Friedrich Herrmann

The Karlsruhe physics course

Lecture notes

Optics

The Karlsruhe physics course

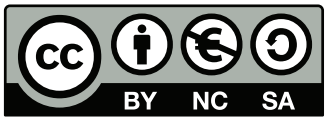
Lecture notes

- Mechanics
- Thermodynamics
- Electromagnetism
- **Optics**

Der Karlsruher Physikkurs

Edition 2020

Prof. Dr. *Friedrich Herrmann*



1

Decomposition of continuous signals

For many technical applications of optics and information theory it is useful to decompose a function that is continuous in time or space. Depending on the application, different decompositions are suitable. We will first discuss the Fourier decomposition (harmonic analysis, Fourier analysis). For example, a function of time is decomposed into sine and cosine functions of different frequencies. In section 1.2 we consider a decomposition into functions that have values which are significantly different from zero only in a limited range on the time axis.

1.1 Harmonic analysis

There are physical systems that decompose a given function into sine or cosine functions:

- An optical filter allows only sinusoidal light waves of certain frequencies to pass.
- A prism deflects sinusoidal light waves differently depending on the frequency.
- An electrical transmission channel only lets pass certain harmonic components of signals.
- A lens sorts spatial sine structures of an object in the focal plane according to the "spatial frequency".
- A piano with a pressed pedal sorts an incoming sound wave according to sinusoidal components.

This does not mean that the "true nature" of light, sound or electrical signals is that they are composed of sine functions. They could just as easily have been broken down into other parts. The decomposition into harmonic parts is useful because nature itself often performs this decomposition. In some cases a decomposition of a time-dependent quantity is appropriate, in others it is a decomposition of a position-dependent quantity. In the following we will use the time as a variable, to be specific. However, all results apply – mutatis mutandis – also to functions of position.

The method of decomposing a function into sine and cosine components is called *Fourier analysis* or *harmonic analysis*. The inverse, i.e. the composition of a non-harmonic function from harmonic components, is called *Fourier synthesis*.

So far we have spoken of devices or physical systems that perform a Fourier analysis or synthesis. Of course, these processes can also be understood as mathematical processes and, for example, a function given on paper can be decomposed by means of mathematics. For the mathematician, each of the natural or technical processes listed above represents a kind of analog computer.

With a little practice, one can often see what harmonic components a function contains. We will learn some rules in the following mathematical treatment of harmonic analysis.

We start with a special case. The function to be analyzed is periodic: $f(t) = f(t + T)$. J. B. Fourier (1768-1830) has shown that any such function can be decomposed into oscillations of frequencies

$$\omega_1 = 1\cdot\frac{2\pi}{T}, \omega_2 = 2\cdot\frac{2\pi}{T}, \ldots \omega_n = n\cdot\frac{2\pi}{T}, \ldots$$

i.e. into a fundamental oscillation and its harmonics.
So it is

$$f(t) = A_0 + \sum_{n=1}^{\infty}\left(A_n\cdot\cos n\omega t + B_n\cdot\sin n\omega t\right) \tag{1.1}$$

with

$$\omega = \frac{2\pi}{T}$$

If $f(t)$ is known, the coefficients A_n and B_n can be calculated:

$$A_0 = \frac{1}{T}\int_0^T f(t)\,dt = \bar{f}(t)$$

$$\left.\begin{aligned} A_n &= \frac{2}{T}\int_0^T \cos n\omega t\,dt \\ B_n &= \frac{2}{T}\int_0^T \sin n\omega t\,dt \end{aligned}\right| \quad n = 1, 2, \ldots$$

The following decomposition is equivalent to this:

$$f(t) = \sum_{n=-\infty}^{\infty} a_n e^{jn\omega t} \tag{1.2}$$

$$a_n = \frac{1}{T}\int_0^T f(t)e^{-jn\omega t}\,dt \quad n = 0, \pm 1, \pm 2, \ldots \tag{1.3}$$

The coefficients in (1.1) and (1.2) are related according to

$A_0 = a_0$
$A_n = \text{Re}\,(2a_n)$
$B_n = -\text{Im}\,(2a_n)$

For $f(t)$ to be real, the following must apply

$a_n = a_{-n}^*$

The contributions of the terms with $(+n)$ and $(-n)$ taken together are then real. If $f(t)$ is an even function, in equation (1.1) the coefficients B_n are zero, and in equation (1.2) the $a_n = a_{-n}$ are real. If $f(t)$ is odd, in equation (1.1) the A_n (including A_0) disappear and in equation (1.2) the $a_n = -a_{-n}$ are imaginary. Note also the meaning of A_0 and a_0: This coefficient simply represents the time average of the function $f(t)$.

As an example we consider the rectangular function of Fig. 1.1:

$f(t) = 1$ for $0 \le t < T/2$
$f(t) = -1$ for $T/2 \le t < T$

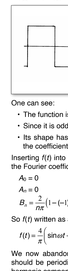


Fig. 1.1
Periodic rectangular function

One can see:

- The function is odd, so it has only sin components.
- Since it is odd, its average value is zero.
- Its shape has a rough similarity with the sine function $\sin\omega t$, so the coefficient B_1 must have a high value.

Inserting $f(t)$ into the equations for A_0, A_n and B_n gives the values of the Fourier coefficients A_n and B_n:

$A_0 = 0$
$A_n = 0$
$B_n = \frac{2}{n\pi}\left(1-(-1)^n\right)$

So $f(t)$ written as a harmonic series is:

$$f(t) = \frac{4}{\pi}\left(\sin\omega t + \frac{\sin 3\omega t}{3} + \frac{\sin 5\omega t}{5} + \ldots\right)$$

We now abandon the restriction that the function to be analyzed should be periodic. In this general case it contains a continuum of harmonic components and instead of equation (1.2) we have

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{j\omega t}\,d\omega \tag{1.4}$$

The resolution for $F(\omega)$, which is mathematically somewhat difficult, results in

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t}\,dt \tag{1.5}$$

The function $F(\omega)$ specifies the continuous distribution of the harmonic components contained in $f(t)$. $F(\omega)$ is called the *spectral function*, or briefly the *spectrum* of $f(t)$.

We can also read equation (1.4) in a way that a function $F(\omega)$ is transformed into a function $f(t)$. It is said that a Fourier transformation is realized. Both equations (1.4) and (1.5) describe a transformation of the same type. So we can say that the spectral function is the Fourier transform of $f(t)$, and $f(t)$ is the Fourier transform of the spectral function. $f(t)$ also tells us which harmonic components the spectral function has. Applying the Fourier transform twice to a function $f(t)$ yields the same function $f(t)$ again, except for a factor 2π.

For $f(t)$ to be real, $F(\omega)$ must again satisfy a condition, namely $F(-\omega) = F^*(\omega)$. And again $F(\omega)$ is purely real if $f(t)$ is an even function.

As an example we consider the rectangular function shown in Fig. 1.2:

$$f(t) = \begin{cases} \dfrac{1}{\Delta t} & \text{for } -\dfrac{\Delta t}{2} < t < \dfrac{\Delta t}{2} \\ 0 & \text{else} \end{cases}$$

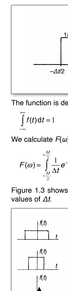


Fig. 1.2
Rectangular function

The function is defined in such a way that

$$\int_{-\infty}^{\infty} f(t)\,dt = 1$$

We calculate $F(\omega)$.

$$F(\omega) = \int_{-\frac{\Delta t}{2}}^{+\frac{\Delta t}{2}} \frac{1}{\Delta t}e^{-j\omega t}\,dt = \frac{e^{+j\omega\Delta t/2} - e^{-j\omega\Delta t/2}}{j\omega\Delta t} = \frac{\sin\left(\frac{\Delta t}{2}\omega\right)}{\frac{\Delta t}{2}\omega}$$

Figure 1.3 shows the original function and its spectrum for different values of Δt.

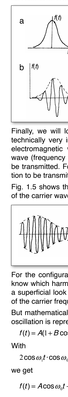


Fig. 1.3
Rectangular functions of different width. Left: Original function, right: spectrum

One recognizes an important property of the Fourier transform: The wider the original function, the narrower the Fourier transform.

Two more examples are shown (without calculation) in Fig. 1.4. A Gaussian curve centered at $t = 0$ is transformed into a Gaussian curve centered at $\omega = 0$, Fig. 1.4a. If the spectral function is a Gaussian curve centered at $\omega_0 \neq 0$, the original function must contain oscillations of the frequency ω_0, Fig. 1.4b.

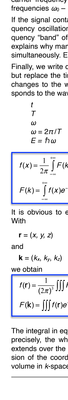


Fig. 1.4
(a) The spectral function is a Gaussian function centered at $\omega = 0$, the original function is a Gaussian function.
(b) The spectral function is a Gaussian function centered at ω_0, in the original function an oscillation with the frequency ω_0 can be seen.

Finally, we will look at an example that is particularly simple but technically very important. For the transmission of messages with electromagnetic waves, a high-frequency electromagnetic carrier wave (frequency ω_0) is used, which is modulated by the function to be transmitted. For the sake of simplicity, we assume that the function to be transmitted is a low-frequency sine wave of frequency ω_s.

Fig. 1.5 shows the case of an *amplitude modulation*: the amplitude of the carrier wave is multiplied by the function to be transmitted.

Fig. 1.5
Amplitude modulation: The amplitude of the carrier wave is multiplied by the function to be transmitted.

For the configuration of the transmission channel it is important to know which harmonic components the synthesized wave has. From a superficial look one might conclude that it contains one oscillation of the carrier frequency and one of the signal frequency.

But mathematical analysis shows that this is wrong. The modulated oscillation is represented by

$$f(t) = A(1 + B\cos\omega_s t)\cos\omega_0 t$$

With

$$2\cos\omega_s t \cdot \cos\omega_0 t = \cos(\omega_0 + \omega_s)t + \cos(\omega_0 - \omega_s)t$$

we get

$$f(t) = A\cos\omega_0 t + \frac{AB}{2}\cos(\omega_0 + \omega_s)t + \frac{AB}{2}\cos(\omega_0 - \omega_s)t$$

The overall oscillation therefore contains a partial oscillation of the carrier frequency ω_0 and two more components with the adjacent frequencies $\omega_0 - \omega_s$ and $\omega_0 + \omega_s$.

If the signal contains a whole spectrum of the width $\Delta\omega$ of low-frequency oscillations, the spectrum of the modulated wave is a frequency "band" of the width $2\Delta\omega$, which is centered around ω_0. This explains why many radio and television programs can be transmitted simultaneously. Each program occupies a different *frequency band*.

Finally, we write down the important equations (1.4) and (1.5) again, but replace the time by the position x. The oscillation period T then changes to the wavelength λ, and the angular frequency ω corresponds to the wave number k

t	$\rightarrow$	x
T	$\rightarrow$	λ
ω	$\rightarrow$	k
$\omega = 2\pi/T$	$\rightarrow$	$k = 2\pi/\lambda$
$E = \hbar\omega$	$\rightarrow$	$p = \hbar k$

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(k)e^{jkx}\,dk \tag{1.6}$$

$$F(k) = \int_{-\infty}^{\infty} f(x)e^{-kx}\,dx \tag{1.7}$$

It is obvious to extend the Fourier transform to three dimensions. With

$\mathbf{r} = (x, y, z)$

and

$\mathbf{k} = (k_x, k_y, k_z)$

we obtain

$$f(\mathbf{r}) = \frac{1}{(2\pi)^3}\iiint F(\mathbf{k})e^{j\mathbf{kr}}\,dk_x\,dk_y\,dk_z \tag{1.8}$$

$$F(\mathbf{k}) = \iiint f(\mathbf{r})e^{-j\mathbf{kr}}\,dx\,dy\,dz \tag{1.9}$$

The integral in equation (1.9) extends over the whole space, or more precisely, the whole position-space. The integral in equation (1.8) extends over the so-called *reciprocal space*, or *k*-space. The dimension of the coordinates in *k*-space is that of a reciprocal length. A volume in *k*-space has the dimension of a reciprocal normal volume.

1.2 The sampling theorem

For certain purposes a different decomposition than the Fourier decomposition is more suitable. In the following we consider the decomposition of a “frequency-band-limited” function $f(t)$ into sinc functions. We will first explain two terms:

sinc function

$$\operatorname{sinc} t = \frac{\sin t}{t}$$

band-limited function

A function whose spectrum does not go beyond a highest frequency $\omega = 2\pi B$.

Suppose $f(t)$ is such a function limited to $\omega < 2\pi B$. Then

$$f(t) = \sum_{n=-\infty}^{+\infty} a_n \cdot \frac{\sin 2\pi B\left(t - \frac{n}{2B}\right)}{2\pi B\left(t - \frac{n}{2B}\right)} \tag{1.10}$$

with

$$a_n = f\left(\frac{n}{2B}\right)$$

This decomposition has some interesting properties.

The coefficients a_n are simply the values of $f(t)$ at equidistant points on the t-axis. The shape of the continuous function is thus unambiguously defined by the function values for these discrete points in time. Of course, this is only possible because of the restriction that the Fourier components of the function do not exceed a maximum frequency.

For the “sample values” $f(n/2B)$ all but one of the summands of (1.10) are zero. If $n = n_0$, then only the summand with $n = n_0$ is different from zero.

The statement that a frequency-band-limited function can be developed according to equation (1.10) is called *sampling theorem*.

The sampling theorem ensures that a discrete sequence of numbers is sufficient for the transmission of a continuous signal, Fig. 1.6. This is applied, for example, for compact disks.

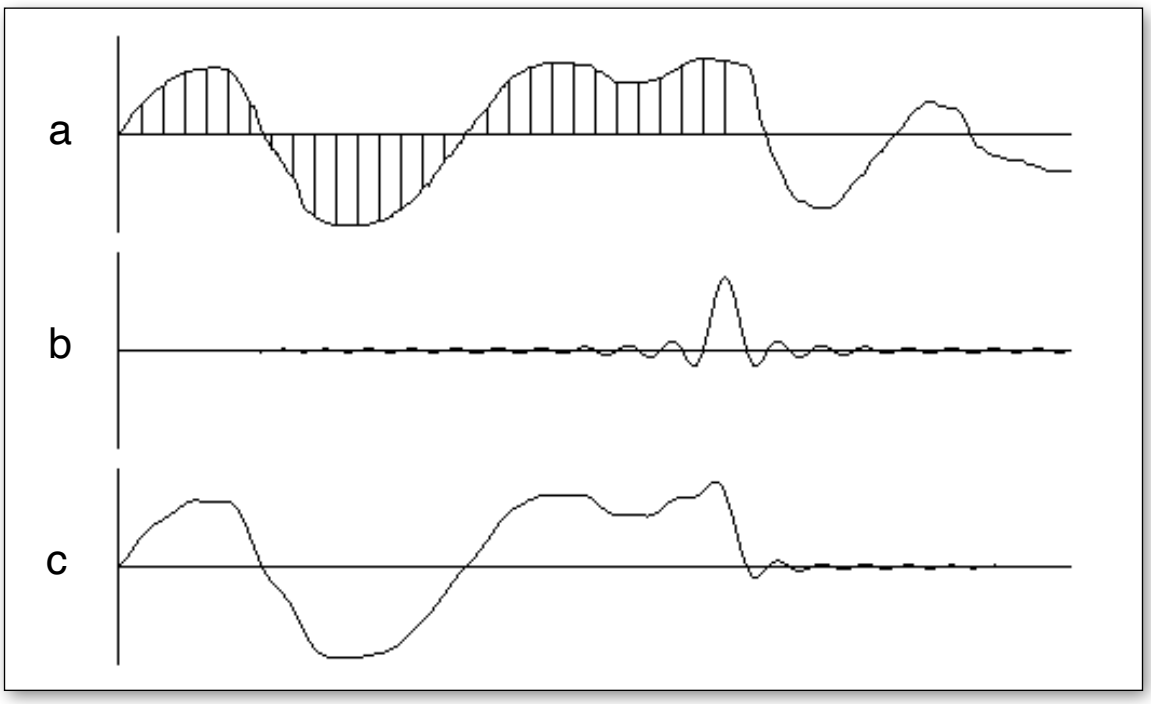


Fig. 1.6
The values of the original function (a) are taken at indicated “sampling points”, multiplied by sinc-functions and added up again, and displayed graphically in (c). The procedure has just arrived at point t_k on the time axis. The last sinc-function added is shown in (b).

We will later learn about the reciprocal of B to be the coherence time. So the sampling theorem tells us that one has to sample twice per coherence time.

2

Light

Although there is also optics of electrons and other radiations, the most important radiation for the realization of images is light. What does the physicist understand by light? We will answer this question step by step, and we will give different answers to it. Here is the first one: light is a kind of substance. It has much in common with a material gas. If you put it in a container, Fig. 2.1, it takes up the entire volume of the container. If you make a hole in such a "radiation cavity", light flows out. One can arrange the opening in such a way that a narrow bundle develops. Just like other gases, light has pressure, volume, energy, entropy and often a temperature.

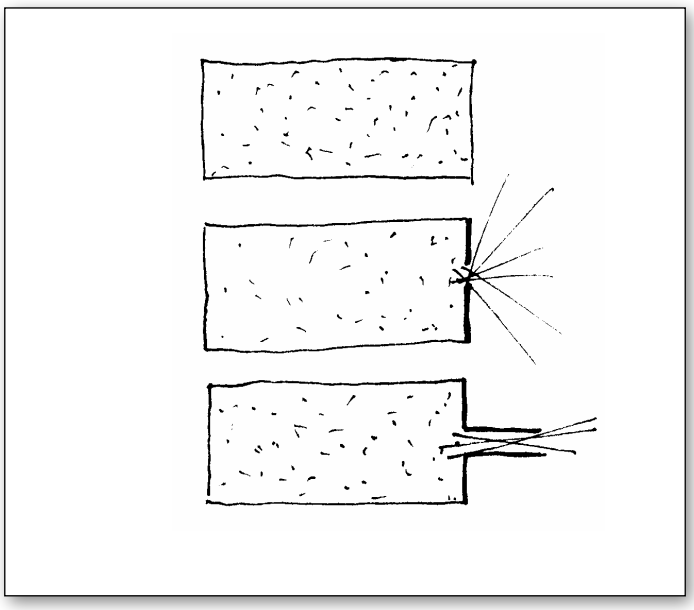


Fig. 2.1
(a) The light gas is confined in a container. (b) The light gas exits through an opening. (c) A light beam is created.

One can therefore regard light as one of many gases. Just as there is an oxygen gas, an electron gas or a neutron gas, there is also a light gas.

Another answer to the question "What is light?" is: light is an electromagnetic field; light is a system described by Maxwell's equations; light is an electromagnetic wave. However, not all of the solutions of Maxwell's equations are called light, e.g. not static electric or magnetic fields. However, the transition between the fields that are called light and those that are no longer called light is blurred.

A typical example of light is the light that comes from the sun. One way to describe this light would be to give the electric and magnetic field strength as a function of position and time, $\mathbf{E}(\mathbf{r}, t)$ and $\mathbf{H}(\mathbf{r}, t)$. But these functions are so complicated that, firstly, it is impossible to specify them and, secondly, they would not be of much use. But which electromagnetic fields are of interest in optics? The solutions of Maxwell's equations that are dearest to the physicist working in the field of optics are the linearly polarized, monochromatic, plane waves that are almost never found in nature. When he is dealing with real light, he breaks it down – in thought or experiment – into such plane waves. And to characterize a particular type of light, he indicates how much of each different type of plane wave is contained in it. Before we get to know this characterization, we need to take a closer look at plane waves.

2.1 Plane waves

A special solution of the Maxwell equations is the “linearly polarized, monochromatic, plane wave”. Despite the long name, it is a very simple solution. The electric field strength ***E*** as a function of position **r** and time *t* is

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}_0 \cos(\omega t - \mathbf{kr} + \Phi)$$

The argument of the cosine function ($\omega t - \mathbf{kr} + \Phi$) is called phase. For $t = 0$ and $\mathbf{r} = 0$ it is equal to Φ. We call Φ the initial phase. If we are not interested in the initial phase, we set it to zero:

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}_0 \cos(\omega t - \mathbf{kr})$$

From the Maxwell equations we get the magnetic field strength

$$\mathbf{H}(\mathbf{r},t) = \mathbf{H}_0 \cos(\omega t - \mathbf{kr})$$

H is perpendicular to **E**, and for the magnitude of the field strengths we have

$$\sqrt{\varepsilon_0} \cdot |\mathbf{E}| = \sqrt{\mu_0} \cdot |\mathbf{H}|$$

E and **H** are therefore connected in a simple, unambiguous way. It is therefore often sufficient to consider only one of these two field strengths. We continue with the investigation of the wave.

$$\mathbf{E}(\mathbf{r},t) = \mathbf{E}_0 \cos(\omega t - \mathbf{kr}) \qquad (2.1)$$

$\mathbf{E}_0$ is the (vectorial) amplitude of the field strength. cos ($\omega t - \mathbf{kr}$) describes a harmonic plane wave propagating in the direction of the *k*-vector. The angular frequency

$$\omega = \frac{2\pi}{T}$$

describes how fast the cosine function oscillates at a fixed position **r**. The magnitude of the wave number vector **k** is a measure of the wavelength:

$$k = |\mathbf{k}| = \frac{2\pi}{\lambda}$$

From Maxwell’s equations it follows that $\mathbf{E}_0$ is perpendicular to **k**. The velocity at which a selected maximum, or a selected zero crossing, is moving, the so-called phase velocity, has the fixed value

$$c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}} \approx 3 \cdot 10^8 \text{ m/s}$$

c is related to ω and *k* according to

$$\boxed{c = \frac{\omega}{k}}$$

The direction of $\mathbf{E}_0$ is called the direction of polarization of the wave.

We summarize the meaning of the constants contained in (2.1):

Magnitude of $\mathbf{E}_0$: Amplitude of the wave
Direction of $\mathbf{E}_0$: Direction of polarization
Magnitude of **k**: Measure of the wavelength
Direction of **k**: direction of propagation of the wave
ω: Measure for the oscillation period

We can now understand the long name of the waves we have studied: plane, monochromatic, linearly polarized waves.

Plane: The wave has a single *k* vector.
Monochromatic: The wave has a single ω value.
Linearly polarized: The wave has a single polarization direction.

With the propagation of the wave an energy flow is associated. The energy flow density j (energy current strength per area) is

$$\mathbf{j} = \mathbf{E} \times \mathbf{H}$$

Since in our case **E** is perpendicular to **H**, and

$$\sqrt{\varepsilon_0} \cdot |\mathbf{E}| = \sqrt{\mu_0} \cdot |\mathbf{H}|$$

we obtain

$$\mathbf{j} = \sqrt{\frac{\varepsilon_0}{\mu_0}} E^2$$

and with

$$c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}}$$

it follows

$$j = c \cdot \varepsilon_0 \cdot E^2$$

j has the same direction as **k**.

Since the energy density is

$$\rho_E = \frac{\varepsilon_0}{2} |\mathbf{E}|^2 + \frac{\mu_0}{2} |\mathbf{H}|^2 = \varepsilon_0 |\mathbf{E}|^2$$

we can write

$$|\mathbf{j}_E| = \rho_E \cdot c$$

This is analogous to the relationship between electric charge density and electric current density:

$$\mathbf{j}_Q = \rho_Q \cdot \mathbf{v}$$

With $\mathbf{E} = \mathbf{E}_0 \cos(\omega t - \mathbf{kr})$ the time average of the energy current density becomes:

$$\overline{\mathbf{j}} = \frac{1}{2} \cdot \sqrt{\frac{\varepsilon_0}{\mu_0}} \cdot |\mathbf{E}_0|^2$$

Often it is useful to represent waves by complex numbers:

$$\mathbf{E} = \mathbf{E}_0 e^{i(\omega t - \mathbf{kr})}$$

However, only the real part has a physical meaning.

This notation has advantages when superimposing waves. Complex numbers can be easily added in the complex number plane: The numbers are represented by arrows and added graphically like vectors.

2.2 Superposition of two plane waves

Later we want to represent light as a superposition of plane waves. We begin by examining the simplest superposition that can be imagined: that of *two* plane waves.

$$\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2 = \mathbf{E}_{1,0}\, e^{i(\omega_1 t - \mathbf{k}_1\mathbf{r})} + \mathbf{E}_{2,0}\, e^{i(\omega_2 t - \mathbf{k}_2\mathbf{r} + \varphi)}$$

There are various possibilities.

Partial waves with different polarization directions

The two partial waves propagate in the *z*-direction, i.e. **kr** = *kz*. They are supposed to have the same frequency ω, and their amplitudes are supposed to be perpendicular to each other: $\mathbf{E}_{1,0}$ = ($E_{1,0}$, 0, 0) and $\mathbf{E}_{2,0}$ = (0, $E_{2,0}$, 0). Furthermore, they should be out of phase with each other by $\pi/2$. So it is

$$\mathbf{E}_1 = \begin{pmatrix} E_{1,0}\cdot\cos(\omega t - kz) \\ 0 \\ 0 \end{pmatrix} \text{ and } \mathbf{E}_2 = \begin{pmatrix} 0 \\ E_{2,0}\cdot\sin(\omega t - kz) \\ 0 \end{pmatrix}$$

Thus the resulting wave is

$$\mathbf{E} = \begin{pmatrix} E_{1,0}\cdot\cos(\omega t - kz) \\ E_{2,0}\cdot\sin(\omega t - kz) \\ 0 \end{pmatrix}$$

Such a wave is called elliptically polarized. For z = const, the **E** vector describes an ellipse in the *x*-*y* plane. If $E_{1,0} = E_{2,0}$, the ellipse becomes a circle and the wave is called a circularly polarized wave.

Partial waves with different frequencies

The two partial waves propagate in the *z*-direction, the polarization direction of both waves is the *x*-direction, the frequencies are $\omega + \Delta\omega$ and $\omega - \Delta\omega$:

$$E_1 = E_{1,0}\cdot\cos\left[(\omega - \Delta\omega)\left(t - \frac{z}{c}\right)\right], \quad E_2 = E_{2,0}\cdot\cos\left[(\omega + \Delta\omega)\left(t - \frac{z}{c}\right)\right]$$

The resulting wave is

$$E = (E_{1,0} + E_{2,0})\cdot\cos\Delta\omega\left(t - \frac{z}{c}\right)\cdot\cos\omega\left(t - \frac{z}{c}\right)$$
$$+ (E_{1,0} - E_{2,0})\cdot\sin\Delta\omega\left(t - \frac{z}{c}\right)\cdot\sin\omega\left(t - \frac{z}{c}\right)$$

If z = const, a modulated oscillation is obtained, Fig. 1.5. If $E_{1,0} = E_{2,0}$, the wave is completely pinched off, it breaks down into *wave trains* or *wave packets*.

Partial waves with different directions of propagation

The two partial waves have the same amplitude, and the same frequency, and they are both polarized in the *x*-direction. Their directions of propagation, however, are inclined in the *y*-*z* plane away from the *z*-direction by equal and opposite angles. We use the complex notation:

$$E_1 = \mathrm{Re}\left[E_0 e^{i(\omega t - k_z z + k_y y)}\right] \quad \text{and} \quad E_2 = \mathrm{Re}\left[E_0 e^{i(\omega t - k_z z - k_y y)}\right] \tag{2.2}$$

The resulting wave is

$$E = \mathrm{Re}\left[E_0 e^{i(\omega t - k_z z)}\cdot(e^{ik_y y} + e^{-ik_y y})\right] = \mathrm{Re}\left[2E_0\cdot\cos(k_y y)\cdot e^{i(\omega t - k_z z)}\right]$$

$$E = 2E_0\cdot\cos(k_y y)\cdot\cos(\omega t - k_z z) \tag{2.3}$$

This is a plane wave moving in the *z*-direction, which is spatially modulated in the *y*-direction. At positions with $k_y y = (n/2)\cdot\pi$, with n = 0, 2, 4, 6, ... its amplitude is $2E_0$, i.e. twice as large as that of the single waves. At positions with $k_y y = (n/2)\cdot\pi$, with n = 1, 3, 5,... the amplitude is equal to zero. This phenomenon is called *interference*. At some points the interference is *constructive*, we have amplification, at others it is *destructive*, we have cancellation. The time average of the energy flux density is

$$\bar{j} = \frac{c\varepsilon_0}{2} 4E_0^2\cdot\cos^2(k_y y) = 2c\varepsilon_0 E_0^2\cdot\cos^2(k_y y)$$

Interference is a phenomenon for which we have no experience whatsoever in dealing with ordinary light. After all, it says the following: At a certain place a light wave number 1 arrives. So energy also arrives there, and “it is bright”. Now we take the light wave 1 away and let another light wave 2 run to the place and again it’s bright. But if we now let both light waves 1 and 2 run at the same time, the flow of energy to the place under consideration disappears, it is dark there. The reason why we have almost no experience with this is that the interference of light can be disturbed very easily.

