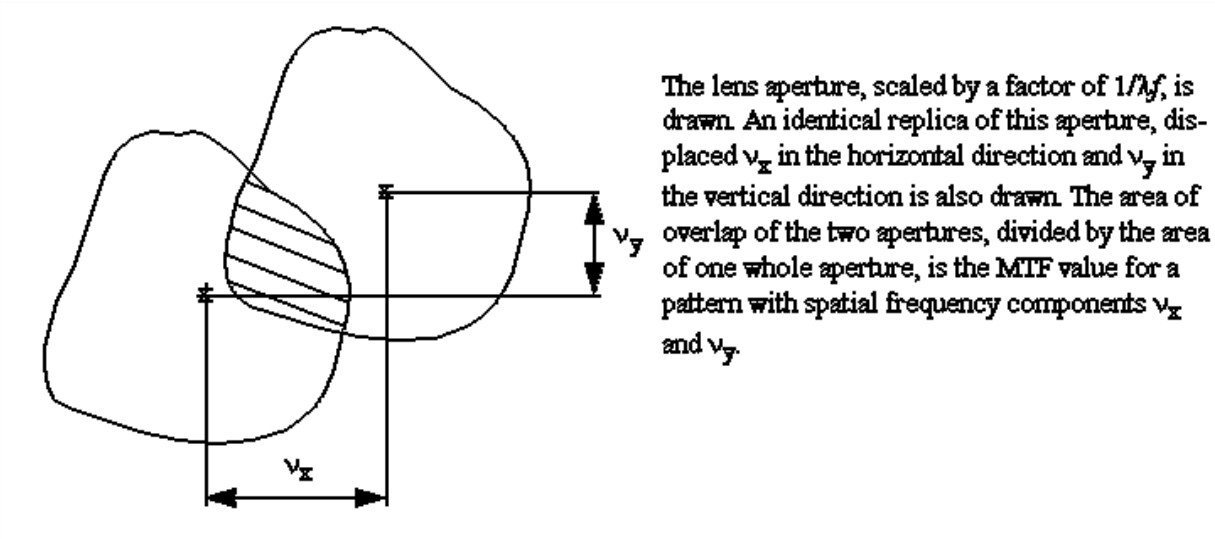
Signal-to-noise ratio: $SNR = \dfrac{\text{mean value}}{\text{RMS noise}}$

Dynamic range: $\dfrac{\text{Maximum output signal}}{\text{Minimum RMS noise}}$

Quantization noise: $\dfrac{1}{\sqrt{12}}$ of least significant bit (ADC unit) RMS

Several independent noise sources: $n_{tot} = \sqrt{n_1^2 + n_2^2 + \ldots}$ (RMS)

***OTF* for diffraction-limited lens:**



The lens aperture, scaled by a factor of $1/\lambda f$, is drawn. An identical replica of this aperture, displaced ν_x in the horizontal direction and ν_y in the vertical direction is also drawn. The area of overlap of the two apertures, divided by the area of one whole aperture, is the MTF value for a pattern with spatial frequency components ν_x and ν_y.

Aliasing: The spatial frequency obtained when aliasing occurs, ν_{alias}, is given by the conditions $\nu_{alias} = |n\nu_S - \nu|$, $n = 1, 2, 3, \ldots$, and $\nu_{alias} \le \dfrac{\nu_S}{2}$, which must be fulfilled simultaneously. ν_s = sampling frequency. ν = true frequency.

Index