We consider two plane waves as in equation (2.2), but allow that there is a phase shift between them, which changes with time. (This is equivalent to the fact that we no longer have purely harmonic waves). We can take this into account by adding the phase angle $\varphi(t)$ to the modulation factor in (2.3):

$$E = 2E_0\cdot\cos(k_y y + \varphi(t))\cdot\cos(\omega t - k_z z) \tag{2.4}$$

The positions y, for which $\cos(k_y y + \varphi(t)) = 0$, now move back and forth with time according to $\varphi(t)$, and if they move fast, they can no longer be recognized.

The time average of the energy flux density of the wave (2.4) is

$$\bar{j} = c\varepsilon_0\overline{E^2} = c\varepsilon_0 E_0^2$$

It is simply equal to the sum of the energy flux densities of the individual waves.

To put it somewhat loosely, we can summarized:

If one has interference, one must add the field strengths. If there is no interference, one has to add the energy current densities.

2.3 Distributions of plane waves

Light coming from any light source can be thought of as composed of linearly polarized, monochromatic, plane waves. Depending on the light source and – for a given light source – depending on the position under consideration, this composition is different. In general, waves of different polarization directions, frequencies and *k*-vectors will contribute to the light. One can characterize the light by giving the following information:

(1) the distribution of polarization directions
(2) the distribution of frequencies (the spectrum)
(3) the distribution of *k*-vector directions.

Because $c = \omega/k$ the frequency is equivalent to the magnitude of the *k*-vector. Points (2) and (3) together therefore specify the entire distribution of the *k*-vectors.

The degree of polarization

If the light contains waves of all polarization directions, it is said to be unpolarized.

One can divide the total energy current density *j* into a linearly polarized part j_p and an unpolarized part j_u:

The degree of polarization *V* is

$$V = \frac{j_p}{j_p + j_u}$$

A polarizing filter is transparent for light of one polarization direction and opaque for light of the perpendicular polarization direction.

If polarized light of the energy flux density j_0 is incident on a polarization filter whose transmission direction is rotated by the angle Θ against the polarization direction of the light, the fraction

$$\frac{j}{j_0} = \cos^2\theta$$

will pass through the filter.

If completely unpolarized light of energy flux density j_0 is incident on the filter, the energy flux density *j* behind the filter is just half the value it is in front of the filter

$$\frac{j}{j_0} = \frac{1}{2}$$

because the unpolarized light can be seen as a mixture of waves of different polarization directions Θ, and the average value of $\cos^2\Theta$ over all angles is 1/2.

With a polarization filter one can determine the degree of polarization of light: One lets the light shine on the filter and twists the filter over an angular range $\Delta\Theta = \pi$. The energy flow of the transmitted light takes on a maximum value j_{max} and a minimum value j_{min}. Now $j_{max} - j_{min}. = j_p$ and therefore

$$V = \frac{j_{max} - j_{min}}{j_{max} + j_{min}}$$

Coherence

The distribution of the *k*-vectors of light is best represented in *k*-space. We consider light propagating in the *z*-direction and in adjacent directions. To characterize the light, we draw in *k*-space the region where the endpoints of those *k*-vector arrows are located, that represent the largest part of the light. Figure 2.2 shows a section through the *k*-space and through this region. In this cross-section the region appears as a surface area.

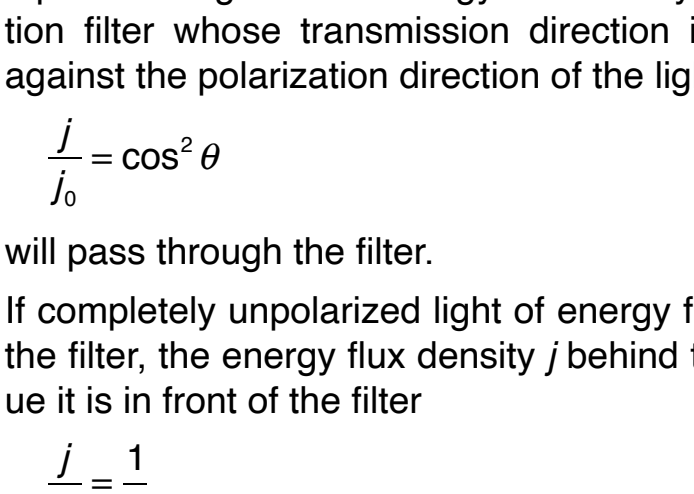


Fig. 2.2
Light distribution, shown in *k*-space

Of course, it is generally not possible to specify an exact boundary of this region. But it is possible, for example, to place the boundary line so that the *k*-vectors within the enclosed area describe 90% of the total light. Or one could draw level lines in the image, i.e. the 10%, 20% line, etc.

For a plane, monochromatic wave, the area shrinks to one point, Figure 2.3a. The larger the area that the light occupies in *k*-space, the more it deviates from such a wave.

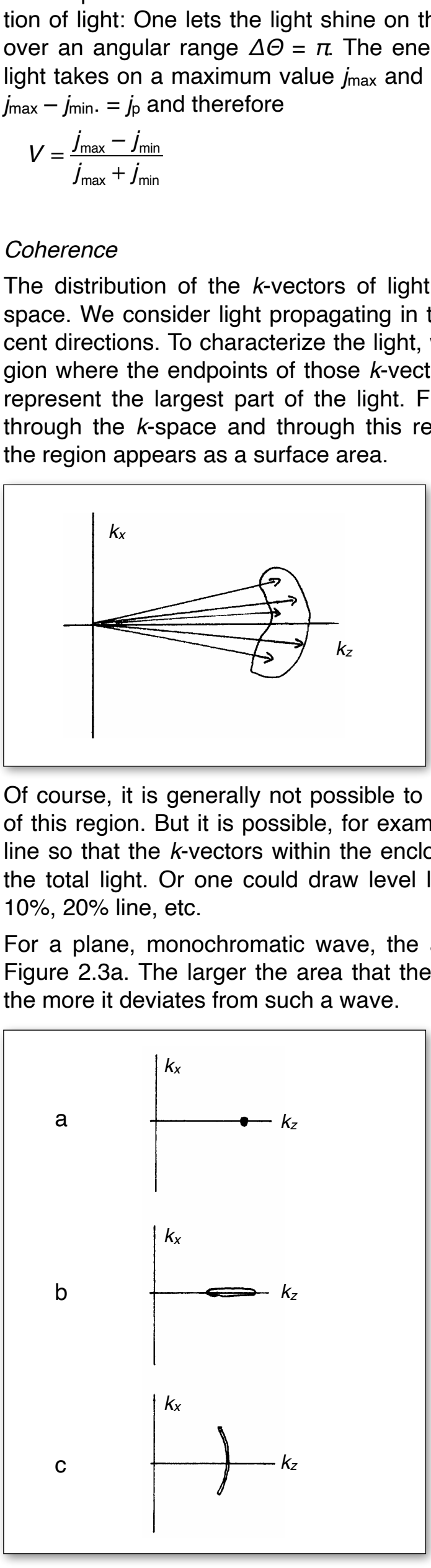


Fig. 2.3
Light distribution in *k*-space for
(a) coherent light
(b) spatial coherent light
(c) temporal coherent light

Fig. 2.3b shows the distribution for a wave that is plane but not monochromatic. The *k*-vectors of its harmonic components all have the same direction, but the magnitudes are different. With $\omega = c \cdot k$ the wave also has a wide frequency range. If we superpose plane waves of different frequencies, which all lie in a narrow frequency range of the width $\Delta\omega$, we obtain a wave that consists of wave trains (cf. section 2.2), Fig. 2.4.

Fig. 2.4
A wave containing Fourier components of different frequencies consists of "wave trains".

These wave trains have an average length of $2\pi/\Delta k$ and they have a period of

$$\frac{2\pi}{\Delta\omega} = \frac{2\pi}{c\Delta k}$$

There is a well-defined phase relationship between the spatial parts of such a wave train. The smaller the interval Δk (or $\Delta\omega$), the longer the wave trains or, as we say, the greater the *temporal coherence*. $2\pi/\Delta k$ is the *coherence length* of the wave and $2\pi/\Delta\omega$ the *coherence time*.

Figure 2.3c shows the distribution for a wave that is monochromatic but not plane. The *k*-vectors have a well-defined magnitude, but their directions are spreading. In such a wave, there are spatial beats transverse to the direction of propagation, see section 2.2. The narrower the angular range in *k*-space, the wider the coherent wave fronts or, as we say, the greater the *spatial coherence*.

In the same way that polarized light can be obtained from unpolarized light by filtering out the light with the "wrong" direction of polarization, one can also get coherent light from incoherent light by filtering out the light with the "wrong" *k*-vectors. And just as with polarization, there are different methods or tricks for doing this.

The simplest way to reduce the frequency range is to use an optical filter. Devices in which both the frequency interval and the average frequency can be set arbitrarily are called *monochromators*.

The angular dispersion of light can be reduced in two very simple ways: Either one moves away from the light source, or one blocks out the light of the wrong direction. Thus, the light from a fixed star (at the location of the earth) is spatially very coherent.

The distributions in Figures 2.2 and 2.3 correspond qualitatively to the following types of light:

Fig. 2.2 Sunlight
Fig. 2.3a Laser light
Fig. 2.3b Light from a star
Fig. 2.3c Light from a spectral lamp (close to the lamp)

Table 2.1 contains some typical numerical values.

Table 2.1

	Frequency interval (1/s)	Coherence time (s)	Coherence length (m)	Solid angle (sterad)	Width of wave trains (m)
Sun	$3 \cdot 10^{15}$	$3 \cdot 10^{-16}$	10^{-7}	10^{-4}	$2{,}5 \cdot 10^{-5}$
Spectral lamp	$5 \cdot 10^{9}$	$2 \cdot 10^{-10}$	$6 \cdot 10^{-2}$	$5 \cdot 10^{-4}$*	10^{-5}*
Argon laser	$5 \cdot 10^{6}$	$2 \cdot 10^{-7}$	60	10^{-8}	$0{,}5 \cdot 10^{-2}$

* at a distance of 0.5 m

2.4 Spherical waves

Besides the plane wave, the spherical wave is an important wave type for optics. While a longitudinal spherical sound wave has a very simple structure, the spherical electromagnetic wave is a complicated phenomenon: it is described in the lecture about electromagnetism in the context of the Hertzian dipole. Since the **E** and **H** vectors are transverse to the direction of propagation, this wave cannot have full spherical symmetry. The wave surfaces, i.e. the surfaces of constant phase, are indeed spherical surfaces. However, the magnitudes of the electric and magnetic field strength, as well as the energy flux density, are direction-dependent. This is due to the fact that the radiating dipole has a specific direction in space.

However, one can imagine that from one point radiating dipoles of different orientation emit wave trains in quick succession. In this case, the temporal average of the energy flux density is spherically symmetrical, and the wave can be treated like a scalar wave.

3

Light in matter

When light propagates in a material medium, an interaction between light and matter takes place. The influence of matter on light becomes easily understandable if one thinks of the light broken down into monochromatic, plane, polarized waves: Matter acts on each such component in a characteristic way. One could also say that it decomposes the light into these components.

The study of the interaction between light and matter has two objectives:

- In order to realize optical images one has to manipulate light by means of material arrangements.
- The light is a means to study the structure of matter. Such investigations are a subject of solid state physics and atomic physics.

3.1 The optical constants

Three things happen with a light wave that is sent on its way in matter:

- Its phase velocity is different in matter than in vacuum.
- Its amplitude decreases in the direction of propagation and the wave is absorbed.
- Its direction of polarization is rotated.

Each of the three effects is described by a material constant: the first by the *refractive index n*, the second by the *absorption index* κ and the third by the *specific rotation*. Actually these "material constants" are not constants at all, because their values depend on the frequency. They are functions of the frequency.

In addition, the optical properties can also depend on the direction of polarization and the direction of the *k*-vector. They are then no longer described by scalars, but by tensors. However, we start with the study of optically isotropic substances, i.e. substances whose optical constants are scalars.

For the quantities ω and k appearing in the solution of the Maxwell equation

$$E(x,t) = E_0\, e^{i(\omega t - kx)} \tag{3.1}$$

the relation

$$\frac{\omega}{k} = \frac{1}{\sqrt{\varepsilon_0 \mu_0}} = c$$

only applies as long as $\varepsilon = 1$, $\mu = 1$ and $\sigma = 0$, i.e. for the vacuum. In matter these conditions are no longer fulfilled. Nevertheless, one can still make a solution approach like (3.1) for matter, but then one gets another relationship between ω and k. In particular, it can happen that k becomes complex, i.e.

$$k = k_1 - ik_2 \tag{3.2}$$

where k_1 and k_2 are real numbers. We want to investigate how such solutions differ from those in the vaccum. For this purpose we insert (3.2) in (3.1):

$$E(x,t) = E_0 \cdot e^{-k_2 x} \cdot e^{i(\omega t - k_1 x)} \tag{3.3}$$

This is a wave with an exponentially decreasing amplitude. Its phase velocity is

$$v_{\mathrm{ph}} = \frac{\omega}{k_1}$$

The ratio between the phase velocity in vacuum and that in matter is called the refractive index n of the matter

$$n = \frac{c}{v_{\mathrm{ph}}}$$

So the refractive index is related to ω and k_1 according to

$$n = \frac{c}{\omega} k_1 \tag{3.4}$$

Also the energy flux density of such a wave decreases exponentially with x. For the time average $\overline{j}$ of j applies

$$\overline{j} = \overline{j}_0 \cdot e^{-ax} \tag{3.5}$$

a is called the *absorption coefficient*. Since j is quadratic with the field strength **E** we have:

$$a = 2k_2$$

It is convenient to define a complex refractive index n'. In (3.4) instead of the real part k_1 we insert the whole complex k

$$n' = \frac{c}{\omega}(k_1 - ik_2) = n\left(1 - i\frac{k_2}{k_1}\right)$$

The quotient

$$\kappa = \frac{k_2}{k_1} = \frac{a}{2k_1} \tag{3.6}$$

is called the *absorption index* of the medium. Thus we have

$$n' = n(1 - i\kappa) \tag{3.7}$$

With $k_1 = 2\pi/\lambda$ we get

$$\kappa = \frac{a\lambda}{4\pi}$$

κ has an obvious physical meaning: From (3.5) it follows that the reciprocal value of a can be understood as the penetration depth of light in matter. The absorption index is therefore a measure of the penetration depth per wavelength.

If a wave runs from a medium with the refractive index n_a into a medium with a different refractive index n_b, its frequency does not change. Therefore, it follows from (3.4) that n and k_1 are unambiguously related for the wave. It is therefore often useful not to consider ω and k (or k_1) as independent parameters, but ω and n. The electric field strength of a plane wave running in x-direction is therefore often written like as

$$E(x,t) = E_0 \cdot e^{i\omega\left(t - \frac{n}{c}x\right)}$$

If the medium is absorbing, i.e. $\kappa \neq 0$, it is sufficient to use the complex refractive index n' instead of n. With (3.7), (3.6) and (3.4) one obtains again (3.3).

3.2 Frequency, direction and polarization dependence of the optical constants

Since (3.1) is a solution of the Maxwell equations, the optical constants n and κ clearly depend on the material quantities ε, μ and σ that appear in the Maxwell equations. From the fact that ε, μ and σ depend on the frequency, the direction of propagation and the polarization of the wave it follows that n and κ also show such dependencies. To relate these dependencies to the structure of matter is an important research topic in solid state physics and atomic physics.

Because n is a function of ω, a wave train of finite length, which contains Fourier components of different frequencies, will run apart on its way. This process is called *dispersion*. Usually n grows with ω, (decreases with λ) Fig. 3.1. This is called normal dispersion. In frequency ranges in which n decreases with increasing ω there is anomalous dispersion.

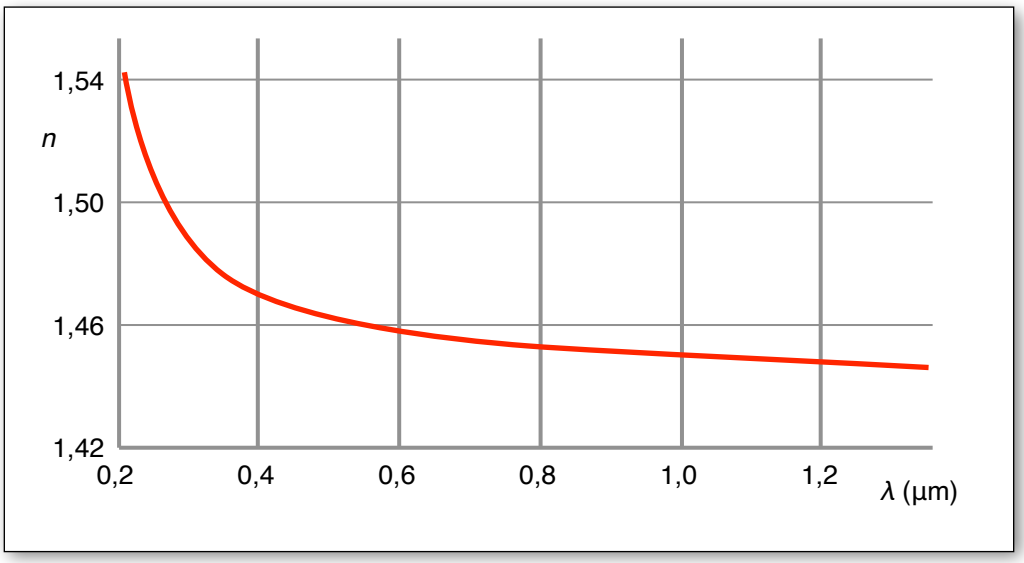


Fig. 3.1
Refractive index for Quartz glass as a function of the wavelength. For normal dispersion, the refractive index decreases with increasing wavelength.

Anomalous dispersion is always accompanied by absorption.

In addition, n and κ in substances of sufficiently low symmetry also depend on both the direction of propagation and the direction of polarization. Therefore, and due to the fact that solids can have a wide variety of symmetries, a large number of different effects result.

If the refractive index depends on the direction of wave propagation, it also automatically depends on the direction of polarization. Crystals for which this is the case are called *birefringent*. If the absorption index depends on the direction of polarization, it is called *dichroism*.

An optically isotropic substance can also be made anisotropic “from the outside”, for example by

- a mechanical stress
- an electric field
- a magnetic field.

Mechanical stress leads to stress birefringence. The birefringence caused by an electric field is known as the *Kerr effect*. A magnetic field causes birefringence if the field strength is transverse to the direction of propagation of the light (Cotton-Mouton effect) or a rotation of the plane of polarization if the light propagates in the direction of the field (Faraday effect).

Further effects occur when crystals that are already anisotropic in themselves are brought into external fields.

3.3 The group velocity

So far in the context of waves we have been dealing with the phase velocity. However, the phase velocity is not a dynamic physical quantity, but a kinematic one. It does not describe the movement of a physical object, but only the movement of a geometric point, such as the zero crossing of the electric field strength in a wave. It is no more similar to a physical movement than the movement of a "car" on a cinema screen, which is only the movement of a shadow.

But with a light flux a real physical transport is associated, namely the transport of energy, momentum, entropy and other substance-like quantities. Does it make sense to describe this transport by a velocity? It makes sense at least as long as the transport has a temporal (and thus a spatial) beginning and an end, Fig. 3.2. Because then the energy and the other quantities are localized, they are in the region of x_1 at a time t_1 and in the region of x_2 at a time t_2. From this one can calculate a transport velocity

$$v = \frac{x_2 - x_1}{t_2 - t_1}$$

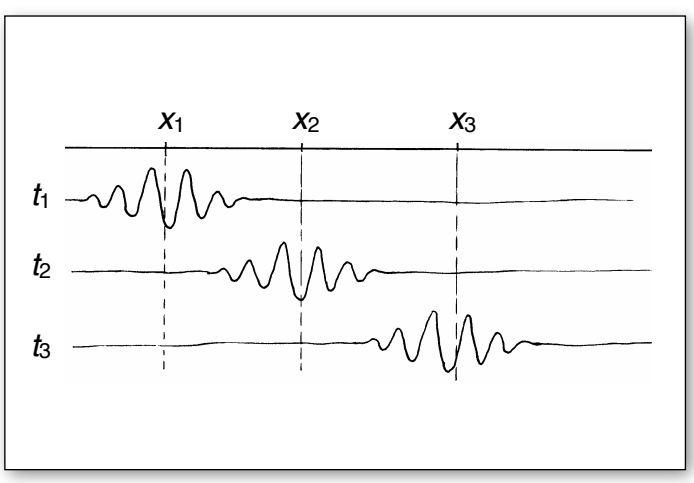


Fig. 3.2
Wave packet at three instances of time. A velocity can be assigned to the transport.

If there is no dispersion, the wave packet moves without changing its shape, because all its Fourier components have the same phase velocity. The dynamic, or group velocity of the wave packet is therefore equal to the phase velocity. Things are different when dispersion is present. Then the partial waves of the wave packet run at different velocities than the packet as a whole. The phase velocity of the partial waves can be higher than c, but the whole packet always runs with a (dynamic) velocity $v \leq c$.

If one looks at the representation of a wave packet, where only the electric field strength is plotted, one could suspect a contradiction, Fig. 3.3.

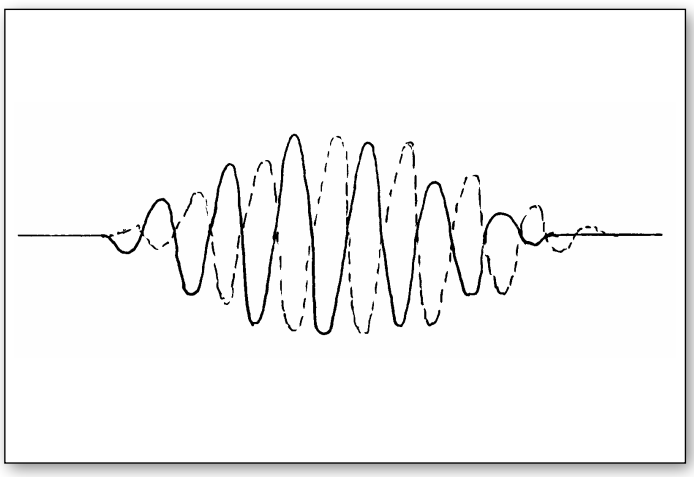


Fig. 3.3
The phase velocity can be greater than the group velocity.

If the maxima within the wave packet run faster than the whole packet, doesn't the energy within the packet have to run at the high velocity as well? But then what happens to it when it reaches the front end of the packet? One can see that in this case the energy density cannot simply be equal to $\varepsilon_0 \mathbf{E}^2$ as it is the case for a wave in a vacuum. This is only true as long as electric and magnetic field strength are in phase. So it follows that **E** and **H** can no longer be in phase as soon as dispersion is present.

Let us calculate the group velocity for a simple special case: the case that the wave group has only two harmonic components with frequencies close together. The total wave is then a kind of beat, a sequence of wave packets, Fig. 1.5.

For the maximum of a wave group, the phases of the two partial waves coincide:

$$\omega_1 t - k_1 x = \omega_2 t - k_2 x$$

or

$$(\omega_2 - \omega_1)\, t = (k_2 - k_1)\, x$$

The velocity at which the group propagates is therefore

$$v_{gr} = \frac{x}{t} = \frac{\omega_2 - \omega_1}{k_2 - k_1}$$

or, as the frequencies should be close together

$$\boxed{v_{gr} = \frac{d\omega}{dk}}$$

If ω is not linearly dependent on k, the velocity of the wave group is no longer the same as the phase velocity of the harmonic waves into which it can be decomposed. Of course, this fact has the consequence that the wave packet will spread apart during its movement.

4

Light at interfaces: Reflection and refraction

4.1 Law of Reflection and Refraction

As is well known, light that passes from one homogeneous material to another is refracted and reflected, Fig. 4.1. The transition from one material to the other must take place in a layer whose thickness is $\ll \lambda$. In this case the law of reflection applies

$$\alpha = \alpha' \tag{4.1}$$

and the law of refraction

$$n_a \sin \alpha = n_b \sin \beta \tag{4.2}$$

apply. α and α' are the angles between the wave normal of the incident or reflected wave and the normal of the boundary surface. β is the angle between the normal of the refracted wave and that of the the boundary surface, and n_a and n_b are the refractive indices of the two materials.

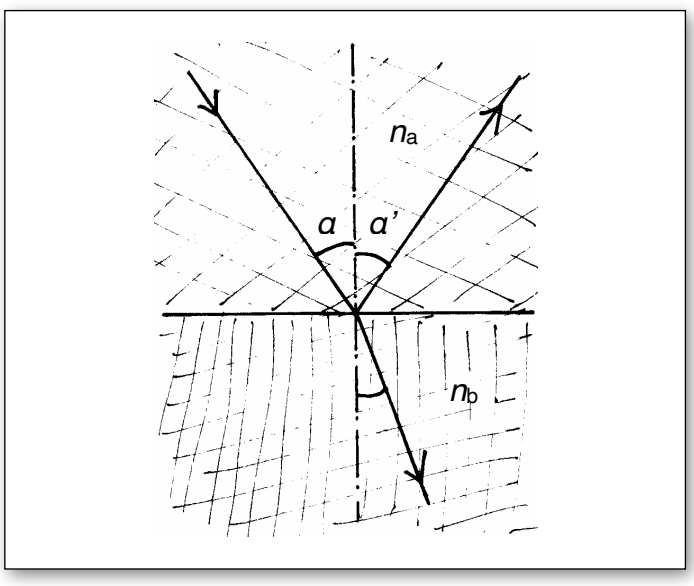


Fig. 4.1
The light wave is reflected and refracted.

The wave front normals of the incident, reflected and refracted light, as well as the boundary surface normal lie in one plane, the plane of incidence.

The two laws (4.1) and (4.2) give information about the direction of the outgoing waves, if one knows the direction of the incoming wave and the refractive indices. They don't tell us which fraction of the light is reflected and which is refracted. This is done by the *Fresnel equations*, which will be discussed in the next section.

We mark quantities that refer to the three waves as follows:

incident wave: index i
reflected wave: index r
refracted (transmitted) wave: index t

The laws of reflection and refraction can be easily derived for monochromatic plane waves. At the surface, the phases of the three waves can only differ by a constant value, i.e. for all time instants t and all positions $\boldsymbol{r}_G$ on the interface must apply:

$$\omega_i t - \mathbf{k}_i \mathbf{r}_G = \omega_r t - \mathbf{k}_r \mathbf{r}_G + \Phi_r = \omega_t t - \mathbf{k}_t \mathbf{r}_G + \Phi_t \tag{4.3}$$

From the fact that this chain of equations must apply to any fixed location $\boldsymbol{r}_G$ for any instant of time, it follows

$$\omega_i = \omega_r = \omega_t$$

So all three waves have the same frequency. The fact that (4.3) must be valid for a given instant of time for every position $\mathbf{r}_G$ of the interface is equivalent to

$$(\mathbf{k}_i - \mathbf{k}_r)\mathbf{r}_G = \text{const for every } \mathbf{r}_G$$

and

$$(\mathbf{k}_i - \mathbf{k}_t)\mathbf{r}_G = \text{const for every } \mathbf{r}_G$$

These relations are fulfilled when $(\mathbf{k}_i - \mathbf{k}_r)$ and $(\mathbf{k}_i - \mathbf{k}_t)$ are perpendicular to the interface. This means that the components of the k-vectors parallel to the interface

$$k_{\parallel i},\ k_{\parallel r} \text{ and } k_{\parallel t}$$

must be equal to each other, Fig. 4.2.

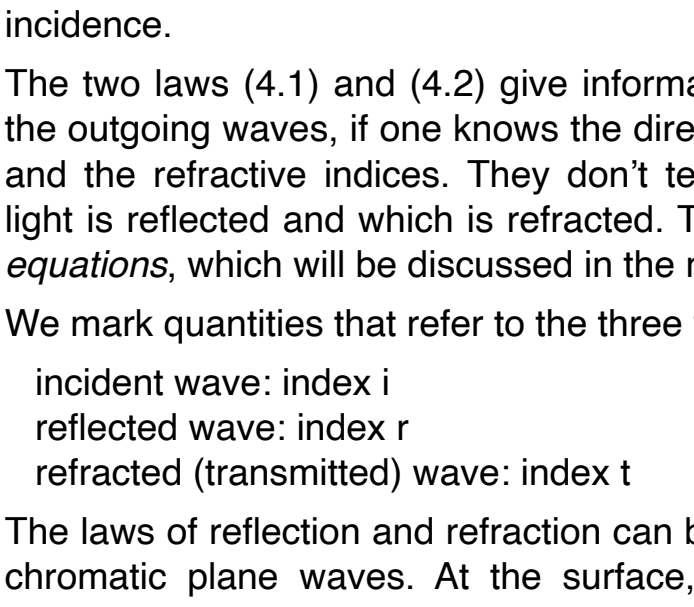


Fig. 4.2
(a) A light wave is reflected and refracted on a plane. (b) The components of the k-vector which are parallel to the plane are equal.

With

$$k_{\parallel i} = |k_i| \sin\alpha$$

and

$$k_{\parallel t} = |k_t| \sin\beta$$

we get

$$|k_i| \sin\alpha = |k_t| \sin\beta$$

and with

$$\frac{|k_i|}{|k_t|} = \frac{\frac{\omega n_a}{c}}{\frac{\omega n_b}{c}} = \frac{n_a}{n_b}$$

we finally obtain

$$n_a \sin\alpha = n_b \sin\beta$$

4.2 The Fresnel equations

The question of how much of a light flux hitting an interface is reflected and how much is refracted can also be answered solely on the basis a knowledge of the refractive indices. However, the result depends on how the incident light is polarized. It is therefore useful to decompose the *E*-vector into a component $E_\perp$ which is perpendicular to the plane of incidence and a component $E_\parallel$ which lies in the plane of incidence, Fig. 4.3.

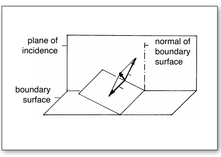


Fig. 4.3
Appropriate decomposition of the vector of the electric field strength

Using the Maxwell equations, a somewhat laborious calculation yields the reflection coefficients $r_\perp$ and $r_\parallel$, and the transmission coefficients $t_\perp$ and $t_\parallel$:

$$r_\perp = \frac{E_{r\perp}}{E_{i\perp}} = \frac{\cos\alpha - \sqrt{n^2 - \sin^2\alpha}}{\cos\alpha + \sqrt{n^2 - \sin^2\alpha}} \tag{4.4}$$

$$t_\perp = \frac{E_{t\perp}}{E_{i\perp}} = \frac{2\cos\alpha}{\cos\alpha + \sqrt{n^2 - \sin^2\alpha}} \tag{4.5}$$

$$r_\parallel = \frac{E_{r\parallel}}{E_{i\parallel}} = -\frac{n^2\cos\alpha - \sqrt{n^2 - \sin^2\alpha}}{n^2\cos\alpha + \sqrt{n^2 - \sin^2\alpha}} \tag{4.6}$$

$$t_\parallel = \frac{E_{t\parallel}}{E_{i\parallel}} = \frac{2n\cos\alpha}{n^2\cos\alpha + \sqrt{n^2 - \sin^2\alpha}} \tag{4.7}$$

Here α is the angle of incidence of the incoming wave and n_b/n_a is abbreviated *n*.

Fresnel had already derived these equations in 1821 using his mechanical theory of light. They are called *Fresnel's equations*.

Since the components $E_{i\perp}$, $E_{r\perp}$ and $E_{t\perp}$ are not parallel to each other, there is no uniform, natural way to define the signs. The signs in both equations (4.6) and (4.7) correspond to directions of positive counting indicated by arrows in Figure 4.4.

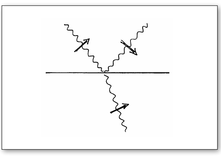


Fig. 4.4
To define the sign in the Fresnel equations

Figure 4.5 shows the curve of the four coefficients (4.4) to (4.7) as a function of the angle of incidence α for the case that $n = n_b/n_a = 1.7$. This corresponds approximately to the transition from air ($n_a = 1$) into glass (the refractive index is between 1.45 and 1.9 depending on the type of glass).

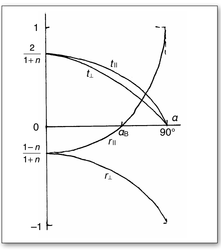


Fig. 4.5
Coefficients of reflection and transmission as a function of the angle of incidence for $n_b > n_a$

Let us discuss these curves.

1. For $\alpha = 0$ we have

$$r_\perp = r_\parallel = \frac{1-n}{1+n}$$

and

$$t_\perp = t_\parallel = \frac{2}{1+n}$$

2. The greater the difference $n_b - n_a$ of the refractive indices, the more light is reflected.

3. for $\alpha \to 90°$, i.e. for grazing incidence, all light is reflected.

4. While the phase of the transmitted light is the same as that of the incident light, the $E_\perp$ component makes a phase jump of π.

5. $E_{r\parallel}$ displays the most interesting curve shape, Fig. 4.6. For angles of incidence smaller than the *Brewster angle* α_B, $E_{r\parallel}$ makes a phase jump. At $\alpha = \alpha_B$ we have $E_{r\parallel} = 0$, and for larger angles $E_{r\parallel}$ is in phase with $E_{i\parallel}$.

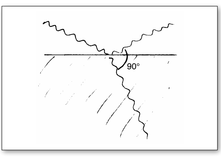


Fig. 4.6
If the light is incident at the Brewster angle, the directions of the reflected and refracted waves are orthogonal to each other.

Zeroing the numerator in (4.6) results in a condition for α_B

$$\tan\alpha_B = n$$

Furthermore, one finds

$$\alpha_B + \beta_B = 90°$$

If the light is incident at the Brewster angle, the reflected light is completely linearly polarized. The vector of the electric field strength is perpendicular to the plane of incidence.

If $n_a > n_b$, i.e. if the light passes from the material with the lower to that with the higher refraction index, a new phenomenon emerges: *total reflection*. Fig. 4.7 shows for this case the value of the four coefficients (4.4) to (4.7) as a function of the angle of incidence. Since $n = n_b/n_a < 1$, for angles of incidence with sin $\alpha > n$ the root

$$\sqrt{n^2 - \sin^2\alpha}$$

becomes imaginary. So the four coefficients become complex. The angle α_G in sin $\alpha_G = n$ is called *critical angle of total reflection*.

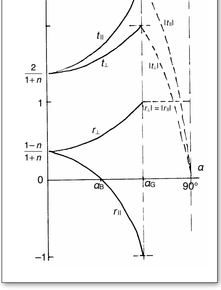


Fig. 4.7
Coefficients of reflection and transmission as a function of the angle of incidence for $n_a > n_b$

The contributions of $r_\perp$ and $r_\parallel$ are equal to 1, i.e. the reflected wave is only out of phase with the incident wave. The magnitudes of $t_\perp$ and $t_\parallel$, however, are not equal to zero. This means that a wave penetrates into the medium with the lower refractive index. The wave fronts of this wave are perpendicular to the interface. However, their amplitude decays exponentially in the direction of the normal of the interface.

The Fresnel equations can still be applied even if the refractive indices are complex.

We consider the case that the incident wave travels in air ($n_a \approx 1$) and hits a metal surface ($n' = n(1 - i\kappa)$ = complex) vertically ($\alpha = 0$).

With (4.8) the reflection coefficient becomes

$$r_\perp = r_\parallel = \frac{1 - n(1 - i\kappa)}{1 + n(1 - i\kappa)}$$

The reflectance *R* indicates which fraction of the incoming energy flows away with the reflected light.

It is

$$R = rr^* = \frac{(1-n)^2 + (n\kappa)^2}{(1+n)^2 + (n\kappa)^2}$$

5

Diffraction

5.1 What is diffraction?

If an obstacle is placed in the way of a plane wave in such a way that part of each wave front passes it and part of it does not, it is found that the wave also runs behind the obstacle into the area from which the source cannot be seen, i.e. which is actually in the shadow. This phenomenon is called diffraction. It is said that the wave is diffracted at the obstacle. This means that it is deflected from the direction in which it would run without an obstacle.

For sound waves, diffraction is a common, everyday phenomenon. Although the effect is usually very weak in light, it plays an important role in optics.

5.2 The Huygens-Fresnel principle

The principle can be formulated at different levels of generalization. The more general the formulation, however, the more unwieldy it becomes. We choose a formulation whose validity is quite limited. On the other hand, it is very transparent, it is easy to use, and it is sufficient to solve the most important problems.

A monochromatic, plane wave hits a flat obstacle with openings in it. The plane of the obstacle is parallel to the wavefronts of the incoming wave. The Huygens-Fresnel principle allows to determine the light distribution behind the obstacle. It tells us that the light wave continues behind the obstacle as if a spherical wave emanated from each point of the opening. The amplitude of the light field at any point behind the obstacle is obtained by superimposing the contributions of all these spherical waves.

The principle can also be interpreted in this way: The light field behind the obstacle is the same, no matter whether a plane wave hits the obstacle or whether there are many emitters oscillating in phase at the positions of the openings.

The Huygens-Fresnel principle follows from the Maxwell equations. However, the derivation is complicated. Statements about boundary conditions must be included in this derivation, and approximations are made, namely

- the light amplitude immediately behind the obstacle is zero;
- the light distribution in the open areas of the obstacle is the same as if the obstacle were not there.

It is difficult to check mathematically whether these conditions are met with sufficient accuracy. We take as legitimation of the principle the fact that it provides a very good prediction of the outcome of optical experiments.

6

Scattering

6.1 What is scattering

So far, we have considered the wave propagation in homogeneous media, or at the interface from one homogeneous medium to another. Of course, “homogeneous” does not mean that the material is homogeneous down to the smallest dimensions. It simply means that the average values of the physical quantities over regions of the order of the wavelength of the radiation under consideration are independent of the position. If this restriction is abandoned, processes known as *scattering* are admitted. For example, light is scattered when it passes through a ground glass screen or when it is reflected by a sheet of white paper. On a clear day, sunlight is scattered by the air in the atmosphere, with the result that we see the sky not black but bright blue.

The simplest situation involving scattering is the following: A plane wave of any radiation hits a small obstacle, i.e. an obstacle that is small compared to the wavelength or of the same order of magnitude as the wavelength of the radiation.

However, scattering usually refers to a slightly different phenomenon: the light wave encounters an ensemble of many irregularly arranged obstacles.

Depending on the size, distribution and nature of the scatterers and depending on the wavelength of the light, other phenomena are observed. These often bear the name of their discoverer: *Rayleigh scattering*, *Mie scattering*, *Thomson scattering*, *Compton scattering*, *Raman scattering*, *Brillouin scattering*, etc. Scattering phenomena can be divided into two classes:

- elastic scattering: the frequency of the light does not change (examples: Rayleigh and Mie scattering)
- inelastic scattering: the frequency of the light changes (examples: Compton, Raman, Brillouin scattering).

6.2 Scattering as an irreversible process

A monochromatic, plane light wave hits a ground-glass screen, Fig. 6.1. The arrows in the figure represent the *k*-vectors of the light.

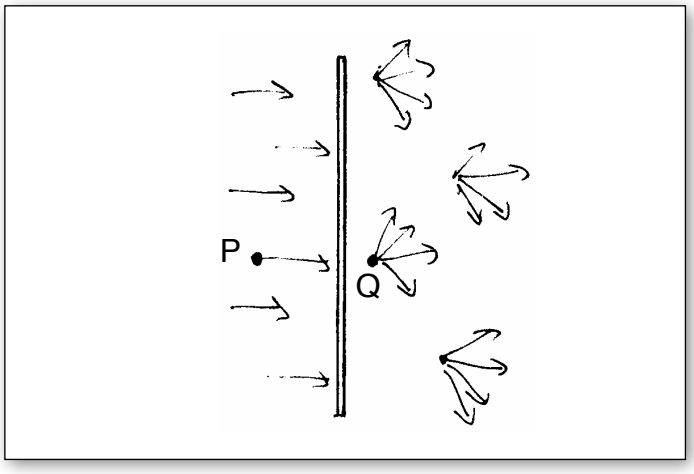


Fig. 6.1
Distribution of the *k*-vectors in front of and behind a ground-glass screen, displayed in position space

Fig. 6.2 shows the distribution of the *k*-vectors in *k*-space at a point P in front of and at a point Q behind the ground-glass screen. The magnitudes of the *k*-vectors, and thus the frequencies, are the same in front of and behind the screen: the scattering is elastic. However, the distribution of the directions of *k* has changed: the spatial coherence has strongly decreased.

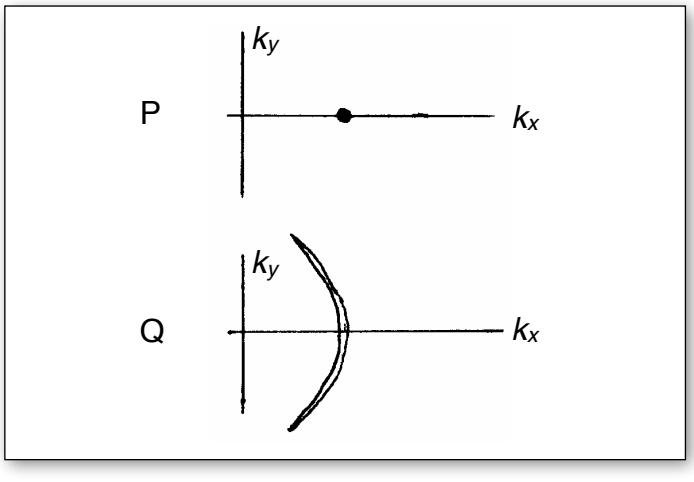


Fig. 6.2
Distribution of the k-vectors in front of and behind a ground-glass screen, displayed in *k*-space

There is no passive optical component (lens, mirror, ground glass ...) with which the scattering process can be reversed. Scattering is an irreversible process.

In thermodynamics, irreversible processes are described in a very simple and comprehensive way. There is one quantity that can be created but not destroyed: entropy. A process is always irreversible if entropy is produced. Reversing the process would require the annihilation of entropy, and that is forbidden. So scattering is a process in which entropy is produced.

6.3 Example: Rayleigh scattering

If a plane light wave hits a single molecule, the molecule becomes polarized. The polarization follows the electric field strength of the incident light, it changes according to $\sin(\omega t)$. This turns the molecule into a Hertzian oscillator and it emits a wave. The energy flux density of this wave is directional: It is zero in the direction of the dipole moment, i.e. in the direction of the electric field strength of the incident wave. Furthermore, it is proportional to the fourth power of the oscillation frequency.

The directional dependence has the consequence that the scattered light that runs away perpendicular to the direction of incidence is linearly polarized, Fig. 6.3. The frequency dependence has the consequence that blue light is scattered much more than red light.

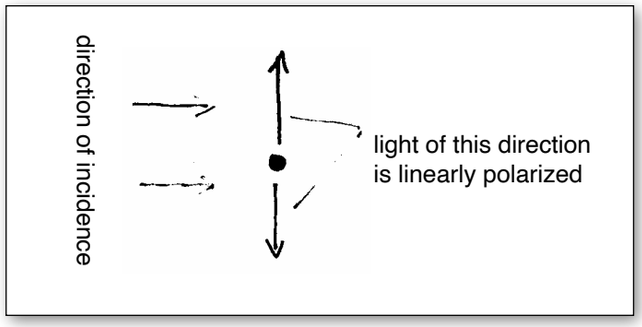


Fig. 6.3
Direction of propagation of linearly polarized light in Rayleigh scattering

These considerations were based on a single molecule.If the light wave now falls on many homogeneously distributed molecules, the scattering disappears, because for each molecule there is a second one at a distance of $\lambda/2$ transverse to the direction of incidence, whose scattered wave interferes with that of the first molecule, Fig. 6.4.

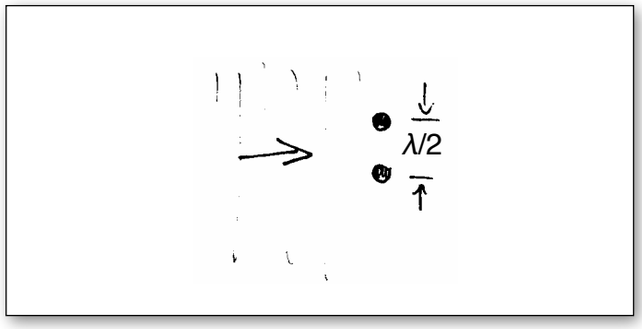


Fig. 6.4
On Rayleigh Scattering

Only when the scattering medium is no longer homogeneous does a scattering effect result again: i.e. when the density of the material changes over distances of the order of λ. Such density fluctuations are always present in gases. That is why gases show this scattering behavior. This scattering is called *Rayleigh scattering*.

Rayleigh scattering can be recognized by the following properties:

- the energy flux density of the scattered light goes with ω^4;
- the light scattered perpendicularly to the direction of the incident light is linearly polarized;
- the energy flux density of the scattered light is symmetrically distributed over the angle against the incident light: it is scattered equally in forward and backward direction.

The blue light of the unclouded sky is Rayleigh scattered light.

6.4 Example: Mie scattering

Things become much more complicated when the size of the scattering centers approaches the wavelength range of the light. The case in which the scatterers are spherical was treated quantitatively by Gustav Mie. The directional dependence of the energy flux density of the scattered light is complicated. However, a qualitative statement can be easily remembered: The larger the scattering centers are, the more light is scattered in the forward direction.

7

Interference phenomena

The superposition of plane, monochromatic waves leads to the phenomenon of interference: extinction of light at certain points, amplification at others (see section 2.2). In this chapter ee will investigate interference phenomena. In each of the experiments to be considered we will be dealing with two problems:

- What does the resulting wave field look like?
- Which trick is used to obtain the plane, monochromatic waves, i.e. the coherent light?

7.1 Elementary light beams

Light, whose *k*-vectors are distributed in the area $\Delta k_x \cdot \Delta k_y \cdot \Delta k_z$, forms spatial wave packets of the extension $\Delta x \cdot \Delta y \cdot \Delta z$ with

$$\Delta x \cdot \Delta k_x = 2\pi \qquad \Delta y \cdot \Delta k_y = 2\pi \qquad \Delta z \cdot \Delta k_z = 2\pi \tag{7.1}$$

As long as the light used for an experiment comes from a single packet, interference can be observed. The relations (7.1) are also called *coherence conditions*. We write them in yet another form. The light forms a beam, which essentially runs in the *z*-direction, Fig. 7.1.

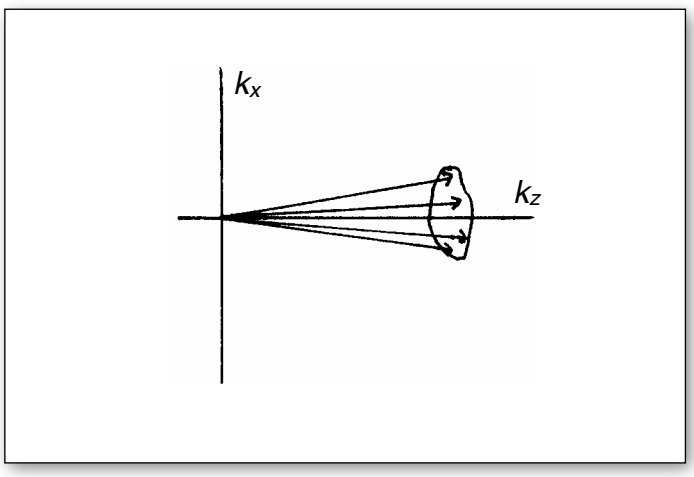


Fig. 7.1
Distribution of the *k*-vectors for light travelling essentially in the *z*-direction.

We then have

$$\Delta k_z \approx \Delta k = \frac{\Delta\omega}{c}$$

Thus the coherence conditions become

$$\Delta x \cdot \Delta k_x = 2\pi \qquad \Delta y \cdot \Delta k_y = 2\pi \qquad \Delta z \cdot \frac{\Delta\omega}{c} = 2\pi \tag{7.2}$$

Δz is the coherence length known from before. With $\Delta z/\Delta t = c$ we can replace it by the coherence time, and we get

$$\Delta x \cdot \Delta k_x = 2\pi \qquad \Delta y \cdot \Delta k_y = 2\pi \qquad \Delta t \cdot \Delta\omega = 2\pi$$

Instead of Δk_x and Δk_y one can also use the angular aperture of the *k*-vector distribution. With

$$\Delta k_x = k \cdot \sin(\Delta\varphi_x) = \frac{2\pi}{\lambda} \cdot \sin(\Delta\varphi_x)$$

and

$$\Delta k_y = k \cdot \sin(\Delta\varphi_y) = \frac{2\pi}{\lambda} \cdot \sin(\Delta\varphi_y)$$

we get

$$\Delta x \cdot \sin(\Delta\varphi_x) = \lambda \qquad \Delta y \cdot \sin(\Delta\varphi_y) = \lambda \qquad \Delta z \cdot \frac{\Delta\omega}{c} = 2\pi \tag{7.3}$$

In general, for any given light beam we have

$$\Delta x \cdot \Delta k_x > 2\pi \qquad \Delta y \cdot \Delta k_y > 2\pi \qquad \Delta z \cdot \Delta k_z > 2\pi$$

However, each beam can be broken down into partial beams defined by (7.1), so-called *elementary beams*.

Such a decomposition can be carried out in many ways: e.g. by taking the whole angular distribution of the *k*-vectors and obtaining very small spatial areas. Or one takes the whole light-filled space and splits the light into parts with very narrow *k*-distributions, or something in between. In the 6-dimensional *phase space*, which is formed by the three spatial and the three wavenumber coordinates, an elementary beam occupies a well-defined (6-dimensional) “volume”, namely

$$\Delta x \cdot \Delta k_x \cdot \Delta y \cdot \Delta k_y \cdot \Delta z \cdot \Delta k_z = (2\pi)^3$$

Figure 7.2 shows a two-dimensional section of the phase space. The whole light beam occupies the space bounded by the large rectangle. The partial images a, b and c show three different decompositions into elementary beams. The area of the projections of the elementary beams is the same in all three cases, namely 2π.

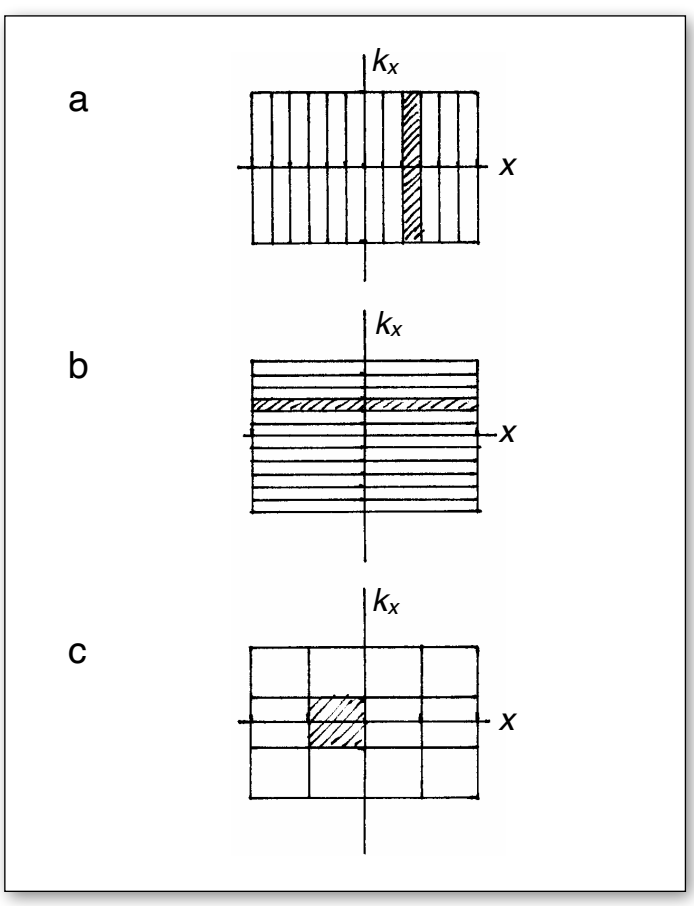


Fig. 7.2
Three different decompositions of light into elementary beams.

7.2 The interference pattern of two spherical waves

Since this case occurs frequently, we will look at it in more detail. Two spherical waves (cf. section 2.4) depart from two points P_1 and P_2, which are at a distance of 3λ, Fig. 7.3. The oscillations at the points P_1 and P_2 are in phase.

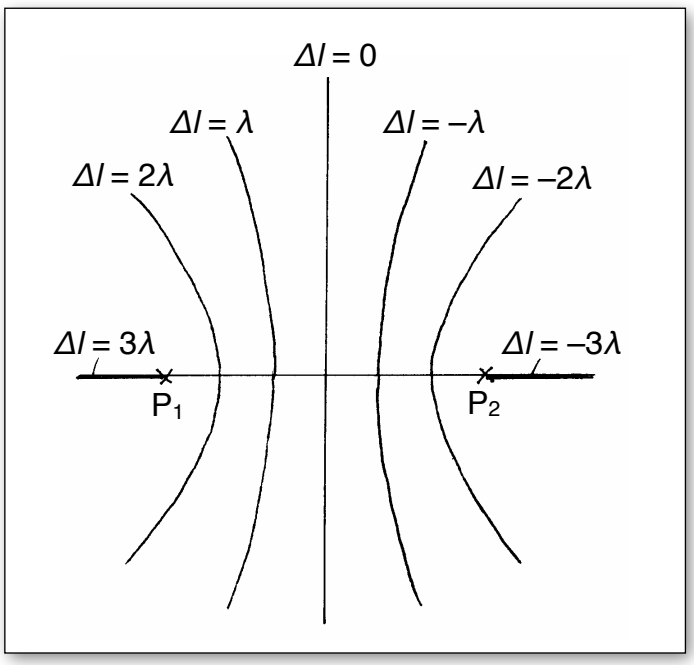


Fig. 7.3
From points P_1 and P_2 spherical waves originate. The oscillations at the positions of points P_1 and P_2 are in phase. The hyperbolas drawn are places of constructive interference.

At any point P there is a phase difference $\Delta\varphi$ between the two waves coming from P_1 or P_2. If the phase difference is an even multiple of π, the waves superpose constructively. Where the phase difference is an odd multiple of π, the waves superpose destructively. If they have the same amplitude at the considered position, they cancel each other completely:

$\Delta\varphi = n\pi$ with $n = 0, \pm 2. \pm 4, \ldots$ amplification

$\Delta\varphi = n\pi$ with $n = \pm 1. \pm 3, \ldots$ attenuation

Instead of the phase difference one often uses the *path difference* Δl between two waves:

$$\Delta l = \frac{\Delta\varphi}{k} = \frac{\Delta\varphi}{2\pi}\lambda$$

The path difference has the dimension of a length. The conditions for amplification and attenuation are thus:

$\Delta l = \frac{n}{2}\lambda$ with $n = 0, \pm 2. \pm 4, \ldots$ amplification

$\Delta l = \frac{n}{2}\lambda$ with $n = \pm 1. \pm 3, \ldots$ attenuation

Fig. 7.3 shows the sections with the drawing plane of the surfaces defined by the phase differences -4π, -2π, 0, 2π and 4π (hyperboloids of revolution). The waves are amplified at these points. The hyperboloids on which extinction takes place lie between these surfaces.

One usually observes a light field by using a flat white screen. If such a screen is placed parallel to the connecting line P_1P_2, one sees hyperbolas as interference patterns. On a small screen at a large distance these become parallel straight lines. If, on the other hand, the screen is perpendicular to the straight line P_1P_2, one obtains circles, or again straight lines if the screen is far off the axis P_1P_2.

7.3 Interference by reflection

7.3.1 The Michelson interferometer

Figure 7.4 shows its setup: S_1 and S_2 are two mirrors, H is a semi-transparent mirror. At first we assume that a plane wave enters from the left. The amplitude of this wave is decomposed by H into the equal parts t and r. The component t is reflected by mirror S_1, the component r by S_2.

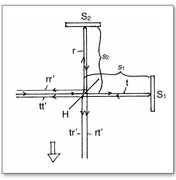


Fig. 7.4
Michelson interferometer

The returning waves are split again at H, namely t into tt' and tr', and r into rt' and rr'. tt' now interferes with rr' and rt' with tr'. If the first of these interferences is constructive, the second is destructive and vice versa. Since it is more convenient, one observes only the light rt' + tr', i.e. the light running away in the direction of the thick arrow in fig. 7.4. Whether the light in this direction is amplified or extinguished depends on the distances s_1 and s_2 of the mirrors from the centre of the apparatus, more precisely: on the difference $s_2 - s_1$. If one of the mirrors S_1 and S_2 is shifted in the direction of its normal by $\lambda/2$, one goes from extinction to amplification or vice versa.

If the incident light is not a plane but a spherical wave, the interference phenomena are more complicated. In fig. 7.5 a spherical wave starts from L. The light field of the outgoing beam is the same as if one had the two point-shaped light sources L_1 and L_2. If the mirror normals are perpendicular to each other and at an angle of 45° to the normal of H, and if S_1 and S_2 have different distances from H, Fig. 7.5a, a circular interference pattern is obtained. If, on the other hand, the distances s_1 and s_2 are the same, and if one of the mirrors S_1 and S_2 is tilted, Fig. 7.5b, then one obtains hyperbolas as interference patterns.

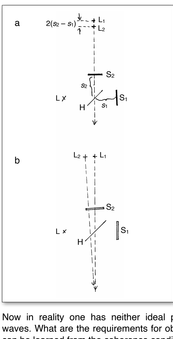


Fig. 7.5
(a) If the virtual light sources L_1 and L_2 are placed one behind the other, a circular interference pattern is created. (b) If the light sources are side by side, hyperbolas are obtained.

Now in reality one has neither ideal plane waves nor spherical waves. What are the requirements for observing interference? They can be learned from the coherence conditions, (7.1), (7.2) or (7.3).

As in Fig. 7.5a, we consider the two "virtual" light sources L_1 and L_2, Fig. 7.6. Their distance is equal to $2(s_2 - s_1)$. We therefore superimpose two light amplitudes belonging to two points of the light field at a distance of $\Delta z = 2(s_2 - s_1)$. According to the 3rd condition in (7.3) the spectral bandwidth $\Delta\omega$ of the light must be

$$\Delta\omega \leq \frac{2\pi c}{\Delta z} = \frac{2\pi c}{2(s_2 - s_1)}$$

The larger the distance $s_2 - s_1$ is chosen, the more monochromatic the light has to be in order to observe interference.

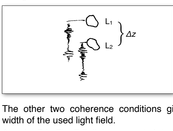


Fig. 7.6
The greater the distance Δz between the virtual light sources, the more monochromatic the light must be in order to observe interference.

The other two coherence conditions give a statement about the width of the used light field.

At point P in Fig. 7.7, light comes to interfere, leaving the same point of the light source in directions different by $\Delta\varphi_x$.

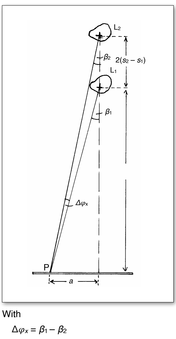


Fig. 7.7
At point P, we get light to interfere, which runs away from one and the same point of the light source in different directions.

With

$$\Delta\varphi_x = \beta_1 - \beta_2$$

and

$$\beta_1 = \frac{a}{l} \qquad \beta_2 = \frac{a}{l + 2(s_2 - s_1)}$$

we get

$$\Delta\varphi_x = \frac{2a(s_2 - s_1)}{l^2}$$

According to the first coherence condition (7.3), the maximum permissible width Δx of the light field becomes

$$\Delta x = \frac{\lambda}{\Delta\varphi_x} = \frac{\lambda l^2}{2a(s_2 - s_1)}$$

Correspondingly, this applies to the width Δy. Here, too, coherence is destroyed if $s_2 - s_1$ is too large. In addition, one can see that the extension of the light source must be small if one wants to observe interference at large distances a from the center of the interference pattern.

The Michelson interferometer has many applications:

- precise length measurements;
- testing the quality of lens and mirror surfaces;
- measurement of the refraction index of gases;
- investigation of the dependence of the speed of light on its direction (Michelson-Morley experiment);
- spectral analysis.

A Michelson interferometer used for spectral analysis is called a *Fourier spectrometer*.

The Fourier spectrometer works as follows: The light to be analyzed is sent into the interferometer. The detector is located in the center of the observation beam. Now one of the two mirrors is moved in the direction of its normal, so that the distance $\Delta s = 2(s_2 - s_1)$ between the virtual light sources changes, and the energy flux density is registered as a function of Δs. The contribution of a frequency ω to the electric field strength at the location of the detector is

$$E_1 + E_1 = E_0(\omega)e^{i\omega t}(1 + e^{i\omega\Delta t}) = E_0(\omega)e^{i\omega t}(1 + e^{i\frac{\omega}{c}\Delta s})$$

From this we get

$$j \propto \int_0^\infty \left|E_+ + E_-\right|^2 d\omega = \int_0^\infty \left|E_0(\omega)\right|^2 \left|1 + e^{i\frac{\omega}{c}\Delta s}\right|^2 d\omega$$

$$= \int_0^\infty \left|E_0(\omega)\right|^2 2\left(1 + \cos\left(\frac{\omega}{c}\Delta s\right)\right) d\omega$$

$$j(\Delta s) \propto 2\cdot\int_0^\infty E_0(\omega)^2 d\omega + 2\cdot\int_0^\infty E_0(\omega)^2 \cos\left(\frac{\omega}{c}\Delta s\right) d\omega$$

The Fourier transformation of the measured function $j(\Delta s)$ returns the spectrum $E_0(\omega)^2$.

The resolution of the spectrometer is the better, the larger the range over which Δs is changed. Since Δs must be measured accurately to fractions of a wavelength, the spectrometer is not suitable for light of very short wavelengths. It is used for spectral analysis in the infrared range.

There are other interferometers that are similar to the Michelson interferometer (Mach-Zehnder interferometer, Sagnac interferometer), and there are simpler experiments and natural phenomena based on the same principle as the Michelson interferometer.

Fig. 7.8 shows the Pohl interference experiment. The light coming from L is reflected at the front and the back of a thin mica plate. L_1 and L_2 are the virtual mirror images of L. The path difference between waves 1 and 2 increases with increasing angle θ. On the screen one can see a ring-shaped interference pattern.

Fig. 7.8
The Pohl interference experiment. The light coming from L is reflected at the front and back of a thin mica plate.

Soap bubbles or a thin film of oil on water appear colored. The light is reflected at the top and bottom of the oil film or bubble skin, and the reflected light is made to interfere on the retina of our eyes. Since the condition for extinction depends on the wavelength, it occurs at a different angle depending on the wavelength.

The coating of optical lenses is also based on this principle: a thin layer (thickness d) of a transparent material (refractive index n_b) is applied to the lens surface. The light is reflected at both the front and the back of the layer. By choosing an appropriate thickness of the layer ($d = \lambda/4n_b$), the reflected light is destructively interfering, and by choosing the appropriate refractive index ($n_b = \sqrt{n}$) for the layer, the amplitudes of the two reflected waves are equal, so that they cancel each other completely.

7.3 Interference by reflection

7.3.2 The Fabry-Pérot interferometer

The Fabry-Pérot interferometer essentially consists of two plane-parallel glass plates which are mirrored in such a way that the degree of reflection is about 0.9, Fig. 7.9.

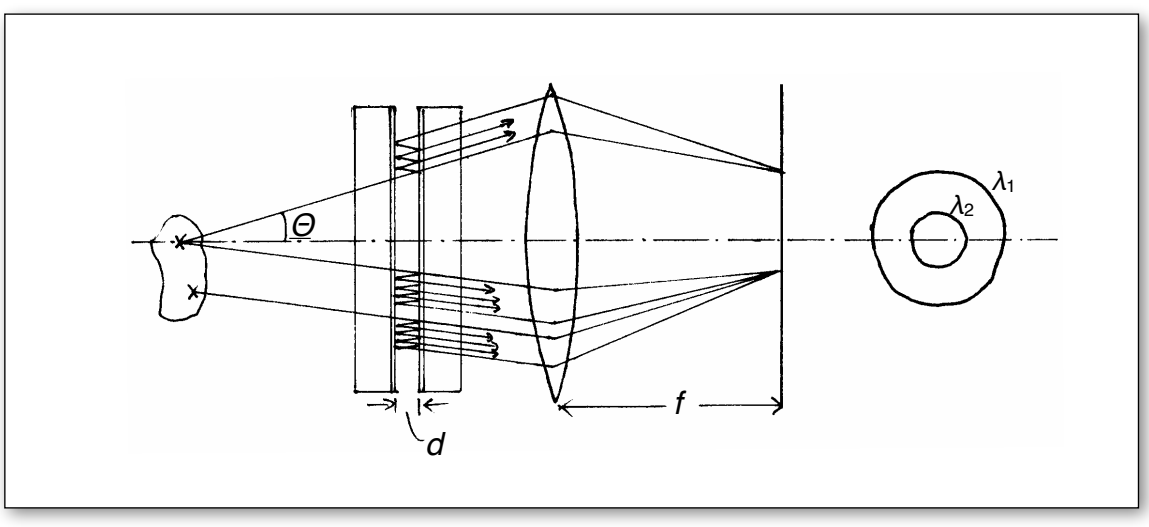


Fig. 7.9
Fabry-Pérot interferometer

On one side of the plates there is an extended light source, on the other a lens, and in the focal plane of the lens is the observation screen. The light that reaches the space between the plates is reflected back and forth several times. With each reflection, a small part of this light leaves the space between the plates. The light that leaves this space in the direction of the lens, falls on the screen, and one there observes interference patterns. The figure shows that in one point on the observation screen, the light that corresponds to a single direction in front of the plates is united. In the point under consideration, many waves (of the same direction) interfere: the wave that was not reflected at all by the glass plates, the wave that was reflected once to and fro, the wave that was reflected twice to and fro, and so on. There is a phase difference between two consecutive waves in this series δ, which consists of two parts: one part

$$\frac{4\pi d}{\lambda \cos\theta}$$

that is caused by the different lengths of the waves' paths and one part that is due to the fact that each further back and forth of the wave causes two phase jumps when being reflected. Successive waves interfere constructively or destructively, depending on the angle Θ and the distance d between the plates. In the observation plane, the locations with the same phase difference lie on circles. One therefore observes circular interference patterns. Before we discuss the interference pattern further, we want to examine the question, which requirements the coherence conditions make on the light source.

In the plane of the screen, waves that are offset by $\Delta z = 2d$ in the direction of propagation interfere. The third relationship of (7.2) therefore requires

$$\Delta\omega \le \frac{2\pi c}{2d}$$

The greater the plate spacing, the more monochromatic the light must be.

At any point on the screen, light of a single direction interferes, i.e. $\Delta k_x = \Delta k_y = 0$. It follows from the first two conditions of (7.1), (7.2) or (7.3) that the light source may be extended laterally to any extent.

We now calculate the amplitude of the transmitted light as a function of the angle of incidence Θ and the plate distance d. The designations are shown in Fig. 7.10.

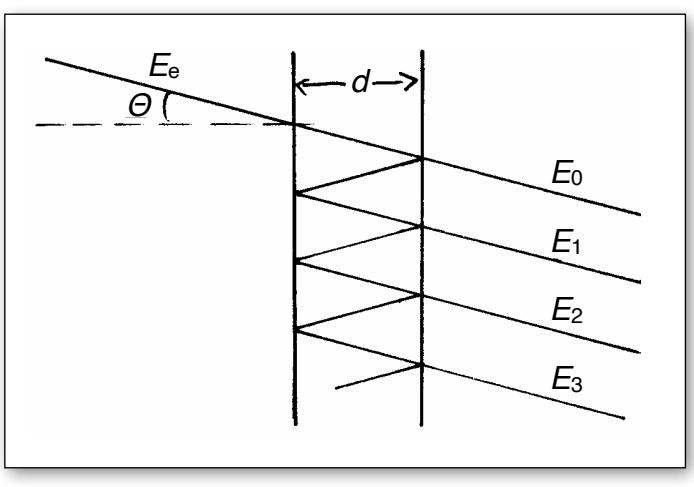


Fig. 7.10
To calculate the amplitude of the transmitted light in the Fabry-Pérot-interferometer

The light with the amplitude E_0 had to pass a mirror twice, whereby its amplitude was reduced by the factor t^2.

Thus we have

$$E_0 = E_e e^{i\omega t} t^2$$

The light with the amplitude E_1 was reflected twice more, and due to the longer distance, it has experienced a phase shift of $k\Delta l$ compared to E_0, where

$$\Delta l = \frac{2d}{\cos\theta}$$

It is therefore

$$E_1 = E_e e^{i\omega t} \cdot t^2 \cdot r^2 e^{ik\Delta l}$$

Accordingly one gets E_2, E_3 etc.

$$E_n = E_e e^{i\omega t} \cdot t^2 \cdot r^{2n} e^{ink\Delta l}$$

That there is still a phase shift in every reflection is expressed by the fact that r is complex. With

$$r = \rho \cdot e^{i\Delta\varphi}$$

we obtain

$$E_n = E_e e^{i\omega t} \cdot t^2 \left(\rho^2 \cdot e^{i(k\Delta l + 2\Delta\varphi)}\right)^n$$

The resulting amplitude is

$$\sum_0^\infty E_n = E_e e^{i\omega t} \cdot t^2 \frac{1}{1 - \rho^2 \cdot e^{i(k\Delta l + 2\Delta\varphi)}}$$

On the screen one does not observe the field strength, but the energy current density

$$j \propto EE^* = \left(E_e t^2\right)^2 \frac{1}{1 - 2\rho^2 \cos(k\Delta l + 2\Delta\varphi) + \rho^4}$$

The independent variables d and Θ are hidden in Δl.

Figure 7.11 shows j as a function of $k\Delta l$ for two different values of ρ. One can see from this figure the benefit of the Fabry-Pérot interferometer: It is a high-resolution spectrometer. For different wavelengths the ring-shaped "spectral lines" have different diameters.

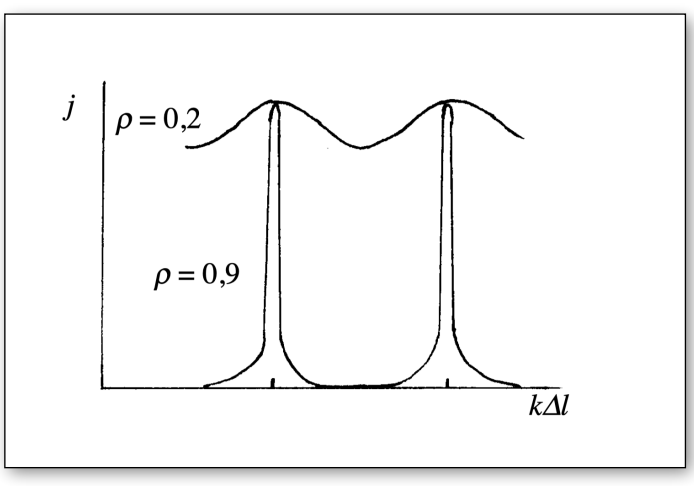


Fig. 7.11
Energy current density as a function of the phase shift for two different values of the reflection coefficient. The closer ρ is to 1, the better the resolution.

One also notices that the resolution gets better the closer the reflection coefficient is to 1.

The condition for the proper operation is, of course, that the coherence condition is respected. If one ignores the coherence condition, the rings that belong to different light frequencies coincide.

Fig. 7.11 shows that the device only transmits light that is incident at specific, sharp angles. The light complementary to the transmitted light is reflected back towards the light source.

Besides being used as a spectrometer, this arrangement is also used as a laser resonator. In this case, the plate distance is as large as the laser is long.

A very simple version of the Fabry-Pérot interferometer is the interference filter: a thin layer of metal or a suitable non-conducting material is applied to both sides of a glass plate. In contrast to filters whose effect is based on absorption, interference filters only allow light of a very small wavelength range to pass through: about 5 - 10 nm.

7.4 Interference by diffraction

7.4.1 Generals

Instead of bringing together light coming from different places within an elementary beam by mirrors and thus causing interference, the phenomenon of diffraction can be used for this purpose.

In fig. 7.12 the light diffracted at the two holes is brought to interference and observed on the screen.

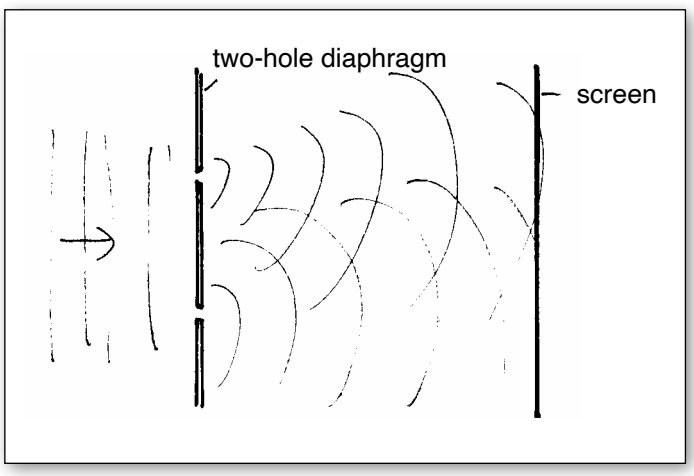


Fig. 7.12
The light diffracted at the two holes is brought to interference.

The mathematics of diffraction becomes particularly simple when we look at the arrangement invented by Fraunhofer, Fig. 7.13: Behind the diffracting object there is a lens. The observation screen is placed in its focal plane.

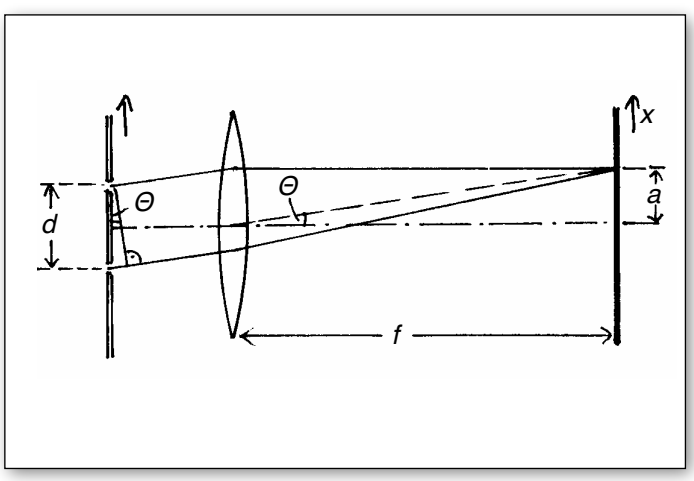


Fig. 7.13
Fraunhofer arrangement for the interference of diffracted light

Here, similar to the Fabry-Pérot interferometer, the light that belonged to a certain direction in front of the lens is brought to interference at a point on the screen: the lens assigns a position (on the screen) to a direction (left of the lens).

Before investigating certain interference patterns, we want to ask again which conditions the equations (7.1) to (7.3) impose on the structure of the arrangement.

On the screen, light from an area of width $\Delta x = d$ is to be brought to interference. The k-vector is therefore only allowed to scatter by an angle

$$\Delta\varphi_x < \frac{\lambda}{d}$$

Laser light is such that this relationship still holds when the entire beam width is taken for Δx. Therefore, one simply says “laser light is coherent”. Laser light is particularly suitable for diffraction interference experiments. The third coherence condition tells us up to which angle against the optical axis one can still observe interference phenomena. On the screen at a distance a from the optical axis light is brought to interference, whose k-vector forms with the optical axis the angle Θ with $\tan\Theta = a/f$. This light comes from two points of the light field which are distant by $\Delta z = d \cdot \sin\Theta$ in the direction of propagation. With $\sin\Theta \approx \tan\Theta$ we get

$$\Delta z = \frac{d \cdot a}{f}$$

Therefore, the third coherence condition (7.3) requires that

$$\Delta\omega \leq \frac{2\pi c f}{d a}$$

The path difference of the interfering waves increases with increasing a and with increasing d. The larger the diameter of the diffracting object and the greater the distance from the optical axis where interference patterns are to be observed, the smaller the frequency range of the light must be.

7.4 Interference by diffraction

7.4.2 The Fraunhofer arrangement as a Fourier transformer

We start with a one-dimensional analysis: The diffracting object is extended in the *x*-direction and is characterized by the transparency function $t(x)$. $t(x)$ indicates by which factor the amplitude of the wave behind the diffracting object is smaller than before. According to the Huygens-Fresnel principle, one can imagine that a spherical wave starts from every point behind the object. Since the lens makes waves of a certain direction converge in a certain point, we ask what contribution the individual Huygens' elementary waves make to the field strength of the wave of a certain direction Θ.

These contributions have a different phase in the point on the screen, depending on the position on the object. In order to calculate the light amplitude on the screen, we have to integrate the contributions of all spherical waves belonging to the direction Θ with the correct phase.

We first look at the contributions of the direction Θ, which start from points P1 and P2, Fig. 7.14.

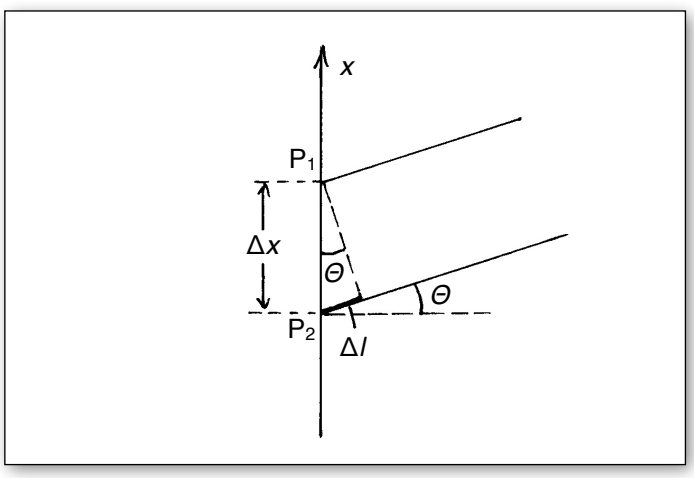


Fig. 7.14
To calculate the path difference of two spherical waves

The Distance between P1 and P2 is Δx. The path difference of the waves is

$$\Delta l = \Delta x \cdot \sin \Theta$$

So their phase difference is

$$\Delta\varphi = \Delta l \cdot k = \Delta x \cdot k \cdot \sin \Theta = \Delta x \cdot k_x$$

where k_x is the *x*-component of the *k*-vector of the diffracted light. We get the total contribution $T(k_x)$ of the spherical waves to the field strength at the considered point on the screen by integration

$$T(k_x) = \int_{-\infty}^{+\infty} t(x) e^{-ik_x x}\, dx \qquad (7.4)$$

To each direction Θ a value of the coordinate x' on the screen is assigned:

$$\tan\theta = \frac{x'}{f}$$

With

$$\sin\theta = \frac{k_x}{x}$$

and $\sin \Theta \approx \tan \Theta$ we get

$$k_x = \frac{k}{f} x'$$

Therefore instead of (7.4) we can also write

$$T(x') = \int_{-\infty}^{+\infty} t(x) e^{-i\frac{k}{f}xx'}\, dx \qquad (7.5)$$

$T(k_x)$ is a measure of the electric field strength on the screen. It has the same dependence on k_x as the field strength, but it is not the field strength itself. It cannot be the field strength, because $t(x)$ is not a field strength either. One would not have been allowed to use the field strength behind the diffracting object instead of $t(x)$, at most a field strength per *k*-direction interval. By the integration a length dimension is added. Furthermore, one should have considered that the field strength of a spherical wave decreases with $1/r$ from the center of the sphere. Fortunately, we can ignore all these complications: We are not interested in the absolute value of the field strength, but only in its change as a function of k_x or of x', which is the same as that of T.

With (7.4) and (7.5) we have now obtained a very simple result: The field strength of the Fraunhofer diffraction pattern is the Fourier transform of the field strength at the location of the object.

The signal which is registered with the usual detectors, for example with a photographic film, is proportional to the square of the field strength, i.e. to the square of the Fourier transform of the transparency function of the object.

7.4 Interference by diffraction

7.4.3 Single-slit and double-slit diffraction

The transparency function of a slit of width a is

$$t(x) = \begin{cases} 1 & \text{for } -\frac{a}{2} < x < \frac{a}{2} \\ 0 & \text{else} \end{cases}$$

We have already calculated the Fourier transform of this function in chapter 1:

$$T(k_x) = a \frac{\sin \frac{a}{2} k_x}{\frac{a}{2} k_x}$$

The graphical representation of this function is shown in Fig. 1.3. The energy flux density, which is proportional to the square of this function, is shown in Fig. 7.15. It is zero for

$$\frac{a}{2} k_x = n\pi \quad n = \pm 1, \pm 2, \ldots$$

or with $k_x = k \sin \Theta$ for

$$\frac{a}{2} k \sin\theta = n\pi$$

or

$$\sin\theta = \frac{n\lambda}{a}$$

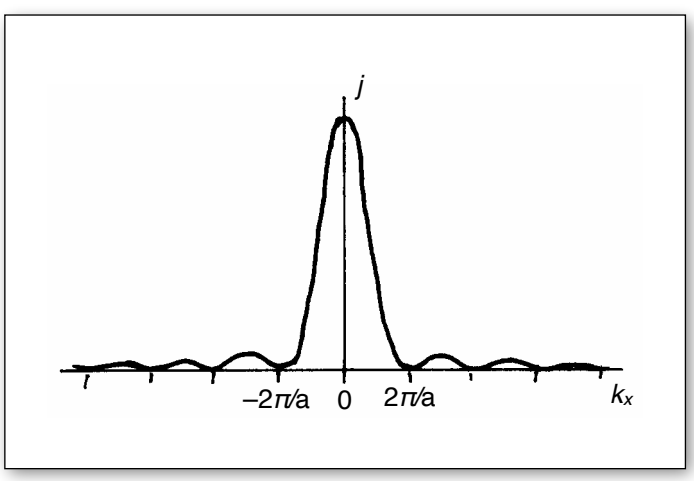


Fig. 7.15
Energy flux density in the diffraction pattern of a single slit

If the slit width becomes narrower, the *main maximum* of the diffraction pattern becomes wider. In the limit case, in which the transparency function is a δ-function, $T(k_x)$ = const (the main maximum is infinitely wide).

As a second example we calculate the diffraction pattern of two very narrow (δ-functional) slits, which lie at a distance d to each other, fig. 7.16. Their transparency function is

$$t(x) = \delta\left(x - \frac{d}{2}\right) + \delta\left(x + \frac{d}{2}\right)$$

Thus we get

$$\begin{aligned} T(k_x) &= \int t(x) e^{-ik_x x} \\ &= \int \delta\left(x - \frac{d}{2}\right) e^{-ik_x x} dx + \int \delta\left(x + \frac{d}{2}\right) e^{-ik_x x} dx \\ &= e^{-ik_x \frac{d}{2}} + e^{ik_x \frac{d}{2}} \\ &= 2\cos\pi\left(k_x \frac{d}{2}\right) \end{aligned}$$

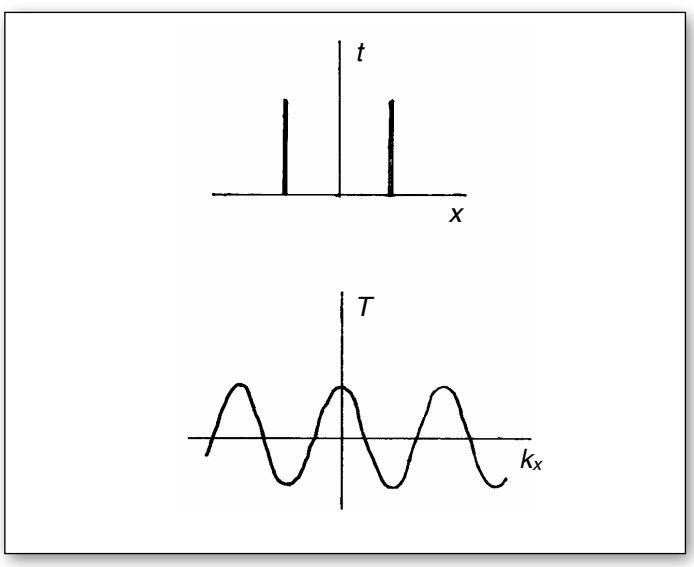


Fig. 7.16
Transparency function (top) and Fourier transform (bottom) of the double slit

7.4 Interference by diffraction

7.4.4 Grating diffraction

Finally we consider the case of the diffraction grating: a series of equidistant narrow slits. We take as transparency function the sum of N equidistant δ functions:

$$t(x) = \sum_{m=1}^{N} \left\{ \delta\left[x - (2m-1)\frac{d}{2} \right] + \delta\left[x + (2m-1)\frac{d}{2} \right] \right\}$$

$$T(k_x) = \sum_{m=1}^{N} 2\cos\left[k_x (2m-1)\frac{d}{2} \right]$$

$$= 2\left(\cos\frac{1}{2}k_x d + \cos\frac{3}{2}k_x d + \cos\frac{5}{2}k_x d + \ldots \right)$$

Here we have got $T(k_x)$ as a series. This representation is especially practical if one wants to follow on the computer screen, which influence each new slit pair has on the diffraction image. But one can also give a closed-form expression for $T(k_x)$. The calculation becomes easiest if we write

$$t(x) = \sum_{m=0}^{N-1} \delta(x - md)$$

Then the Fourier transform is

$$T(k_x) = \int t(x) e^{-ik_x x}\, dx$$

$$= \sum_{m=0}^{N-1} \int \delta(x - md) e^{-ik_x x}\, dx$$

$$= \sum_{m=0}^{N-1} e^{-ik_x dm}$$

$$\frac{1 - e^{-ik_x dN}}{1 - e^{-ik_x d}}$$

The fact that this expression is complex only means that the field strength is out of phase with the contribution with $m = 0$.

We are interested in the energy flux density:

$$j \propto \left(\frac{1 - e^{-ik_x dN}}{1 - e^{-ik_x d}} \right) \cdot \left(\frac{1 - e^{-ik_x dN}}{1 - e^{-ik_x d}} \right)$$

$$= \frac{1 - \cos(k_x dN)}{1 - \cos(k_x d)} = \frac{\sin^2 \frac{k_x dN}{2}}{\sin^2 \frac{k_x d}{2}}$$

Figure 7.17 shows this function for $N = 6$.

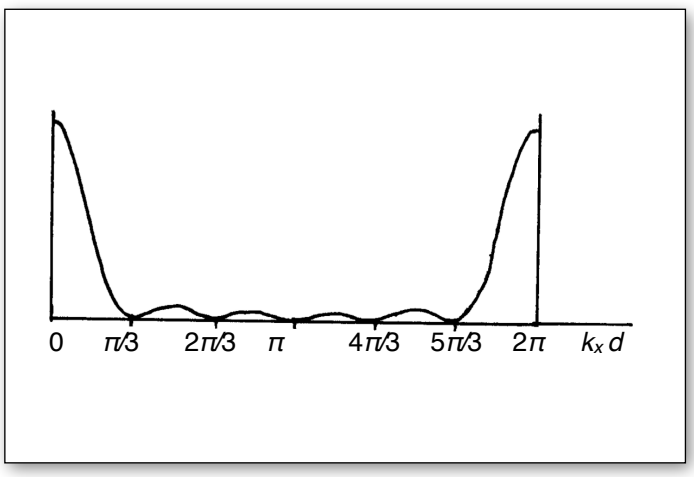


Fig. 7.17
Energy flux density in the diffraction pattern of a grating

The energy current density has high maxima proportional to N^2 for k_x values for which the denominator becomes zero, i.e. when

$$\frac{k_x d}{2} = n\pi \qquad n = 0, \pm 1, \pm 2, \ldots \tag{7.6}$$

These maxima are called *main maxima* of zero, first, second etc. order. Between each two main maxima there are $N - 2$ much smaller *side maxima*.

The diffraction grating is the most important component of a grating spectrometer. In such a device, light with different frequency components is sent on the grating and a superposition of the corresponding diffraction images is created (the energy flux densities add up). In order to separate the light with two different frequencies, the corresponding main maxima must be clearly separated. From (7.6) follows that the separation of two maxima of the same order, which belong to two different frequencies, is proportional to the number n. In addition, two maxima can be resolved the better the narrower they are. Now the width of a main maximum is about $1/N$ of the distance between two adjacent main maxima. So the larger the number N of illuminated slits, the better the frequency resolution.

Altogether, therefore, the product $n \cdot N$ is a measure for the resolving power of the grating spectrometer.

7.4 Interference by diffraction

7.4.5 Convolutions

In optics, a mathematical operation is often useful, which is called convolution. The convolution is defined by

$$F(x)=\int_{-\infty}^{+\infty} f(x')\cdot\phi(x-x')\,dx' \qquad (7.7)$$

One says F is the convolution of f and Φ.

One can consider the convolution as a limit of the following sum:

$$a_1\,\Phi(x-x_1)+\ a_2\,\Phi(x-x_2)+a_3\,\Phi(x-x_3)+\ldots$$

The function $\Phi(x)$ is thus shifted on the x-axis by the amounts x_1, x_2, x_3,…, and all these functions are added after multiplication by the weight factors a a_1, a_2, a_3,…

In expression (7.7), the weight factors are $f(x')dx'$.

We will use convolutions to calculate diffraction patterns. But they play an important role in optics in other contexts as well.

We consider an example for the application of a convolution, which belongs to optics, but has nothing to do with the topic we are currently working on: the pinhole camera. The propagation of light can be described here with light rays, Fig. 7.18.

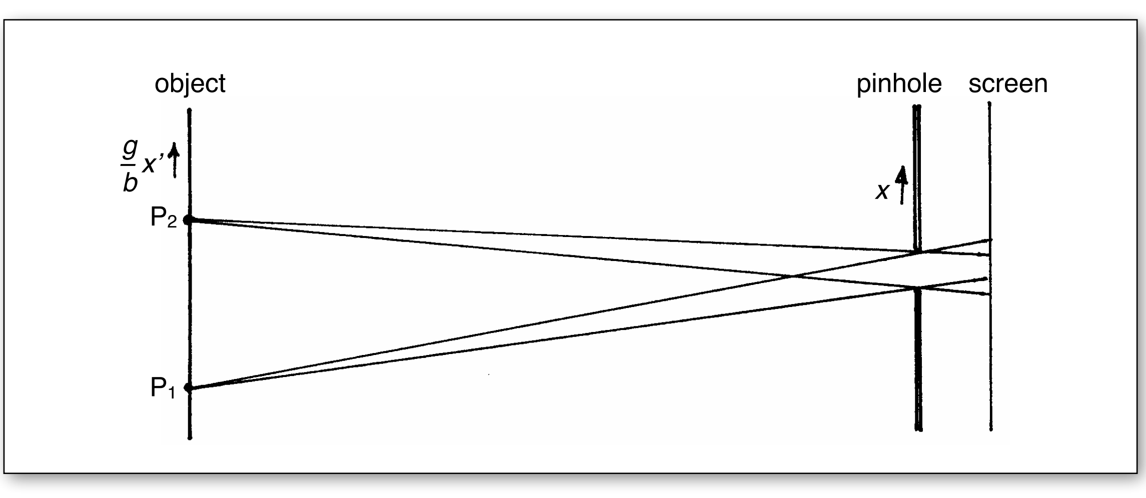


Fig. 7.18
Pinhole camera

Each point P_i of the object generates a light distribution on the screen that corresponds to the transparency function $\Phi(x)$ of the pinhole. The two object points P_1 and P_2 individually create images $\Phi(x-x'_1)$ and $\Phi(x-x'_2)$ that are shifted against each other. Both together create the weighted sum

$$a_1\,\Phi(x-x'_1)+\ a_2\,\Phi(x-x'_2)$$

a_1 and a_2 are measures of the energy flux density of the light coming from points P_1 and P_2 respectively. In order to determine the effect of not only two but of all the object points on the screen, one must form the integral of a sum:

$$F(x)=\int f(x')\cdot\phi(x-x')\,dx'$$

$F(x)$ is the energy flux density distribution on the screen, $f(x')$ is a measure of the energy flux density distribution in the object plane, and $\Phi(x)$ describes the transparency of the hole. The unit in which x' is measured in the object plane is larger by a factor of b/g than the unit in which the same variable x' is measured in the image plane.

Often a simple special case of a convolution is important: the case where $f(x')$ describes a series of sharp peaks at the positions x_1, x_2,..., i.e.

$$f(x')=\delta(x'-x_1')+\delta(x'-x_2')+\ldots$$

The convolution then reduces to the addition of a discrete set of functions which differ only in that they are shifted on the x-axis by finite distances from each other, Fig. 7.19.

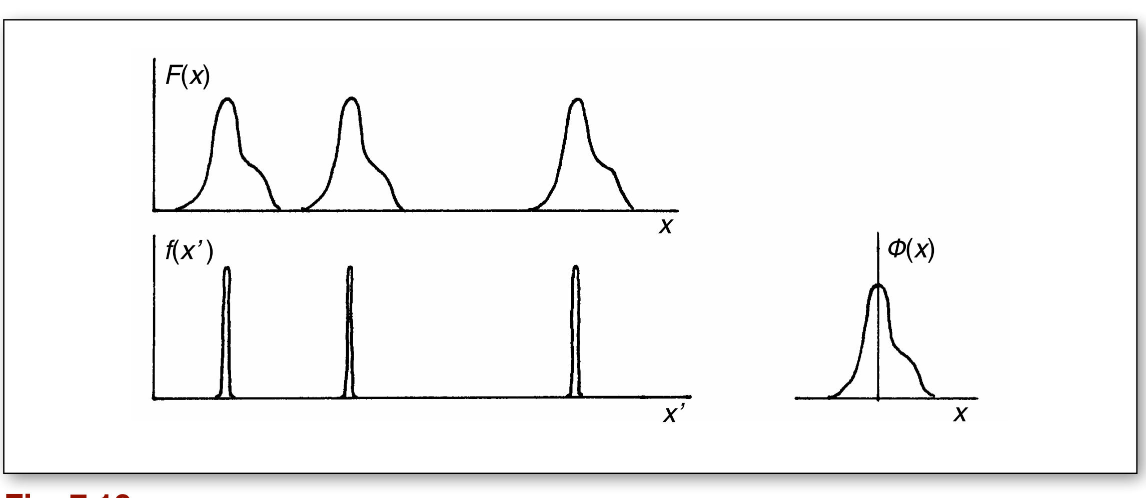


Fig. 7.19
Convolution $F(x)$ of a set of δ functions $f(x')$ with the function $\Phi(x)$

It can be seen that this case is suitable for describing a set of objects that all have the same spatial structure but are located at different places in space: e.g. atoms in a crystal lattice or chairs in a seminar room.

We now return to the examination of Fraunhofer diffraction patterns. Whenever the diffracting object consists of a set of arbitrarily arranged openings of the same kind, the transparency function can be written as a convolution of the transparency function of a single opening with a sum of δ functions indicating the locations of the openings. This description is useful because there is a simple mathematical theorem about the Fourier transform of a convolution integral:

The Fourier transform of the convolution of two functions f and Φ is equal to the product of the Fourier transform of f and that of Φ.

As a proof we calculate the Fourier transform $T(k)$ of the function

$$t(x)=\int f(x')\cdot\phi(x-x')\,dx'$$

We obtain

$$T(k)=\int t(x)e^{-ikx}\,dx$$
$$=\iint f(x')\cdot\phi(x-x')e^{-ikx}\,dx'dx$$

If one sets $x-x'=y$, then

$$T(k)=\int f(x')\cdot\phi(y)e^{-ik(x'+y)}\,dx'dy$$
$$=\int f(x')e^{-ikx'}\,dx'\cdot\int\phi(y)e^{-iky}\,dy$$

q. e. d.

We use this theorem to calculate the diffraction pattern of a double-slit. The transparency function $t(x)$ of the double-slit is the convolution of the transparency function $\Phi(x)$ of the single-slit with the transparency function $f(x')$ of the δ-shaped double-slit (see section 7.4.3):

$$\left.\begin{array}{l}\phi(x)=\begin{cases}1 & \text{for } -\frac{a}{2}<x<\frac{a}{2}\\ 0 & \text{else}\end{cases}\\ f(x')=\delta\left(x'-\frac{d}{2}\right)+\delta\left(x'+\frac{d}{2}\right)\end{array}\right\}\quad t(x)=\int f(x')\cdot\phi(x-x')\,dx'$$

The Fourier transforms FT_Φ of Φ and FT_f of f are

$$FT_\phi(k_x)=a\frac{\sin\frac{a}{2}k_x}{\frac{a}{2}k_x}$$

$$FT_f(k_x)=2\cdot\cos k_x\frac{d}{2}$$

According to our theorem, the Fourier transform of $t(x)$ is the product of FT_Φ and FT_f:

$$T(k_x)=a\frac{\sin\frac{a}{2}k_x}{\frac{a}{2}k_x}\cdot 2\cdot\cos\frac{d}{2}k_x$$

Figure 7.20 shows the square of $T(k_x)$ for the case that $d=3a$.

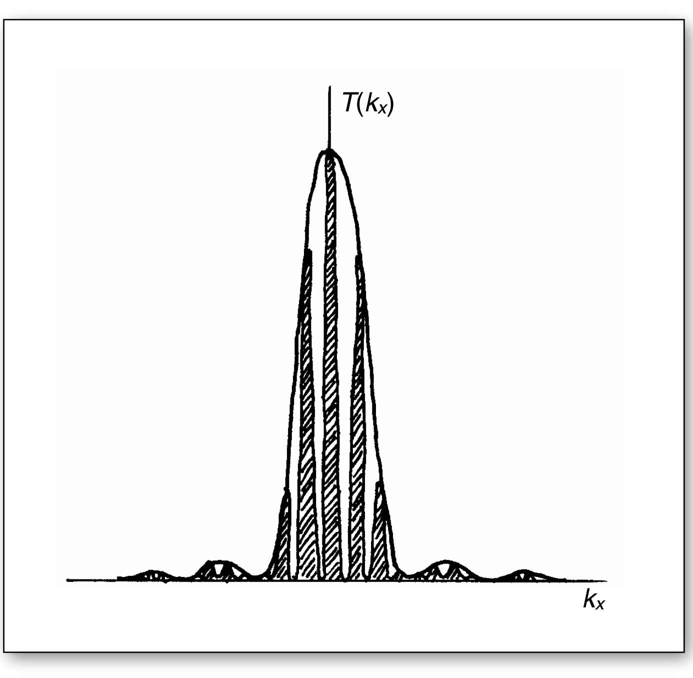


Fig. 7.20
Energy flux density in the diffraction image of the double slit. One can see that the function can be represented by two factors: One corresponds to the diffraction pattern of a single slit, the other to a double slit of two δ-functions.

This function can also be described as follows: the diffraction pattern of two δ-shaped slits (a fast oscillating cosine function) is modulated with the diffraction pattern of the wide (not δ-shaped) single slit.

To obtain the diffraction image of a real grating, i.e. a grating whose slits have a finite width, the diffraction pattern of the δ-grating (section 7.4.4) must be modulated with the diffraction pattern of the single slit.

The coherence condition

$$\Delta\varphi\leq\frac{\lambda}{d}$$

can be used to determine the angular distance of very closely neighbouring stars, such as the partners of a binary star system, or the angle of aperture at which the diameter of a single star is seen. The arrangement used for this is the Michelson stellar interferometer, Fig. 7.21.

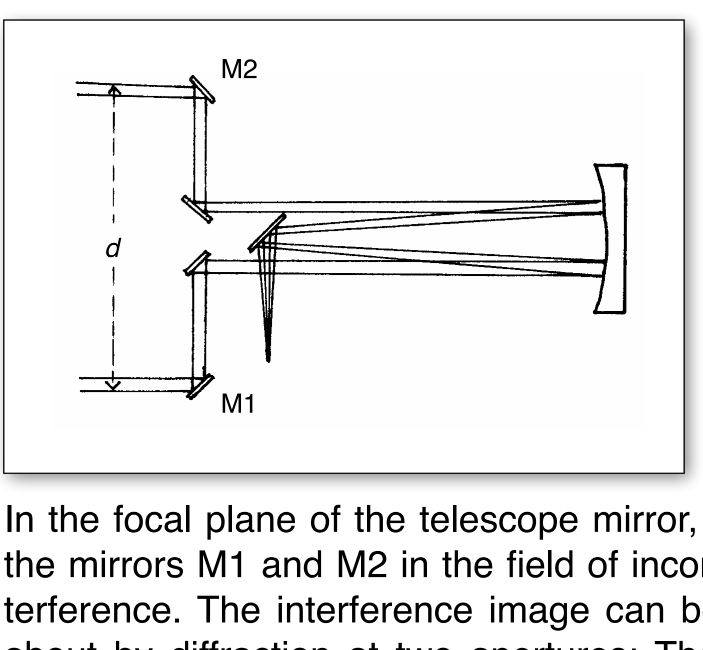


Fig. 7.21
Michelson stellar interferometer

In the focal plane of the telescope mirror, light from the locations of the mirrors M1 and M2 in the field of incoming light is brought to interference. The interference image can be imagined to have come about by diffraction at two apertures: The mirrors M1 and M2 are equivalent to two pinholes in a screen that is placed in the way of the incoming light. The diffraction pattern is the product of the diffraction pattern of one of these “apertures” with that of δ-shaped slits at a distance d. The rotational symmetric structure in fig. 7.22 is caused by the shape of the single mirrors, the vertical stripes by the δ double slit.

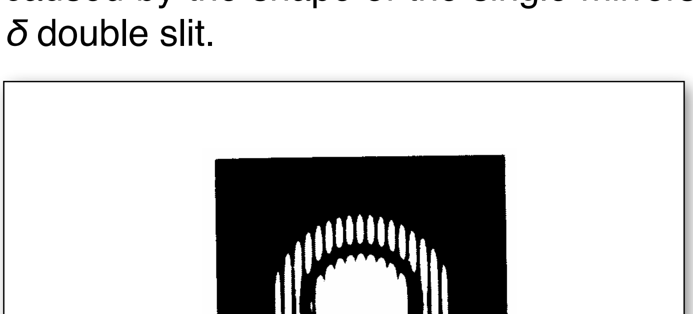


Fig. 7.22
Diffraction pattern of a Michelson stellar interferometer. The rotational symmetric structure is the diffraction pattern of the single mirrors. The vertical stripes correspond to a δ double slit.

M1 and M2 are now moved outwards until the interference fringes disappear. The corresponding distance d of the mirrors is used to calculate the opening angle $\Delta\Phi$ of the light field.

According to this method, a star diameter was determined for the first time in 1920 (Betelgeuse, in Orion top left).

8

Ray optics

8.1 Light rays

We all know the description of light by rays. Rays are imaginary lines. Light moves along these lines like particles on a trajectory.

The description of light by rays is only possible under certain conditions. In order to formulate these conditions, we have to investigate what the special feature of the description by rays is, if we assume that light should actually be described by waves.

The description by rays implies, on the one hand, that light casts a sharp, “geometric” shadow. In Fig. 8.1, light is emitted from the small pinhole L1, and creates a sharp shadow of the large hole L2 on the screen. The shape of the shadow is obtained by the construction known to everyone. It is also said that light propagates in straight lines. But the sharp shadow is only obtained if the diffraction of the light at L2 can be neglected, and this is the case if the hole is large compared to the wavelength. Thus, our first condition for the validity of ray optics is:

The wavelength must be small compared to the dimensions of obstacles.

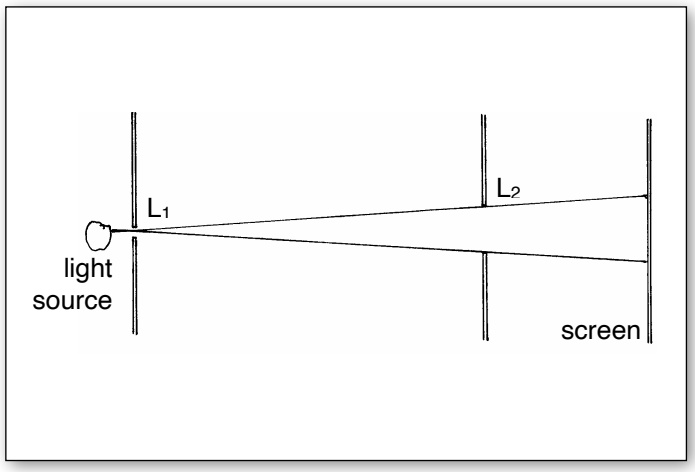


Fig. 8.1
If the wavelength of the light is small against the hole L_2 , a sharp shadow of the hole appears on the screen.

On the other hand, the description by rays implies that it is possible to add up the energy flows corresponding to two rays, Fig. 8.2, but this is only valid if the light is sufficiently incoherent. The light whose energy currents are added must not originate from the same elementary beam, otherwise interference patterns will occur. Our second condition for the validity of ray optics is therefore

The light must be sufficiently incoherent.

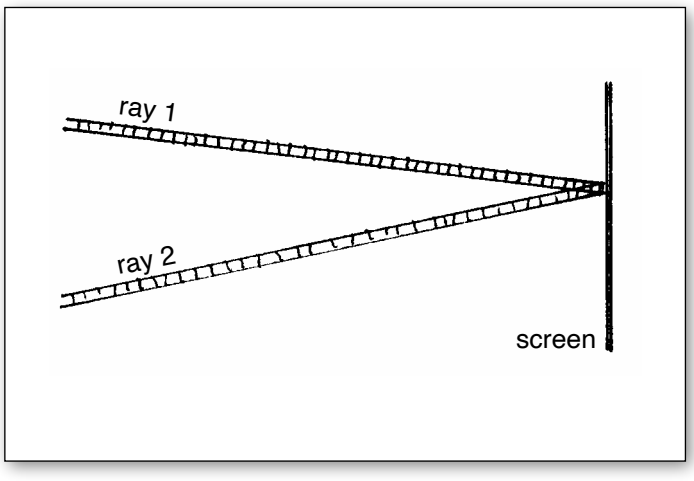


Fig. 8.2
The energy fluxes can only be added if the light is sufficiently incoherent.

The approximation of ray optics behaves to wave optics in the same way as the approximation of classical point mechanics to quantum mechanics. The concept of a light ray in ray optics corresponds to the concept of the path of a mass point in Hamiltonian mechanics.

If one applies ray optics, one only asks for the path of the light. One does not ask for the velocity at which the light travels on the rays. One also does not care about the polarization and therefore one should not ask about which part of the light is reflected and which part is refracted on a glass surface.

8.2 Fermat's principle

We assume that the conditions for operating with light rays are fulfilled and turn to the rules of geometric optics.

If one brings into a distribution of light that comes from the left, Fig. 8.3a, two pinholes mounted one behind the other, a so-called collimator, Fig. 8.3b, a narrow beam of light is created that propagates similarly to a ray. Therefore, such a light beam is often called a ray – in accordance with the colloquial use of the word ray. Such a light beam makes it possible to examine the rules that apply to light rays.

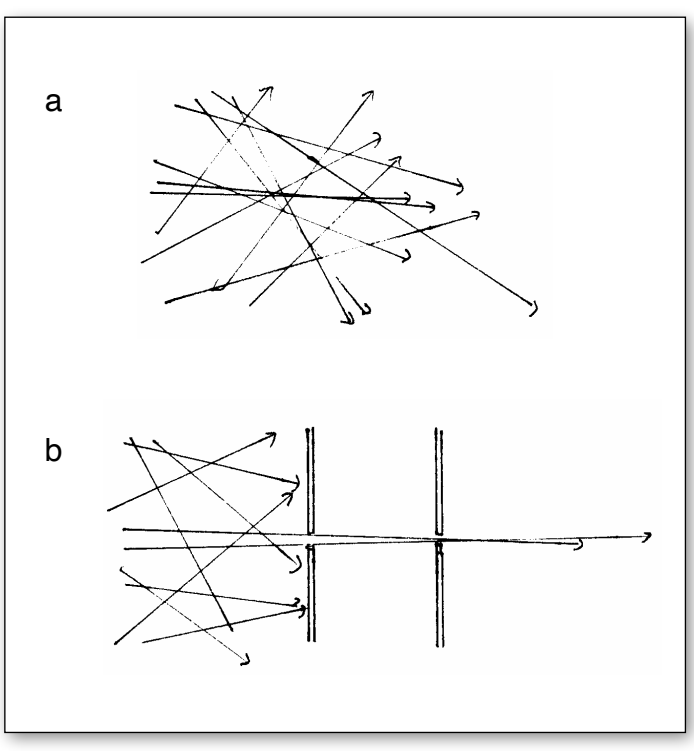


Fig. 8.3
(a) Light without a well-defined direction of propagation. (b) A light ray is generated with the collimator.

The typical task of geometrical or ray optics is as follows:

Given is a point P at the position ***r*** and a direction ϑ, Φ. What is the further path of the beam passing the point P in the direction ϑ, Φ? Fig. 8.4 illustrates this task for the case that the ray travels in the drawing plane.

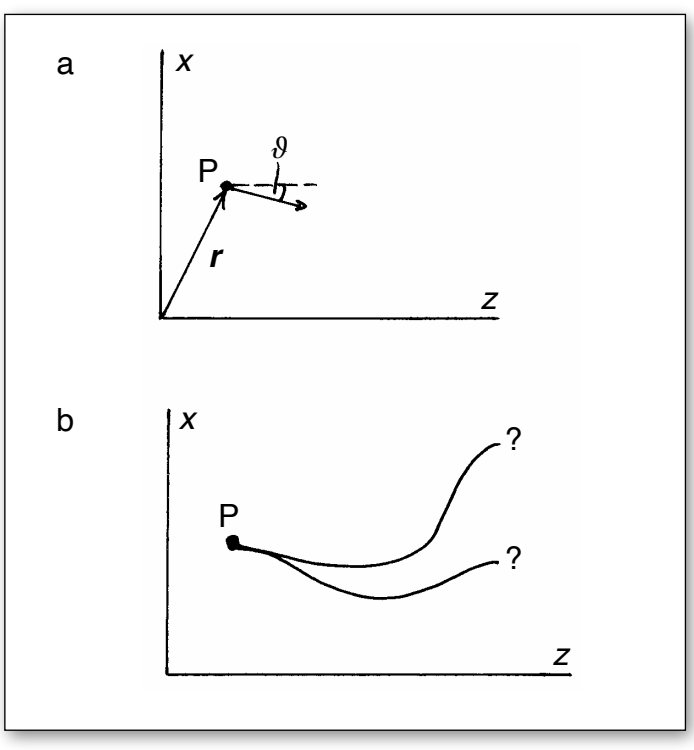


Fig. 8.4
(a) The light starts in a certain direction. (b) Where does it go next?

As long as the refractive index is spatially constant and changes step-wise only at well-defined interfaces, the following three known rules can be applied:

- Light propagates in a straight line.
- Law of reflection ($\alpha = \alpha'$)
- Law of refraction ($n_a \sin \alpha = n_b \sin \beta$)

These rules are sufficient for the treatment of many optical devices. Tracking a beam through a sequence of refractive and reflective interfaces is called *ray-tracing*.

Now, the three rules can be replaced with a single, generally valid rule: *Fermat's principle*. For this purpose, we first define the *light path w* between two points A and B:

$$w_{AB} = \int_A^B n\, ds$$

Here *ds* is an infinitesimal section of the light beam and *n* is the refractive index. Fermat's principle states that the actual light path between two given points A and B is minimum compared to hypothetical neighboring paths between these points:

$$\delta(w_{AB}) = 0$$

Now, we allow the refractive index to change continuously in space.

The variational calculus deals with the general solution of such an expression.

That the law of reflection follows from Fermat's principle can be seen easily, Fig. 8.5. Besides B, the point B', which is mirror-symmetrical to B, is marked in the figure. One can see that the path APB is equal to the path APB'. It is obvious that APB' is minimal if $\alpha = \alpha'$ is chosen.

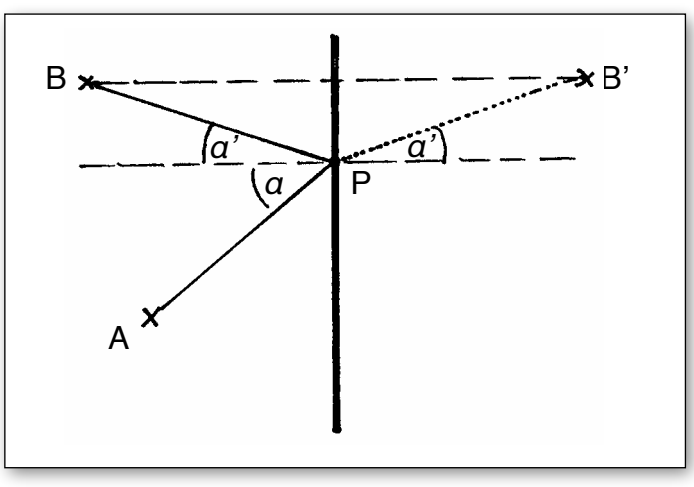


Fig. 8.5
The length of the path APB is minimal when $\alpha = \alpha'$.

The derivation of the law of refraction from Fermat's principle is somewhat more complicated.

A ray always starts at a light source or a scattering object and ends on an absorbing or scattering object. One notices the special role that scattering objects play, for example white surfaces or ground glass: the rays of the incident light end here, and new rays of light begin, but their directions cannot be determined according to Fermat's principle from the directions of the incident rays.

As an example, we consider a lens, Fig. 8.6. Here, we refer to a lens as a body of glass whose surface shape is such that all light rays emanating from a point A merge into a point B.

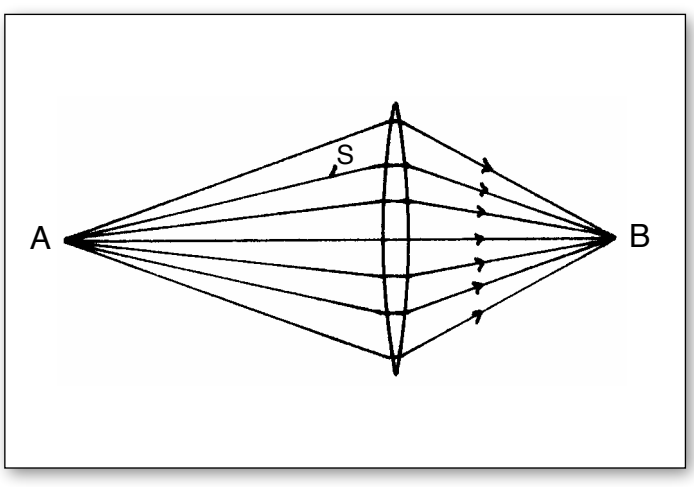


Fig. 8.6
The optical lengths of all rays are the same.

Notice that although such a glass body can be manufactured exactly within the framework of geometric optics, one should not expect the surfaces to be spherical surfaces, as is the case with most technical lenses.

The fact that in our arrangement not only one ray S but also many other rays adjacent to S run from A to B means that the light paths of all these rays are equal because of Fermat's principle. This is an important property of every optical imaging: If one passes from an object point A to an image point B, the light path is the same on all rays.

It should be mentioned already here that, if the lens is designed to image A into B, there is generally no other point A' that is imaged into any point B'. The rays emanating from A' do not intersect at a common point.

8.3 Radiance

Radiance is a quantity used to describe a light field in the context of geometric optics. We introduce it step by step. We first choose a measure for the amount of light: energy. (The following considerations would be the same with other extensive quantities, such as the number of photons or entropy).

The light in the box in Fig. 8.7a has a well-defined energy. In Fig. 8.7b light is leaving the box through the hole, and thus an energy current of strength P flows outwards. If this current is divided by the surface element dA through which it flows, we obtain the magnitude j_E of the energy current density. It is

$$P = \int \mathrm{j}_E \, d\mathbf{A}$$

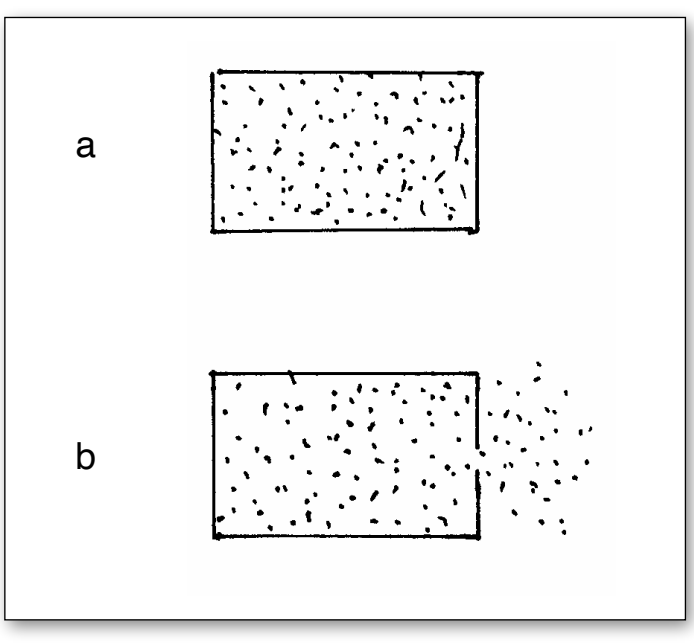


Fig. 8.7
(a) The box contains light. (b) Through the hole, light flows out.

Now the light rays pass through each surface element dA in the various directions. We therefore divide j_E by the solid angle element $d\Omega$ and obtain the energy flux density per solid angle, or in short, the radiance L_E. It is

$$P = \int_A \int_\Omega L_E \, d\Omega \, dA \qquad (8.1)$$

The quantity L_E depends

- on the position in the light field;
- at a fixed position on the direction.

So we have

$$L_E = L_E(\mathbf{r}, \vartheta, \Phi)$$

where the direction in space is characterized by the angles ϑ and Φ.

Figure 8.8 shows a measuring device for L_E. The surface of the lens corresponds to the area dA in (8.1), the photocell area defines $d\Omega$.

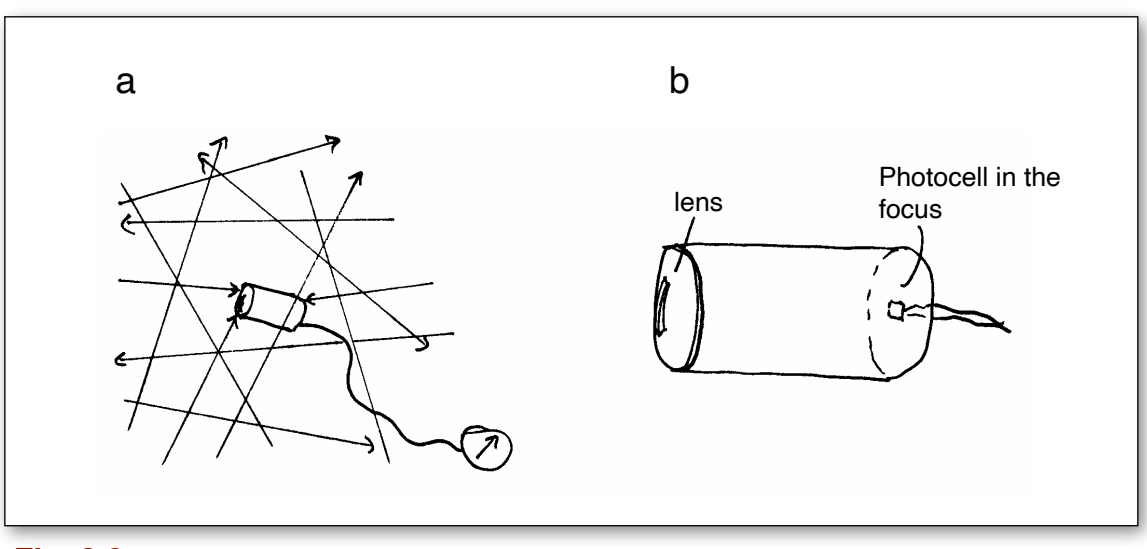


Fig. 8.8
(a) The radiance meter registers light that belongs to a position and direction. (b) Design of the instrument

Often the L_E distribution is rotation-symmetric with respect to an *optical axis*. In this case, two coordinates in the positional space and one angular coordinate are sufficient.

Figure 8.9 shows an example of a radiance distribution. The light comes from a sharply delimited, uniformly radiating surface F, Fig. 8.9a. Fig. 8.9b shows the distribution of L_E at the position z_0 of the z-axis over x and the angle ϑ against the z-axis in perspective. Fig. 8.9c shows a projection onto the x-ϑ plane. In the hatched area L_E has a finite, constant value, outside $L_E = 0$.

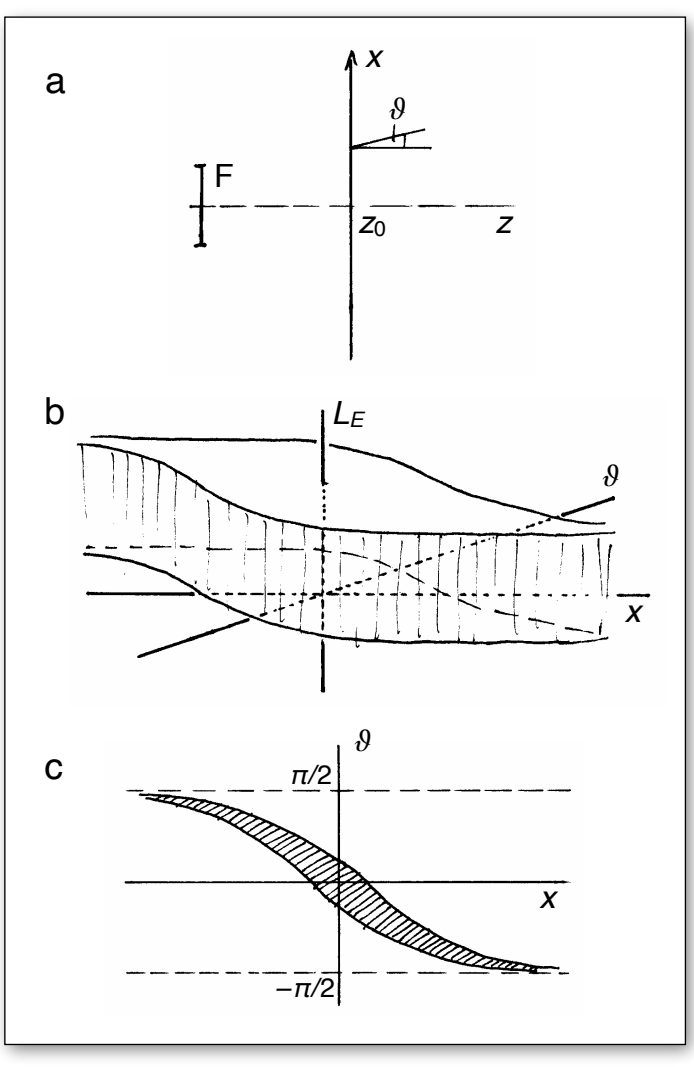


Fig. 8.9
Example of a radiance distribution. (a) The light comes from the uniformly radiating surface F. (b) The radiance is plotted over the position and the direction. (c) Projection into the x-ϑ plane

Figure 8.10 shows a sequence of representations of L_E over ϑ for $x = 0$ at various distances from the light source on the z-axis.

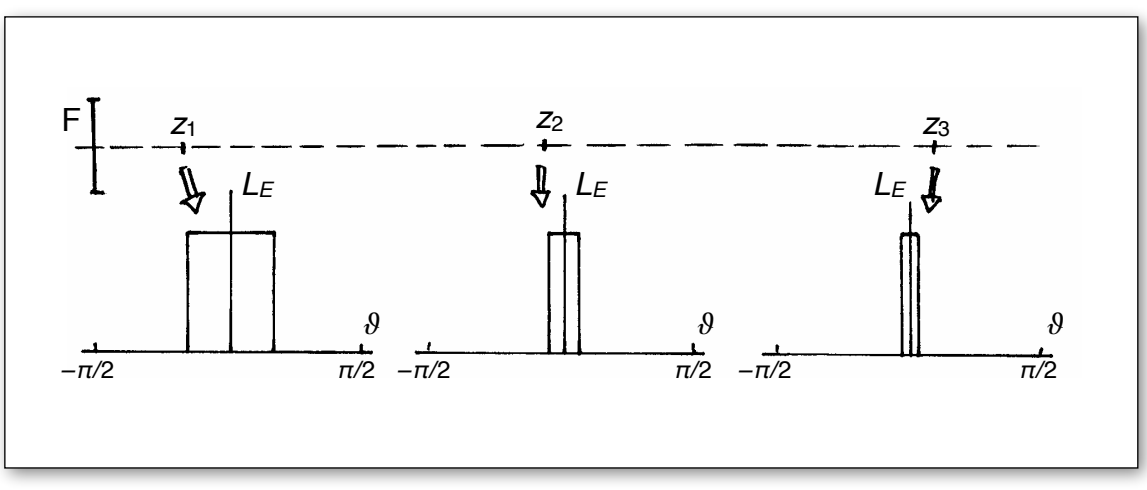


Fig. 8.10
Radiance versus direction for different distances from the illuminating surface F

One can see that the single images differ only in the width of the distribution. The value of L_E in the direction of the z-axis ($\vartheta = 0$) does not change with increasing distance. This is an effect of the following rule: The radiance has the same value at all points of a beam in the direction of the beam.

In this form, however, the rule only applies as long as n is the same everywhere on the beam. The rule can be generalized:

> The quantity L_E/n^2 has the same value at all points on a beam in the direction of the beam.

Here is another consequence of this rule:

One might expect that with a sufficiently large lens one could concentrate as much light from the sun as one wants in one point. If one places an object at this point, one could thus bring it to an arbitrarily high temperature. But this contradicts the 2nd law of thermodynamics. Our theorem $L_E/n^2 = \text{const}$ shows us immediately that this is not possible.

The sequence of images in Fig. 8.11 shows that at best it is possible to arrive at a situation where L_E in P has the same value L_{E0} over the entire solid angle. If one has achieved this, however, the point is in an environment identical to the one directly at the solar surface. Because also on the sun L_E has the same value L_{E0} according to our rule. The point can therefore at most assume the temperature of the surface of the sun; it is then in thermal equilibrium with the sun – and by the way, it radiates back to the sun via the lens and the mirror as much light as it receives from there.

Fig. 8.11
By enlarging the parabolic mirror, light from all directions arrives at point P. The radiance is not changed by the mirror.

Fig. 8.12 finally shows qualitatively what happens to the L_E distribution when the sky is cloudy. On the way from z_1 to z_2 through the clouds the narrow $L_E(\vartheta)$ distribution is smeared over the whole half space.

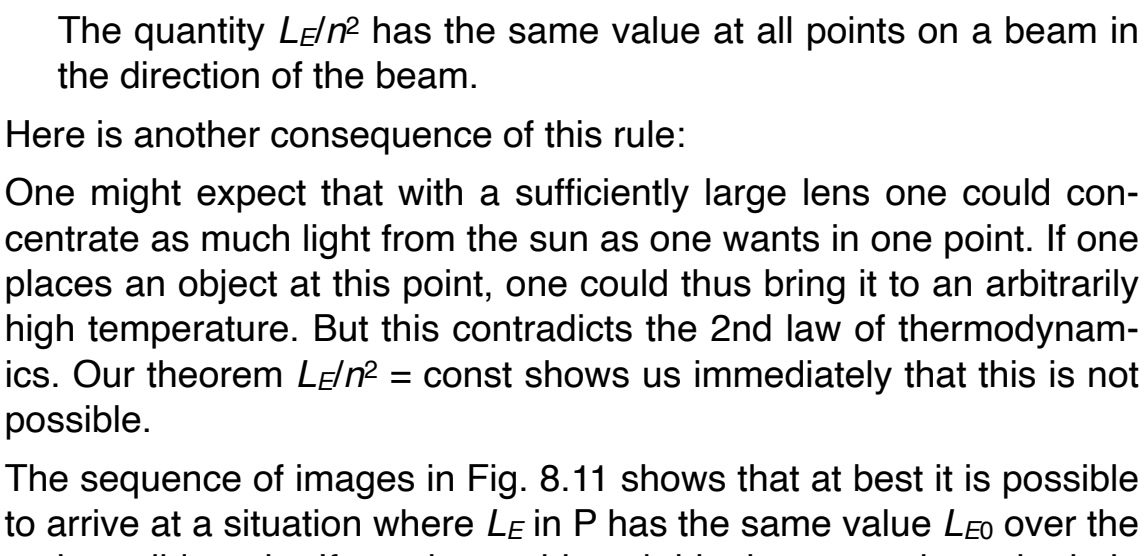


Fig. 8.12
Change in the radiance distribution of sunlight caused by a cloud

9

Imaging

9.1 Collinear imaging

By a mapping of the space one understands a transformation $\boldsymbol{r}' = \boldsymbol{r}'(\boldsymbol{r})$, which unambiguously assigns an image point $\boldsymbol{r}'$ to each point $\boldsymbol{r}$. One says that the two points are "conjugated" to each other. Optics is interested in those mappings that leave geometric figures as undistorted as possible. If one requires that planes are mapped back onto planes (and thus also straight lines onto straight lines), one arrives at the *collinear mapping* or imaging. A collinear mapping is mathematically described by the equations:

$$x' = \frac{a_1 x + b_1 y + c_1 z + d_1}{ax + by + cz + d}$$

$$y' = \frac{a_2 x + b_2 y + c_2 z + d_2}{ax + by + cz + d}$$

$$z' = \frac{a_3 x + b_3 y + c_3 z + d_3}{ax + by + cz + d}$$

However, these mappings still allow for strong distortions. We therefore make further restrictions. The ideal would be the special case of a collinear mapping, where each figure is transformed into a geometrically similar figure, e.g. the following

$$x' = ax \qquad y' = ay \qquad z' = az \tag{9.1}$$

However, this imaging cannot be realized with the means of optics. In contrast, optics can achieve, albeit only approximately, a *centered collinear mapping*. Let us take a closer look at this mapping. It distinguishes

- one axis
- two planes perpendicular to the axis and
- two points on the axis.

If this kind of imaging is realized with light beams, the axis is called the *optical axis*, the planes are called the *principle planes* H and H', the two points are called the *focal points* F and F', the distances HF and H'F' are called the *focal lengths f* and *f'*.

We restrict further to the case where one focal point has the same distance from one plane as the other from the other plane, Fig. 9.1.

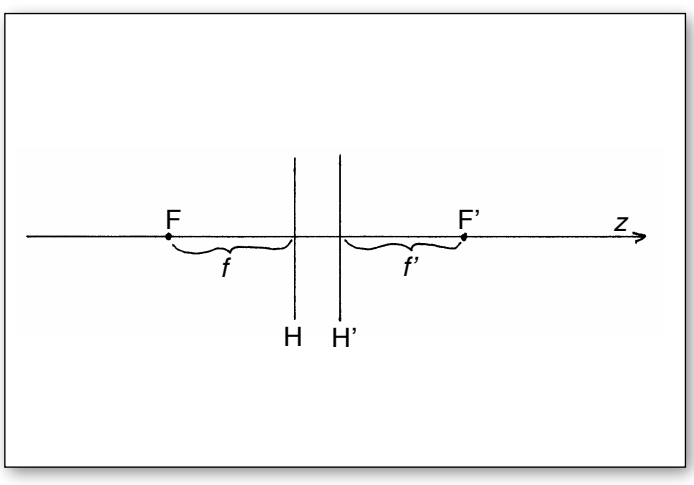


Fig. 9.1
The centered collinear imaging is described by the optical axis, two principal planes and two focal lengths.

This mapping has the property that it images an object plane that is perpendicular to the optical axis into an image plane that is again perpendicular to the optical axis. Furthermore, figures in the image plane are geometrically similar to those in the object plane. If the optical axis is called the *z*-axis, the transformation equations are

$$x' = f\frac{x}{z} \qquad y' = f\frac{y}{z} \qquad z' = -\frac{f^2}{z} \tag{9.2}$$

The coordinates z and z' are measured from the respective focal points. It can be seen that the first two equations (9.2) have the structure of the first two equations (9.1). That means the *x*-*y* plane is not distorted by the mapping. Distances in the z direction, however, are distorted, as can be seen in the third equation (9.2).

From (9.2) follows the well-known procedure of constructing an image point P' from the corresponding object point P, Fig. 9.2.

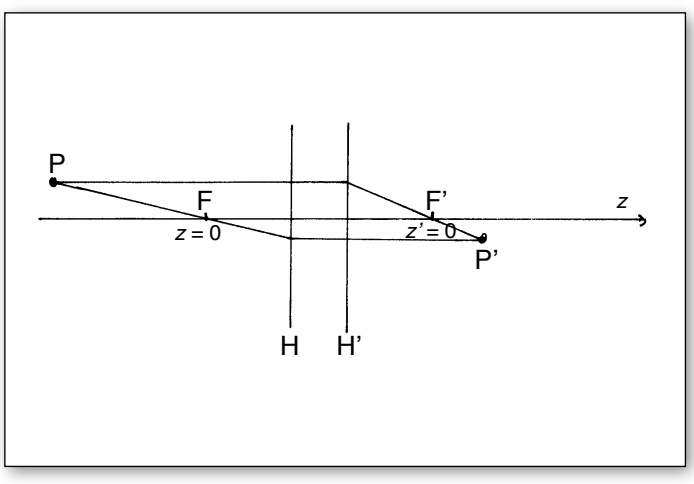


Fig. 9.2
To construct an image point P' from an object point P

In addition, from (9.2) follow some other well-known equations. With the object distance g, the image distance b, the object height G and image height B defined in fig. 9.3 the coordinates of the conjugate points P and P' become

$$x = G \qquad x' = -B \qquad z = -(g - f) \qquad z' = b - f \tag{9.4}$$

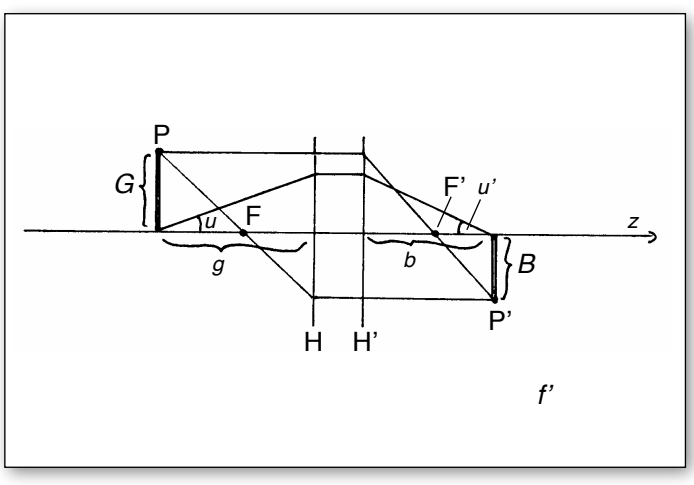


Fig. 9.3
The definition of the object distance g, the image distance b, the object height G and the image height B

With (9.2) we obtain

$$B = f \cdot \frac{G}{g - f} \tag{9.5}$$

and

$$b - f = \frac{f^2}{g - f} \tag{9.6}$$

From (9.6) follows:

$$(b - f) \cdot (g - f) = f^2$$

$$bg = f(b + g)$$

and

$$\frac{1}{r} = \frac{1}{b} + \frac{1}{g} \tag{9.7}$$

Transforming (9.5) results in

$$\frac{G}{B} = \frac{g - f}{f} = \frac{g}{f} - 1$$

The right side of this equation can be replaced using (9.7) by g/b

$$\frac{G}{B} = \frac{g}{b} \tag{9.8}$$

We also infer from Fig. 9.3 the relationship:

$$g \cdot \tan u = b \cdot \tan u'$$

With (9.8) and (9.4) we obtain

$$x \cdot \tan u = x' \cdot \tan u' \tag{9.9}$$

9.2 Realizing collinear imaging

An imaging to which equations (9.5) to (9.9) apply can be approximated by light rays passing through a system of spherical mirrors and refracting surfaces.

Spherical surfaces are chosen because they are much easier to produce than other shapes, but an exact collinear image cannot be achieved even with non-spherical lenses or mirrors, because it would violate fundamental laws of nature: either the first or the second law of thermodynamics.

But what does that mean: a mathematical image can be realized with light? While the mathematical image simply assigns a point P' to a point P by a mathematical operation without any physical connection between the points, the optical image means that light rays that start from point P pass through the lens system at various points and then meet again at a point P'. Of course, this does not mean that the corresponding light is only to be found at points P and P'. It is rather everywhere in space. That the optical imaging takes place can be seen by moving a screen (or other detector) around in space. If it is positioned in such a way that it contains P', the light distribution is "point-like". If the screen moves away from this point, the light distribution expands, "the image of P becomes blurred".

The limitations of the collinear imaging by lenses is not only that the image of an extended object is distorted, but above all that it is impossible to combine the light from more than one single object point into an image point.

Calculating the parameters of such a mapping from the refractive index and the geometry of a lens is somewhat laborious. Here we only quote the most important result for the special case of a "thin lens", i.e. a lens in which the distance between the principal planes is small compared to the focal length:

$$\frac{1}{r} = (n-1)\left(\frac{1}{r_1} - \frac{1}{r_2}\right) \tag{9.10}$$

f is the focal length of the lens, n the refractive index of the material of the lens. r_1 and r_2 are the radii of curvature of the lens. In optics, the radius of curvature of surfaces curved to the left is counted positively, that of surfaces curved to the right negatively. Thus in fig. 9.4a $r_1 = +3$ cm, $r_2 = -4$ cm, in Abb. 9.4b ist $r_1 = -5$ cm und $r_2 = -7$ cm.

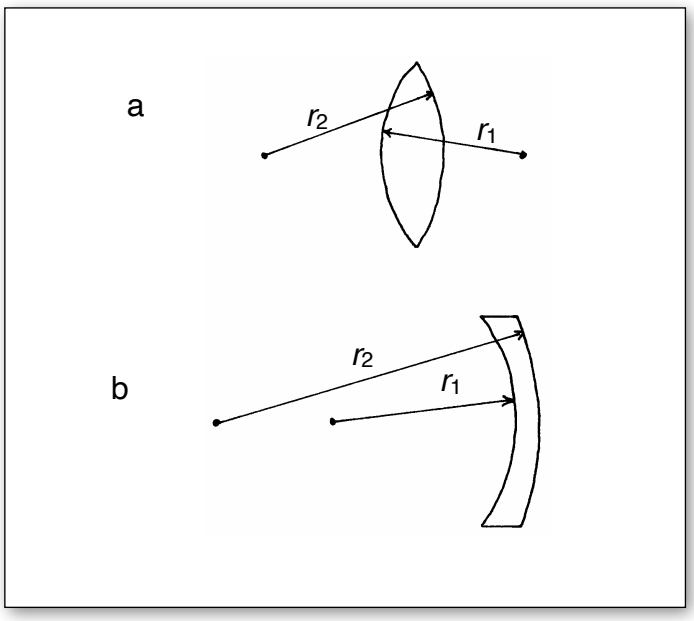


Fig. 9.4
The radius of curvature of surfaces curved to the left is counted positive, that of surfaces curved to the right negative.

We infer from equation (9.10) that the focal length of a lens is positive if it is thicker in the center than at the edge, otherwise it is negative.

Several lenses placed one behind the other form an *optical system*. An optical system also produces a collinear imaging. To construct the image produced by an optical system, as with a single lens, knowledge of a single focal length and the position of two main planes is sufficient. If the collinear imaging by lenses were perfect, two lenses would be sufficient to realize any optical system: with the two single focal lengths and the distance of the lenses, one has enough parameters to give any value to the focal length and the distance of the main planes of the system.

Nevertheless, optical systems often consist of many more than two lenses: the so-called *optical aberrations* are corrected by additional lenses with sometimes different refractive indices.

Optical systems have different names depending on their function and properties: objective, condenser, eyepiece, beam expander etc.

9.3 Diffraction limit of imaging

In the previous section we have seen that light rays realize approximately a collinear imaging. We had assumed from the outset that the approximation $\lambda = 0$ is justified. We now want to investigate the influence of the fact that λ is not equal to 0. Again, we will represent the light by rays. They now represent the normals to the wavefronts. To avoid several complications at the same time, we will assume that the laws of collinear imaging are exactly fulfilled for these “rays”.

An object that is partially transparent is illuminated from the left with coherent light and imaged by means of a lens on a screen, Fig. 9.5.

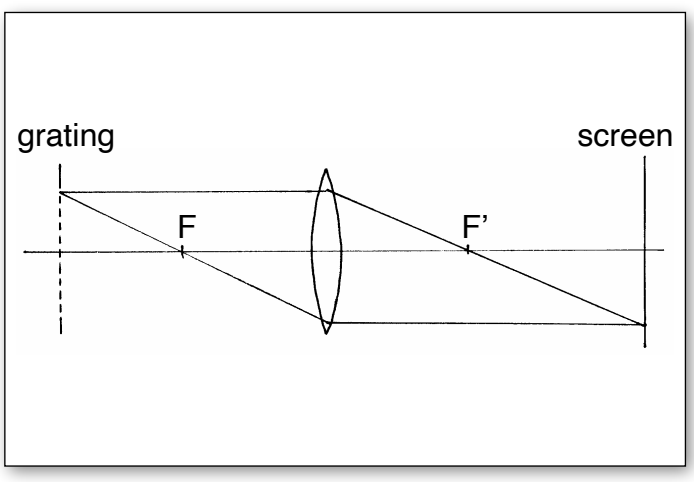


Fig. 9.5
The grating is imaged in the plane of the screen.

For the sake of clarity we choose a diffraction grating as the object. It is immediately obvious that the arrangement is identical with the Fraunhofer diffraction arrangement, Fig. 7.13, except that the screen is not in the focal plane but in the image plane of the grating. But this does not change the fact that the diffraction image is situated in the focal plane.

We now ask for the limitations of the imaging, that originate from the wave nature of the light. For this purpose we imagine to make the lattice constant d of the lattice smaller and smaller. The angle at which the light is diffracted at the grating will then increase. This causes the higher diffraction orders, one after the other, to move out of the lens. In fig. 9.6 only the zeroth and the first order hit the lens. The higher orders are missing, their contribution to the diffraction image in the focal plane disappears, and they do no longer contribute to the image formation in the image plane.

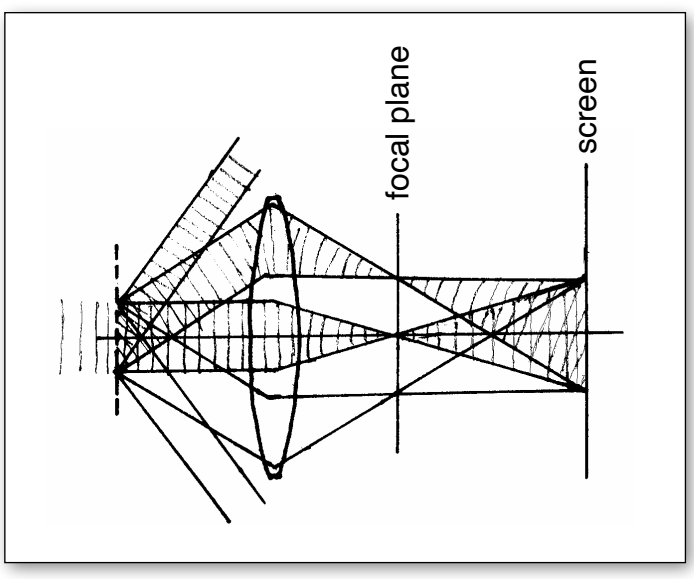


Fig. 9.6
Only the light of the zeroth and first diffraction order still goes through the lens. The higher orders do not contribute any more to the formation of the image.

If the lattice constant is reduced even further, the first-order beams will eventually move out of the area of the lens and only the zero-order will remain.

Which consequence has the disappearance of the higher diffraction-orders for the image? The answer to this question is given by the Fourier theory. To reconstruct the perfect image we need all its Fourier components. As the higher components are taken away, sharp changes of the energy flux density in the image plane are smeared out.

If only the 0th and the 1st diffraction order contribute to the image formation, the image has a sinusoidal intensity distribution, which shows no more than the periodicity of the original grating. If finally only the 0th order is left, the image plane becomes uniformly bright. If one looks at it, one only gets to know something about the average brightness of the object.

This can also be expressed as follows: the lens only allows the low *spatial frequencies* to pass through. The communications engineer calls a component that only allows low temporal frequencies to pass a low-pass filter. So the lens is a low-pass filter for spatial frequencies.

Let us look at these things from yet another angle. We imagine that instead of the grating there are the frames of a movie, alternating in rapid succession. Now a data flow (the quantity measured in bit/s) is flowing through our optical system. However, the data flow that comes out at the back of the lens is smaller than the one that arrives at its front. Part of the data does not get through, it falls on the edge of the lens.

One can see that the data flow would be limited even with an infinitely large lens. If one considers details of the object whose size is equal to the wavelength of the light, i.e. structures of the size $d = \lambda$, the angle of the 1st diffraction order becomes equal to 90° because of

$$\sin\vartheta = \frac{\lambda}{d} = \frac{\lambda}{\lambda} = 1$$

The ultimate limit of the imaging has been reached. Thus, it is not possible to image structures whose size is smaller than the wavelength of the radiation. With a light microscope a resolution of about 1 µm can therefore be achieved. In order to examine smaller structures, one must use other radiations than light: electrons, protons, etc. Their wavelength decreases with increasing energy. Therefore, these particles are often brought to very high energy in so-called accelerators.

But the fact that (with coherent illumination) one encounters the diffraction image in the focal plane does not only allow to determine the limits of the imaging. It also gives us a means to manipulate images: by masking parts of the diffraction image. Some satellite images are composed of many parallel stripes. The stripe structure disturbs when viewing the image. We now make an image of the satellite photo with coherent light. The fringe pattern is expressed in the diffraction image in the focal plane by a sequence of equidistant points. If these points are suppressed in the diffraction image, the fringe pattern disappears in the image plane.

9.4 The resolving power of optical instruments

We will look at a result of the previous section.

An object is to be imaged which has a structure of the size. We assume that the object consists of two light-emitting points with a distance G between them. In order to obtain images of the points that are still distinct, we need a lens that allows at least the first order of interference to pass.

In this case one says that the two object points are just *resolved*.

We now ask what this requirement means for two extreme cases of optical imaging, namely the image in the microscope, Fig. 9.7a, and the image in the telescope, Fig. 9.7b.

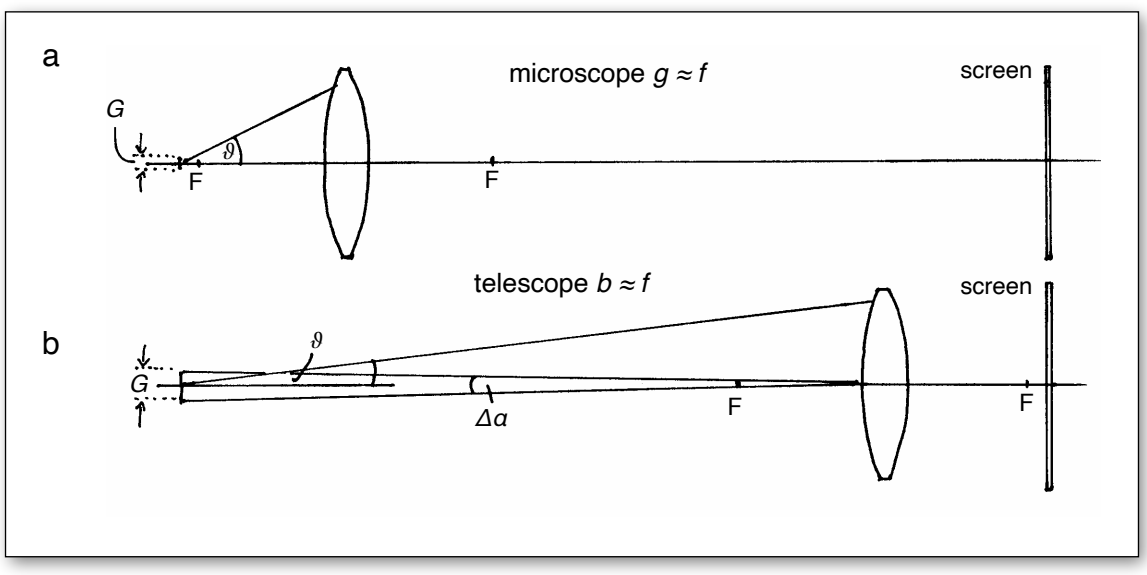


Fig. 9.7
Resolving power of (a) the microscope and (b) the telescope

In the microscope, the object is almost in the focal plane, so we have $g \approx f$, and the image is at a distance that is large compared to the focal length.

In the telescope, the object is at a distance large relative to f, and the image distance is practically equal to the focal length: $b = f$.

With a microscope one usually asks for the minimum distance G_{min} that two object points may have in order to be resolved, with a telescope, one asks for the minimum angle $\Delta\alpha_{min}$ at which they appear when viewed from the telescope.

In both cases we obtain for the angle ϑ of the 1st interference order

$$\sin\vartheta = \frac{\lambda}{G_{min}} \tag{9.11}$$

If R is the lens radius, we have:

$$\tan\vartheta = \frac{R}{g}$$

We now set $\sin\vartheta \approx \tan\vartheta$. This approximation is very well fulfilled for the telescope. But it is also justified for the microscope, because we want to determine G_{min} only approximately.

From (9.11) and (9.12) we thus obtain

$$\frac{\lambda}{G_{min}} = \frac{R}{g}$$

For the microscope we have $g \approx f$, and we get

$$G_{min} = f \cdot \frac{\lambda}{R} \quad \text{microscope} \tag{9.13}$$

For the telescope we ask for the angle $\Delta\alpha_{min} = G_{min}/g$. We thus get

$$\Delta\alpha_{min} = \frac{\lambda}{R} \quad \text{telescope} \tag{9.14}$$

What happens if the two light emitting points are closer together than the relationship (9.13) or (9.14) suggests?

The image of a single point P is not a point, but a small spot, which can be imagined to be formed by the diffraction of light from P at the edge of the lens. If two light sources are closer together then corresponding to equation (9.13) or (9.14), their *diffraction disks* overlap in such a way that they can no longer be recognized as images of separate points.

Equations (9.13) and (9.14) are of fundamental importance. We formulate them again in the form of a rule:

The resolution is all the better

- the larger the opening diameter of the imaging system;
- the smaller the wavelength of the radiation used.

Equation (9.14) represents a statement of information theory. It is not restricted to a specific method of measuring angles. So it does not only set an upper limit for each telescope, but also for the Michelson stellar interferometer (section 7.4.5).

The statement that one detector can distinguish between two objects that have the angular distance $\Delta\alpha_{min}$ is equivalent to the statement that the detector can localize a single object with an angular accuracy $\Delta\alpha_{min}$. We want to illustrate this statement by making an acoustic experiment. Hidden behind a curtain is a loudspeaker that emits a sound of about 600 Hz and a “bandwidth” of about 100 Hz. With the help of two microphones and a two-beam oscilloscope we want to determine the direction in which the loudspeaker is located, Fig. 9.8.

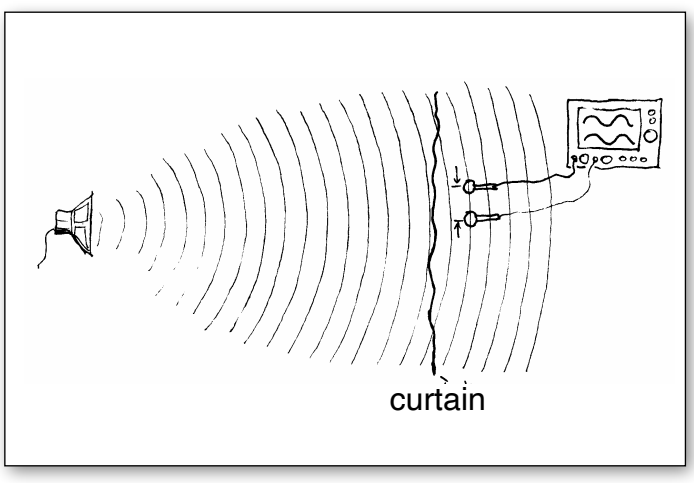


Fig. 9.8
Two microphones are used to determine the direction in which the loudspeaker is located.

The wavelength of the acoustic wave is about 1/2 m. If the distance d between the microphones is much smaller than 1/2 m, both will always register the same signal, no matter how one turns them against each other. If the distance is greater than 1/2 m, however, the microphones deliver different signals depending on the direction of their connecting line. By rotating this line one can determine the direction of the wave normal with an accuracy of about

$$\Delta\alpha = \frac{\lambda}{d}$$

These considerations apply to the observation of the sky with a telescope as well as to the localization of an airplane with radar, an earthquake focus with the help of seismometers or a radio transmitter by radio direction finding.

10

Optical instruments

10.1 The camera

We want to make a two-dimensional image of a three-dimensional object, i.e. a projection. In addition to the limits of the collinear imaging mentioned so far (first: a limit given by the 2nd law; second: a limit coming from the finite wavelength of the light), another limitation emerges: even with a perfect collinear imaging, one would obtain a sharp “image” in three dimensions and not in two, Fig. 10.1.

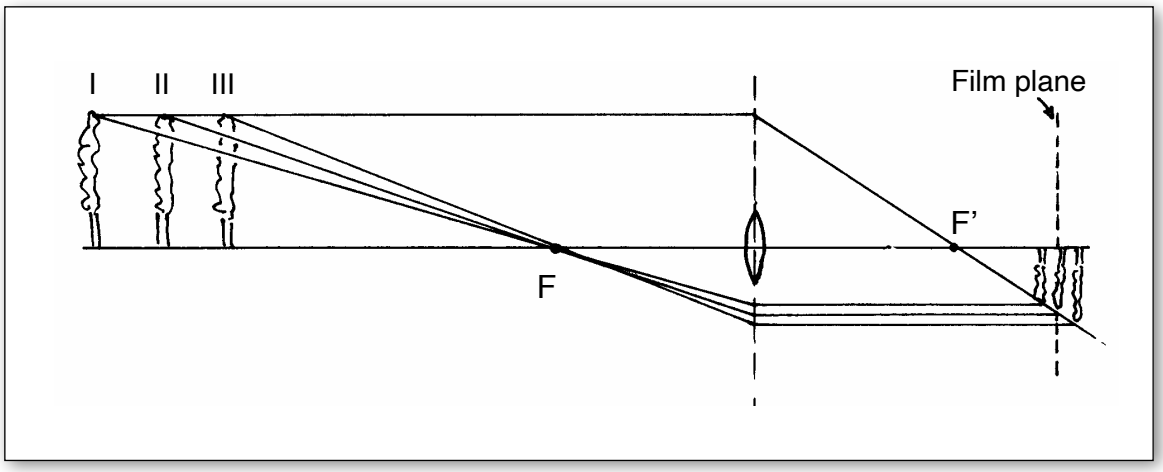


Fig. 10.1
Even with a perfect collinear image, a sharp image is only produced in one plane.

In the plane of the film only the tree number II is sharp. Tree I and tree III are out of focus. The depth range that is still sufficiently sharp is called the *depth of field*. The depth of field increases as the aperture diameter of the lens decreases. For very small apertures, the collinear image becomes the special case of a projection.

A normal photo lens should be able to image an angular range of about $2\varphi = 30°$. The film is located approximately in the focal plane, Fig. 10.2.

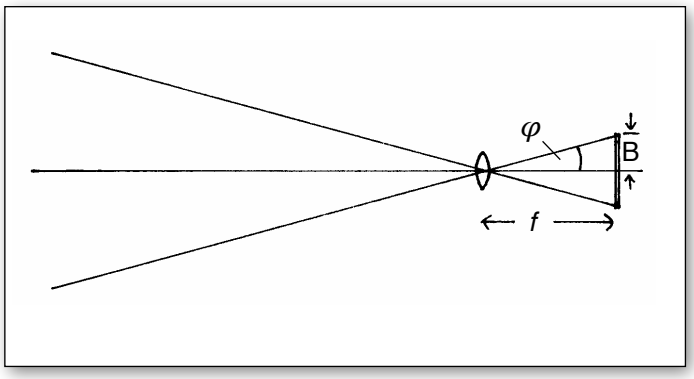


Fig. 10.2
In the camera, the film is located near the focal plane.

From the desired size of the negative follows the focal length that the lens must have. For a 35 mm film with $B \approx 15$ mm, this results in $f = B/\tan\varphi \approx 50$ mm.

The diameter of a photographic lens should be as large as possible: Since one wants to take pictures of moving objects, the exposure time must be small; therefore, the energy required to expose the film must pass through the lens in a short time. Since a larger lens diameter results in a smaller depth of field, the effective lens diameter can be adjusted with the aperture and thus any compromise between energy flow and depth of field can be chosen.

The aperture scale carries the series of numbers …2.8; 4; 5.6; 8; 11; 16.... These numbers do not indicate the objective diameter D itself, but the ratio f/D. The numerical values are chosen so that the energy flow doubles from one number to the next. So when the aperture is 5.6, twice as much energy passes through the lens as with aperture 8, and at aperture 4 twice as much as with aperture 5.6.

Diffraction at the lens aperture is not a major limitation of the camera.

10.2 Image projectors

We begin by looking at a projector that is out of use today: the opaque projector, Fig. 10.3.

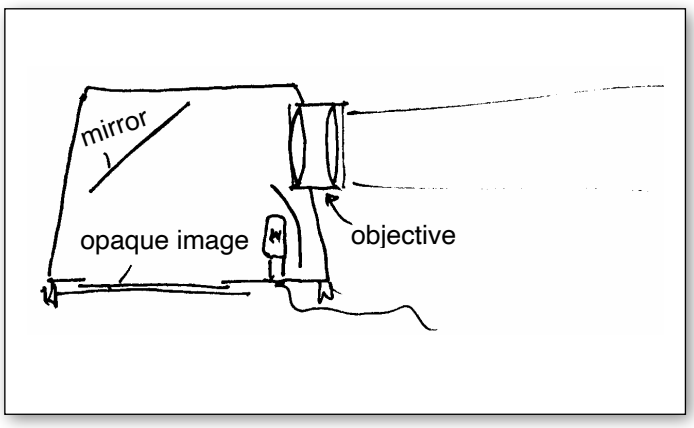


Fig. 10.3
Opaque projector

It was used to project drawings or printed paper images onto the wall. The lens provides an optical image of the paper image on the projection screen. The problem with this projector is that it is difficult to get enough light onto the screen. Even though very strong lamps are used for the illumination of the object and a lens with a large diameter for the projection, the image on the wall is rather faint. The light that comes from the lamps is diffused by the object in all directions and the objective, despite of its big diameter, only gets a small part of it.

This is why in the case of the overhead projector and the slide projector, one makes sure that all light that falls on the object also passes through the objective. To achieve this, two things are necessary.

First, the object must not scatter the light. Slides and overhead projector transparencies let the light pass straight through in those places that should appear bright, in other places it is absorbed. But it is never scattered.

Secondly, one has to make sure that all the light that has passed through the object reaches the lens. This is achieved by means of the condenser, Fig. 10.4. The condenser is a lens as large as the object and is located close behind the object. It images the very small light source onto the objective aperture. Thus, all light that passes through the object also passes through the objective. This type of projector not only produces much brighter images than the opaque projector, it is also much cheaper.

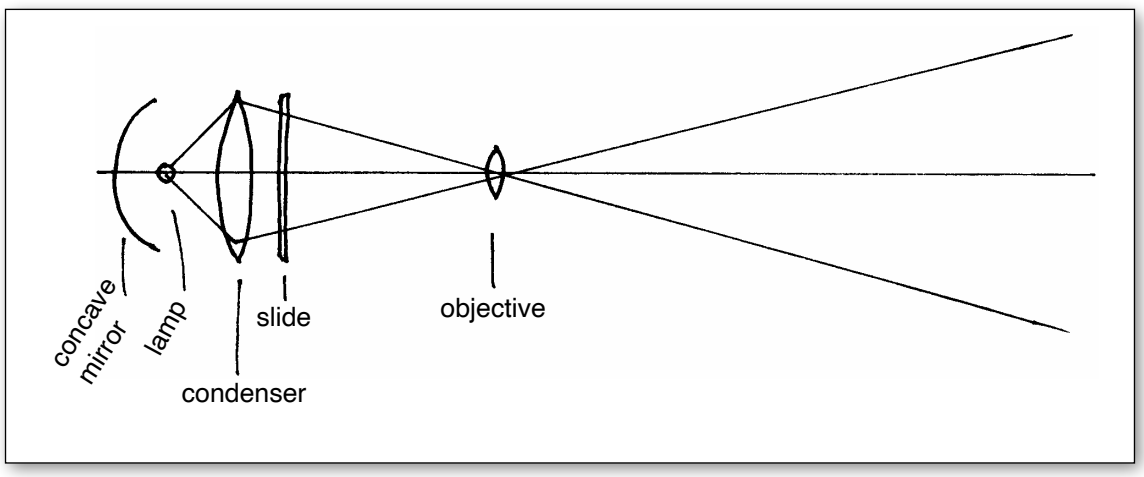


Fig. 10.4
The condenser ensures that all the light that hits the slide passes through the objective.

Why is it cheaper? On the one hand, one can use a very small diameter lens, i.e. a cheap lens. On the other hand, the condenser lens does not contribute to the imaging of the object on the screen. Thus, the condenser has to be big, but it does not need to be corrected and therefore it is not expensive. In the case of the overhead projector, the condenser is simply a Fresnel lens made of plastic.

A Fresnel lens can be imagined to have originated from ring-shaped parts of an ordinary lens, whereby in each ring as much as possible of the superfluous glass was taken away. The refractive surfaces have the same angle against the optical axis as with the real lens, fig. 10.5. Fresnel lenses can also be found in car headlights and lighthouses.

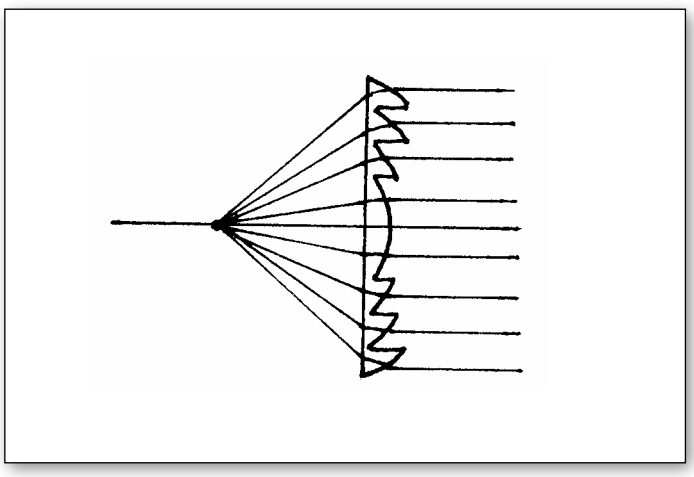


Fig. 10.5
Cross section of a Fresnel lens

10.3 The telescope

It is used to detect the radiation emitted by stars. It is supposed to take an image of a section of the sky and always collect as much radiation as possible.

A distinction is made between lens and reflecting telescopes. Lens telescopes have the advantage that image errors can be easily corrected. They are therefore suitable for imaging large fields of the sky. Due to the mechanical instability, however, lens telescopes with large diameters cannot be built.

The large telescopes are all mirror telescopes, mostly with a parabolic mirror. A parabolic mirror can image a sky field of up to 10 arc-minutes extension sufficiently well.

Telescopes have long focal lengths: from about 1 m to over 100 m. It depends on the focal length what section of the sky is “seen” by the detector. The diameter D of large telescope mirrors is several meters (Hobby Eberly telescope, MacDonald observatory: D = 11 m). The larger the cross-sectional area of the mirror, the greater is the light flux captured from a star.

Although the resolving power of the mirror should theoretically become increasingly better with an increasing diameter, the diameter of the mirror is not important for the resolution. The resolution is much more limited by density fluctuations of the air in the atmosphere, and for diameters larger than 12 cm the resolving power does no longer increase with the diameter. The reason why the mirrors are so large is to collect enough energy from a star within a reasonable period of time.

With a telescope one can therefore see many more stars in the sky than without. With the naked eye one can see about 6000 stars in the whole celestial sphere, with a telescope one can register millions of stars.

Astrophysics retrieves as much data as possible from the sky. It does not limit itself to the visible spectral range, but studies radiation of all wavelengths for which the atmosphere is transparent. This is the case not only in the visible range but also in the radio range with wavelengths from 1 mm to 30 m. Fig. 10.6 shows the height at which the radiation falling on the earth from outside is attenuated by a factor of 1/e.

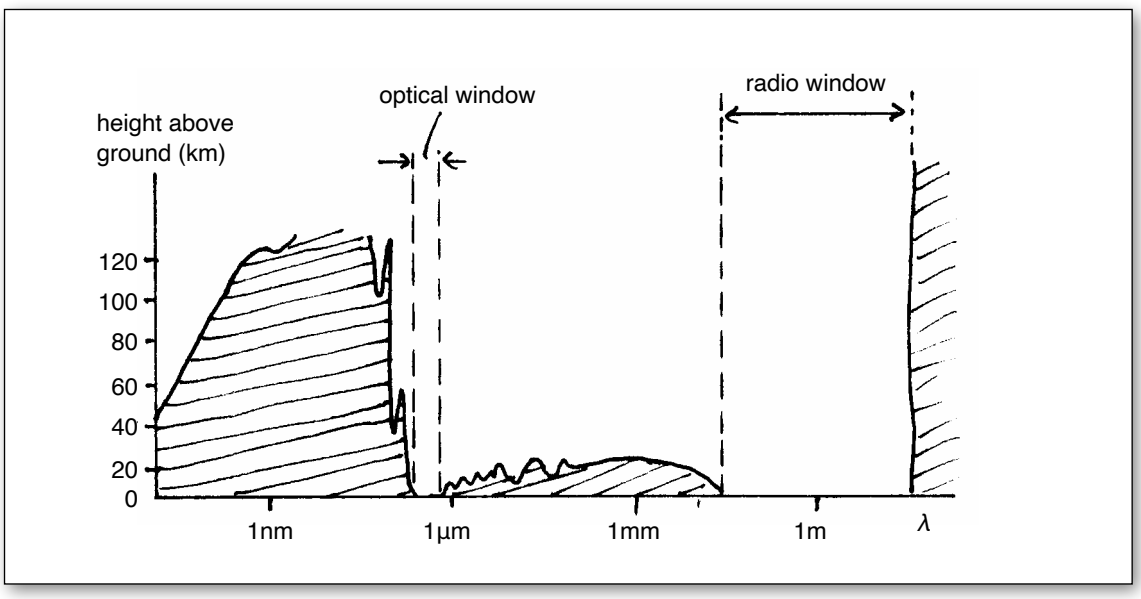


Fig. 10.6
Absorption spectrum of the earth’s atmosphere

Most radio telescopes are designed just like optical telescopes: the most important component is a parabolic mirror. This mirror is made of metal, and it is much larger than in optical telescopes. Because of the longer wavelength, its surface does not need to be as precisely parabolic as in an optical telescope. It may even contain holes, as long as they are not larger than about $\lambda/20$.

The detector in the focal “point” of the mirror of a radio telescope is not larger than the diffraction disk. In order to obtain an extended image of a section of the sky, one must therefore scan this section with the telescope.

The largest parabolic reflector radio telescope in the world is located in Kedu (China, province Guizhou). Its mirror has a diameter of 500 m. For λ = 21 cm (the wavelength of an emission line of neutral hydrogen which is important for astrophysics) this mirror has a resolution of about 2.8 arc minutes. Its resolving power is therefore much worse than that of an optical telescope. However, in the radio range one sees objects and phenomena that are invisible at optical wavelengths.

The resolving power can be improved by bringing the signals of two telescopes placed at a great distance from each other to interference. Since the detector registers the amplitude and phase of the beams, this interference can be managed electronically. By correlating the signals from telescopes in different parts of the earth, a resolution of up to 10 arc seconds can be achieved.

10.4 Beam expander

Most lasers make a very thin parallel beam of light, but for many purposes a wide beam is needed. Therefore it is necessary to use a beam expander, Fig. 10.7.

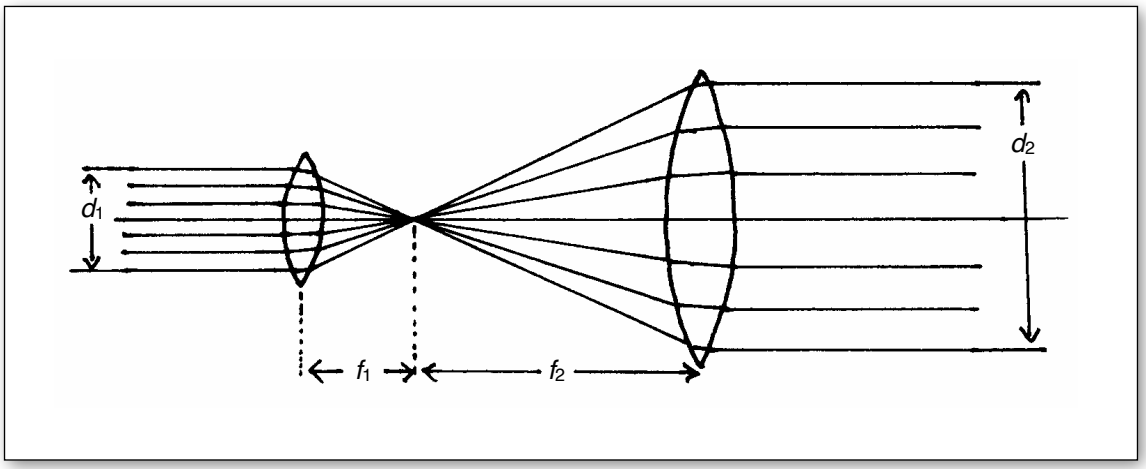


Fig. 10.7
Beam expander

Two lenses are placed in such a way that the distance between them is equal to the sum of their focal lengths. From fig. 10.7 it can be seen that

$$\frac{d_1}{d_2} = \frac{f_1}{f_2}$$

The focal length of the entire system is, like that of a plane-parallel glass plate, infinite. For beam expanders, however, the main planes also lie at infinity.

10.5 The system “eye + magnifying glass”

A magnifying glass is a lens which, together with the eye, forms a lens system with which a small object is to be imaged onto the retina of the eye. The object is placed in one focal plane of the magnifying glass. The object can be seen sharply when the eye is accommodated to infinity, Figure 10.8a.

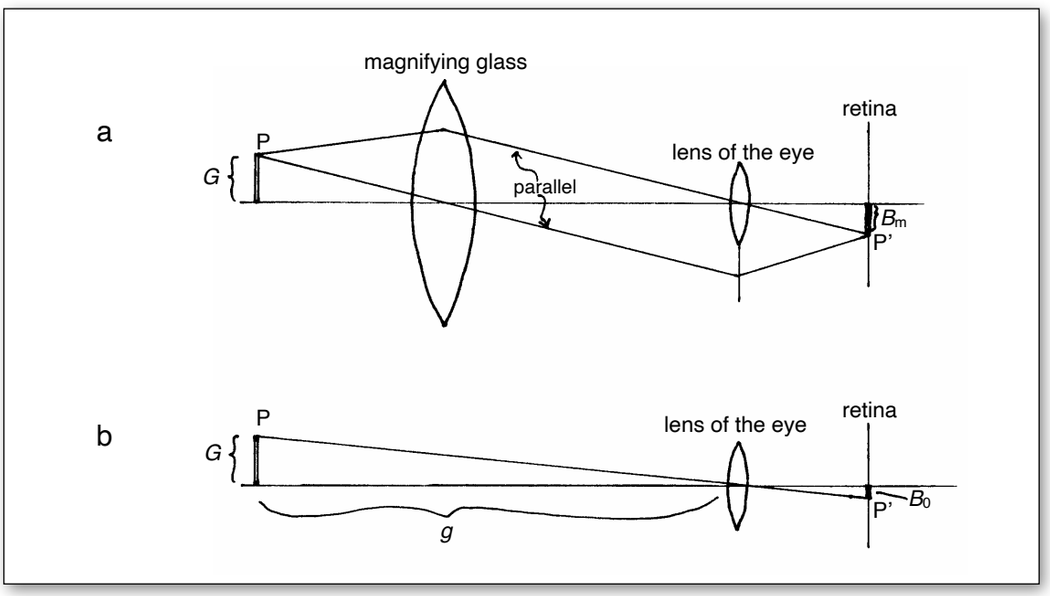


Fig. 10.8
Imaging an object by means of the lens of the eye (a) with and (b) without a magnifying glass

The benefit of a magnifying glass can be seen by comparing the size B_m of the retinal image with an additional lens with B_0, the image size without the additional lens. To construct the image point P’ of P, the rays through the center of the lens were used: these do not change their direction when passing through the lens. From figure 10.8a one can see:

$$\frac{B_m}{G} = \frac{f_{eye}}{f_{mag}}$$

Without a magnifying glass, Fig. 10.8b, one obtains

$$\frac{B_0}{G} = \frac{f_{eye}}{g}$$

If the distance g between eye and object is not changed, the magnifying glass causes a magnification

$$\frac{B_m}{B_0} = \frac{g}{f_{mag}}$$

Since the light coming from an object point is parallel between the magnifying glass and the lens of the eye, the size and sharpness of the image does not change when one approaches the magnifying glass with one’s eye; only the field of view becomes larger.

Thus, in order to use a magnifying glass correctly, one has to pay attention to two points:

- The object must be in the focal plane so that the eye is relaxed.
- The eye should be close to the magnifying glass so that the section of the image is as large as possible.

10.6 The eyepiece

Figure 9.7 shows the most important part of the microscope and telescope. Instead of the screen in this figure one may think of placing a photo film. However, if one wants to look at the image directly instead of taking a photo, one has to extend the instruments further. If one really places a white screen at the position shown in fig. 9.7, one will not see much. The light arriving there is scattered in all directions and only a tiny fraction of it gets through the pupils into our eyes: the image is very faint. Therefore, a so-called eyepiece is placed in the place of the screen. An eyepiece consists of (at least) two lenses with clearly separate functions: the eye lens and the field lens, Fig. 10.9a.

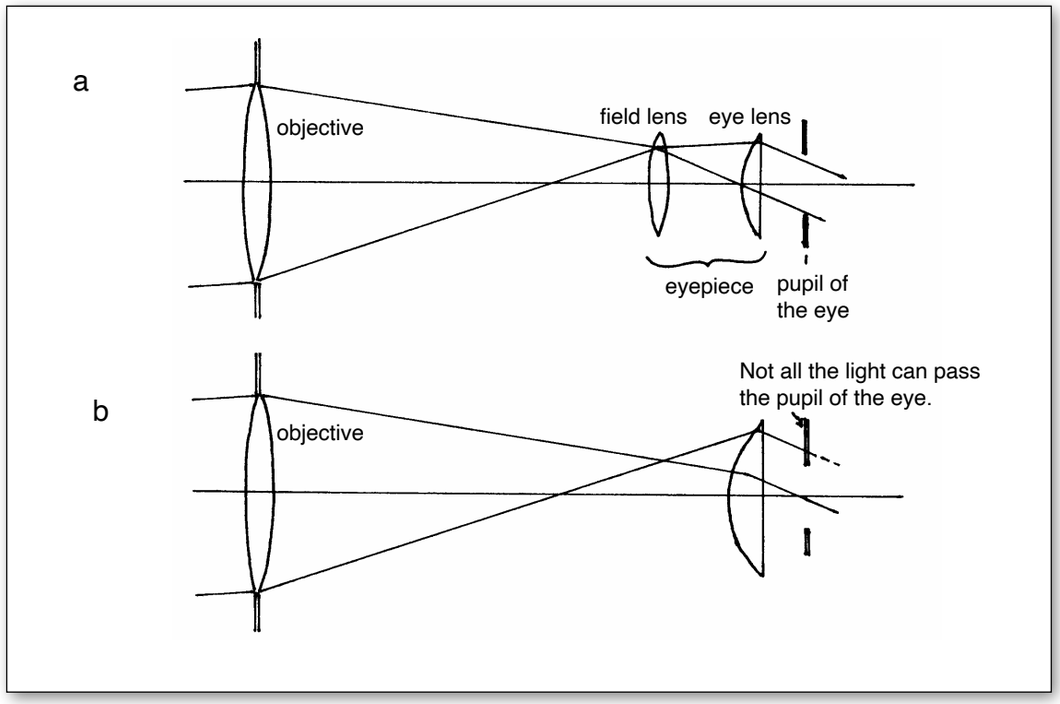


Fig. 10.9
(a) The eyepiece consists of two lenses with clearly distinct functions: the eye lens and the field lens. (b) Without a field lens, part of the light does not reach the pupil of the eye.

The eye lens is nothing more than a magnifying glass with which one views the image created by the objective. The function of the field lens can be compared to that of a condenser. Without it, Fig. 10.9b, the eye lens would have to be very large, and to see the different parts of the image brightly, one would have to move the eye back and forth in front of the lens. The field lens directs the entire energy flow towards the small eye lens, but without changing anything in the optical imaging by the objective and eye lens.

Here we have constructed the image created by the telescope from the images obtained by the subsystems (objective) and (eye + eyepiece).

The telescope alone, i.e. the system (objective + eyepiece) is essentially identical to the beam expander discussed in 10.4, but operated in reverse, i.e. as a beam compressor. This approach reveals an important function of binoculars: to collect light of a certain direction on a wide area and compress it so that it fits through the small pupil opening of the eye.

11

Special procedures

11.1 Radar and scanning electron microscope

An image of a non-self-luminous object can be created using two different methods.

Either one illuminates the whole object and analyzes the light backscattered from the object by direction: One measures the intensity of the light as a function of the direction from which it comes, Fig. 11.1.

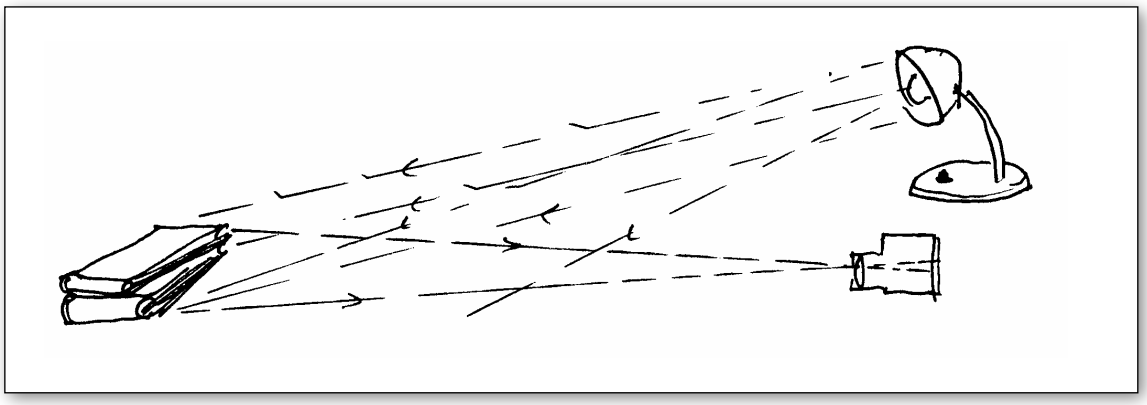


Fig. 11.1
The whole object is illuminated. The scattered light is analyzed by direction.

Examples:

- a photo taken with a flash;
- a television shot where the scene is lit by lamps;
- the observation of an object through an ordinary microscope.

With the second method, the illumination is directional: the object is scanned with the thinnest possible beam. The detector, on the other hand, does not have to distinguish the directions from which the radiation backscattered by the object comes, Fig. 11.2.

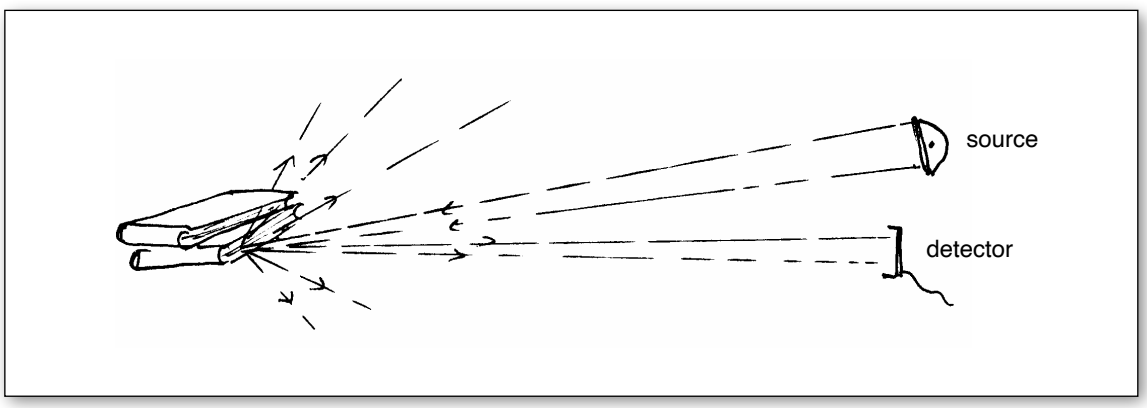


Fig. 11.2
One spot of the object after the other is illuminated. The scattered light does not need to be analyzed by direction.

Examples:

- Radar (RAdio Detection And Ranging). The radiation used is electromagnetic waves with wavelengths from a few mm to a few m. The beam is generated by a parabolic mirror. The scanning is done by rotating the mirror. The radiation is emitted in pulses (pulse repetition frequency some 100 Hz) and from the propagation time of the pulses one determines the distance of the object. The Doppler effect can also be used to determine the velocity of the object.
- The scanning electron microscope. Electrons with wavelengths between 0.004 nm and 0.02 nm are used as radiation. The beam diameter is about 10 nm. There are different possibilities for the detector: either the secondary electrons emitted by the object are registered or the luminescence radiation. The images produced by these scanning methods are real projections of the object. Therefore one obtains a very large depth of focus. This is one of the important properties of scanning electron microscopes.

11.2 Phased array antennas

The surface elements of a parabolic antenna mirror can be imagined to be single antennas whose signals are brought to interference at the focal point. Instead of arranging these single antennas on a parabolic surface, they can also be arranged on any other surface, e.g. a plane. In this case, one only has to manage the interference correctly by some other means. This is done by placing a large number of small antennas on a large flat surface and making their signals interfere electronically. This type of antenna is only suitable for radio waves, because there are no phase sensitive detectors for light waves. Furthermore, the high frequencies of light cannot be processed electronically. Such array antennas are especially used as transmitting antennas for radar systems. By superimposing the waves generated by the individual antennas, a sharp beam of a certain direction is created, just as a narrow beam of light is created behind a diffraction grating out of many spherical elementary waves. By controlling the phase relation between the individual antennas, the resulting beam can be oriented in any direction. This control is much faster than the mechanical orientation of the beam in ordinary radar.

11.3 Optical fibers

Light can be transmitted through thin fibers made of optically highly transparent material. It follows the fiber even if the fiber is curved. As long as the diameter of the fiber is large against the wavelength, it is convenient to consider the propagation process as a sequence of total reflections on the inner surface of the guide, Fig. 11.3.

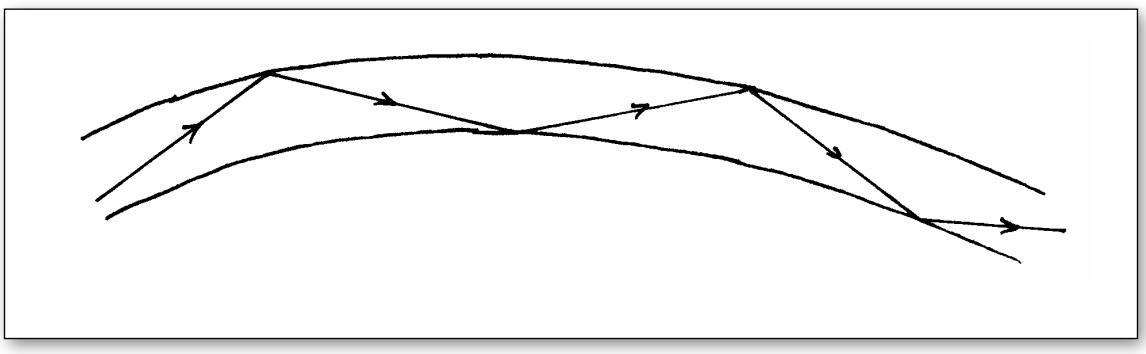


Fig. 11.3
Wave guide. If the diameter is large against the wavelength of the light, the propagation process can be considered as a sequence of total reflections.

To ensure that the light does not leave the fiber, its angle against the normal to the surface must not be less than the critical angle of total reflection, defined by the relationship

$$\sin \alpha = n$$

(cf. section 4.2) .

The diameters d of technical fibers are often only a few µm and are no longer large compared to the wavelength. Therefore, the propagation process is essentially the same as in waveguide technology: an electromagnetic wave is guided by a tube. This means that the electromagnetic field at the tube walls must satisfy certain boundary conditions. One finds that the wave exists in the tube in the form of discrete modes: For each mode, the field strength distribution across the cross section of the tube has a specific shape. Figure 11.4 shows the field strength for the 0th, 1st and 2nd mode.

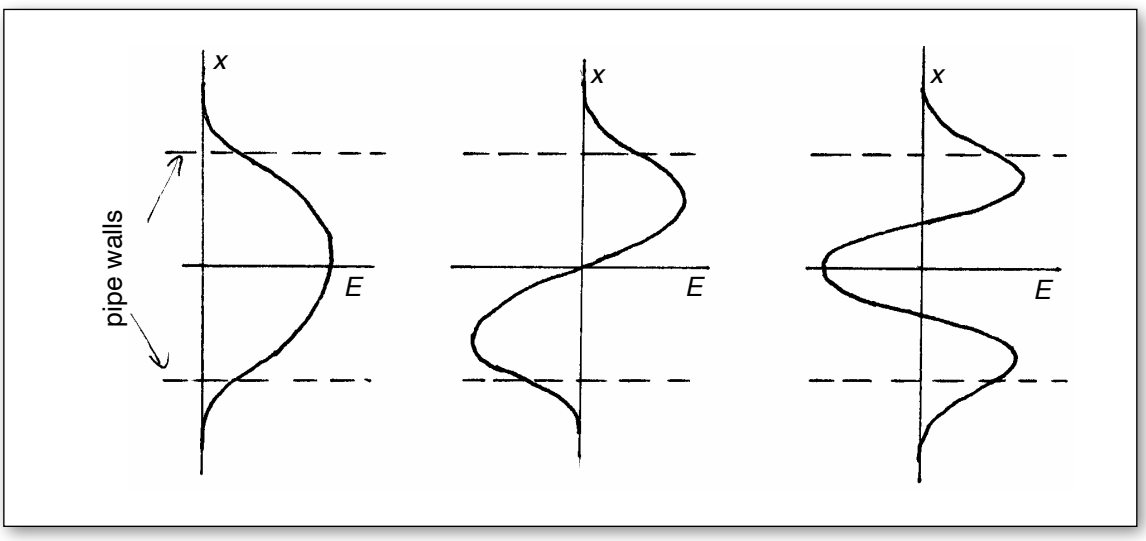


Fig. 11.4
Field strength distribution over the cross-section of the conductor for the 0, 1 and 2 modes.

The larger the ratio λ/d is, the fewer modes fit into the fiber. If d is so small that only the zero-mode goes through the fiber, it is called a *monomode fiber*, otherwise it is called a multimode fiber. Fiber optics have different applications.

If many fibers are combined into a bundle in such a way that the arrangement of the fibres over the bundle cross-section is the same everywhere, images can be transmitted. Such light guide bundles are used in medicine for endoscopy, for example to inspect the inside of the stomach wall. Two light guide bundles are needed: one for illumination and one for image transmission.

A second important application is data transmission over a single fiber. This method has advantages over data transmission using wires or free electromagnetic waves:

- because of the high frequency of light, the maximum data flow is very large (up to several Gbit/s);
- the attenuation is very low (a factor of 1.6 per km of optical fibre length, i.e. 2 dB/km)
- the transmission is neither disturbed by the weather nor by electromagnetic fields coming from outside.

The maximum data flow in a fiber optic cable is limited by dispersion: a square wave signal spreads out on its way through the cable. Therefore, two consecutive rectangular pulses, after they have travelled a long distance, are no longer detectable as two separate pulses. The most important cause of this divergence is mode dispersion. Light propagates at a different speed in the direction of the conductor depending on the mode. To eliminate this type of dispersion, monomode fibers are used for data transmission. The low attenuation is achieved by using very pure quartz as material.

11.4 Holography

On a photo of a landscape we recognize the landscape. But the photo is only a poor substitute for a window of the same size through which we look at the real landscape, Fig. 11.5, because the light field in a plane just above the photo is very different from the light field in a plane just above the window.

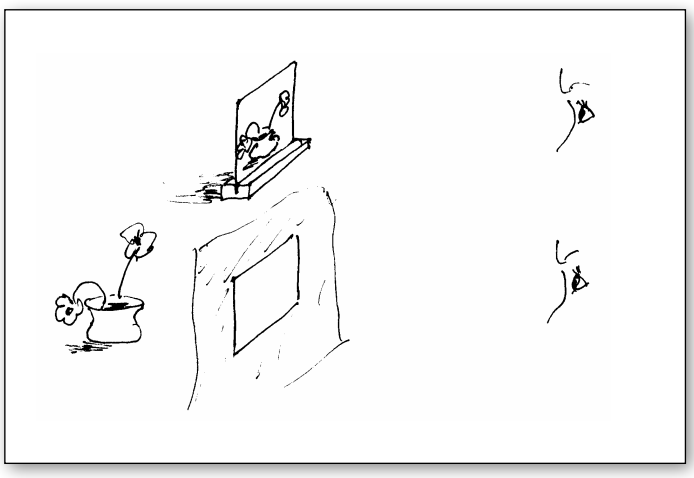


Fig. 11.5
A photo is only a poor substitute for what one would see through a rectangular window.

A hologram is a photographic image which, when illuminated with coherent light, reconstructs the field of light that has been created by the original landscape, not only close above the hologram, but in a large spatial area on one side of the hologram.

How are holograms created? How does the reproduction work?

The object from which a hologram is to be created is illuminated with coherent light. The light scattered back from the object is incident on the film. In addition, a plane wave, the so-called reference wave, is sent onto the film. The scattered light and the reference wave together create an interference pattern that is registered by the film.

If the developed film is then irradiated with a plane wave coming from the same direction as the reference wave when the hologram is recorded, the reconstruction wave, a wave field is created behind the hologram by diffraction of the reconstruction wave that is identical to the wave field that the original object would have created.

In order to understand the process, let us first consider the case where the “object” consists of a single, very distant, point. A wave emanates from the object, which is a plane wave at the location of the film. Interference with the reference wave produces a stripe-shaped interference pattern. It can be arranged in such a way that the *blackening amplitude* of the film is proportional to the amplitude of the object wave. The blackening distribution perpendicular to the stripes is then sine-shaped in the case we are looking at.

If now the reconstruction wave is sent onto the hologram, two diffracted waves are created. One of them is identical with the original wave coming from the object, the other one is symmetrical to it (with respect to the zeroth diffraction order), fig. 11.6.

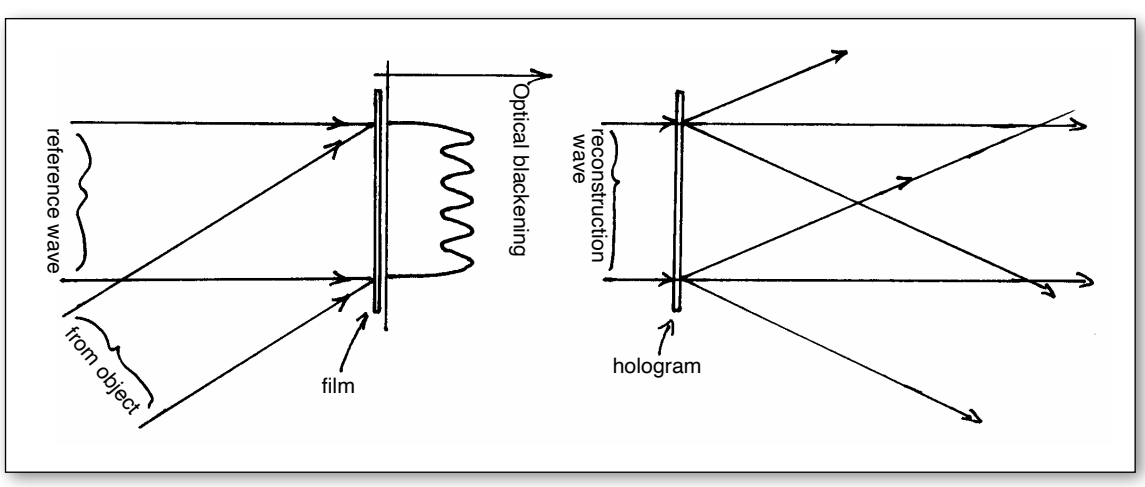


Fig. 11.6
Left: Recording the hologram of a very distant point with object and reference wave. Right: Reproduction using the reconstruction wave

If the point of the object is not located at a great distance, the wave emanating from it is a spherical wave, and the hologram is a system of rings. The diffraction of the reconstruction wave provides firstly the same spherical wave as the object would have provided and secondly a spherical wave converging on a point. The arrangement of the reference wave and the object is chosen so that the wave fields of the divergent and convergent spherical wave do not interfere with each other during reproduction, Fig. 11.7.

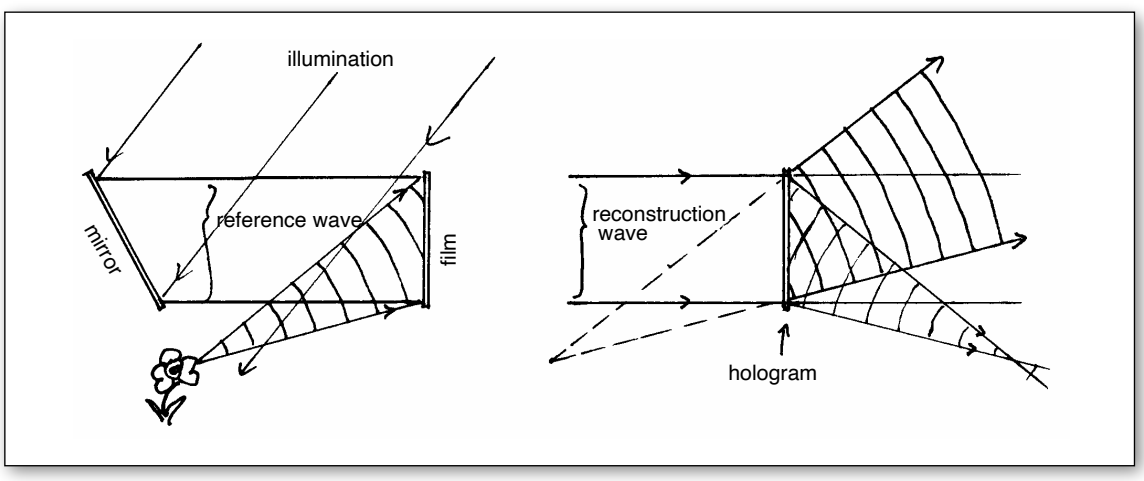


Fig. 11.7
Recording (left) and reproduction (right) of the hologram of a close object

The possibility of reproducing the original wave field with a hologram is due to the fact that in the hologram not only the amplitude but also the direction of the wave is stored at any point of the hologram plane. The value of the amplitude and the direction are encoded in the hologram in different ways: the amplitude in the amplitude of the spatial changes of the blackening and the direction in the distance and orientation of the interference fringes.

11.5 Tomography

The imaging methods discussed so far are based on the fact that every light beam in the object space has a well-defined beginning. From this starting point of the light beam an image point is generated. In many cases, however, the conditions are more complicated. The structure of the illuminated object to be imaged manifests itself in the fact that the radiation penetrates into the object and is absorbed there gradually. An example is the human body “illuminated” by X-rays. In the past, to learn about the inside of the body, one simply made a single projection. Here, the absorption along the entire path of an X-ray contributed to one pixel. It was difficult to distinguish between the various organs that are crossed by the x-rays on the picture.

The so-called computer tomography does not have this disadvantage. With this method it is possible to generate an image of any cross-section through the body. An X-ray source generates a fine beam. The receiver is located at a fixed distance from the source on the beam axis. The source-receiver pair is now moved through the cross-sectional area to be recorded perpendicular to the beam direction, Fig. 11.8. The receiver records an *absorption profile* in the process. This process is then repeated for many other orientations in the same cross-sectional area. Successive recording directions differ by a few degrees. From all profiles together, the local distribution of the absorption coefficient in the whole sectional area can be calculated.

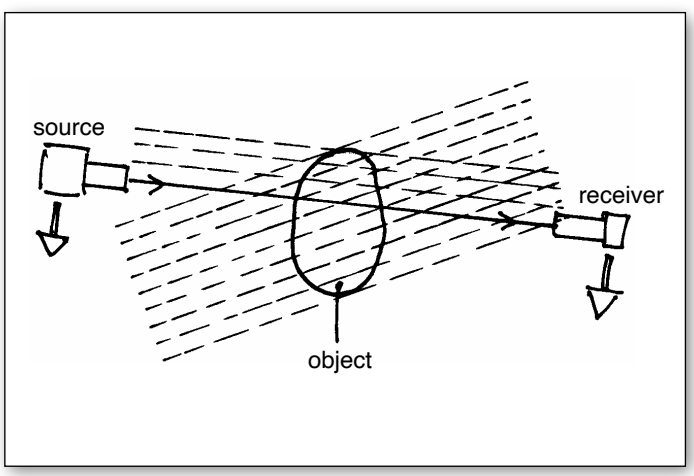


Fig. 11.8
To take a tomogram, the beam is moved through the object to be examined at a right angle to its own direction. This process is repeated for different orientations of the source-receiver arrangement.