2 Fundamental Principles of Holography

2.1 Light Waves

Light can be described as an electromagnetic wave or as a current of particles called photons. The model to be referred to depends on the experiment under investigation. Both models contradict each other, but are necessary to describe the full spectrum of light phenomena. Interaction of light with the atomic structure of matter is described by quantum optics, the theory dealing with photons. Refraction, diffraction and interference are perfectly described by the wave model, which is based on the theory of classical electromagnetism.

Interference and diffraction form the basis of holography. The appropriate theory is therefore the wave model. The oscillating quantities are the electric and the magnetic fields. The field amplitudes oscillate perpendicularly to the propagation direction of light and perpendicularly to each other, i.e. light waves are transverse phenomena. Light waves can be described either by the electrical or by the magnetic field.

Light propagation is described by the wave equation, which follows from Maxwell equations. The wave equation in vacuum is

$$\nabla^2 \vec{E} - \frac{1}{c^2} \frac{\partial^2 \vec{E}}{\partial t^2} = 0 \tag{2.1}$$

Here $\vec{E}$ is the electric field and ∇^2 is the *Laplace operator* defined as

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \tag{2.2}$$

c is the speed of light in vacuum:

$$c = 299792458\, m/s \tag{2.3}$$

The electrical field $\vec{E}$ is a vector quantity, which means it could vibrate in any direction, which is perpendicular to the light propagation. However, in many applications the wave vibrates only in a single plane. Such light is called *linear polarized light*. In this case it is sufficient to consider the scalar wave equation

$$\frac{\partial^2 E}{\partial z^2} - \frac{1}{c^2} \frac{\partial^2 E}{\partial t^2} = 0 \tag{2.4}$$

where the light propagates in z-direction.

It could be easily verified that a linearly polarized, harmonic plane wave described by

$$E(x,y,z,t) = a\cos\left(\omega t - \vec{k}\vec{r} - \varphi_0\right) \tag{2.5}$$

is a solution of the wave equation.

E(x,y,z,t) is the modulus of the electrical field vector at the point with spatial vector $\vec{r} = (x,y,z)$ at the time *t*. The quantity *a* is named *amplitude*. The *wave vector* $\vec{k}$ describes the propagation direction of the wave:

$$\vec{k} = k\vec{n} \tag{2.6}$$

$\vec{n}$ is a unit vector in propagation direction. Points of equal phase are located on parallel planes that are perpendicular to the propagation direction. The modulus of $\vec{k}$ named *wave number* is calculated by

$$\left|\vec{k}\right| \equiv k = \frac{2\pi}{\lambda} \tag{2.7}$$

The angular frequency ω corresponds to the frequency *f* of the light wave by

$$\omega = 2\pi f \tag{2.8}$$

Frequency *f* and wavelength λ are related by the speed of light *c*:

$$c = \lambda f \tag{2.9}$$

The spatially varying term

$$\varphi = -\vec{k}\vec{r} - \varphi_0 \tag{2.10}$$

is named *phase*, with phase constant φ_0. It has to be pointed out that this definition is not standardized. Some authors designate the entire argument of the cosine function, $\omega t - \vec{k}\vec{r} - \varphi_0$, as phase. The definition Eq. (2.10) is favourable to describe the holographic process and therefore used in this book.

The vacuum wavelengths of visible light are in the range of *400 nm* (violet) to *780 nm* (deep red). The corresponding frequency range is $7.5 \cdot 10^{14}\,Hz$ to $3.8 \cdot 10^{14}\,Hz$. Light sensors as e. g. the human eye, photodiodes, photographic films or CCD's are not able to detect such high frequencies due to technical and physical reasons. The only directly measurable quantity is the *intensity*. It is proportional to the time average of the square of the electrical field:

$$I = \varepsilon_0 c\left\langle E^2\right\rangle_t = \varepsilon_0 c \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} E^2 dt \tag{2.11}$$

$\langle\ \rangle_t$ stands for the time average over many light periods. The constant factor $\varepsilon_0 c$ results if the intensity is formally derived Maxwell equations. The constant ε_0 is the vacuum permittivity.

For a plane wave Eq. (2.5) has to be inserted:

$$I = \varepsilon_0 c a^2 \left\langle \cos^2\left(\omega t - \vec{k}\vec{r} - \varphi_0\right)\right\rangle_t = \frac{1}{2}\varepsilon_0 c a^2 \tag{2.12}$$

According to Eq. (2.12) the intensity is calculated by the square of the amplitude.

The expression (2.5) can be written in complex form as

$$E(x,y,z,t) = a\,\mathrm{Re}\left\{\exp\left(i\left(\omega t - \vec{k}\vec{r} - \varphi_0\right)\right)\right\} \tag{2.13}$$

where 'Re' denotes the real part of the complex function. For computations this 'Re' can be omitted. However, only the real part represents the physical wave:

$$E(x,y,z,t) = a\exp\left(i\left(\omega t - \vec{k}\vec{r} - \varphi_0\right)\right) \tag{2.14}$$

One advantage of the complex representation is that the spatial and temporal parts factorize:

$$E(x,y,z,t) = a\exp(i\varphi)\exp(i\omega t) \tag{2.15}$$

In many calculations of optics only the spatial distribution of the wave is of interest. In this case only the spatial part of the electrical field, named *complex amplitude*, has to be considered:

$$A(x,y,z) = a\exp(i\varphi) \tag{2.16}$$

The equations (2.15) and (2.16) are not just valid for plane waves, but in general for three-dimensional waves whose amplitude a and phase φ may be functions of x, y and z.

In complex notation the intensity is now simply calculated by taking the square of the modulus of the complex amplitude

$$I = \frac{1}{2}\varepsilon_0 c|A|^2 = \frac{1}{2}\varepsilon_0 c A^* A = \frac{1}{2}\varepsilon_0 c a^2 \tag{2.17}$$

where * denotes the conjugate complex. In many practical calculations where the absolute value of I is not of interest the factor $1/2\varepsilon_0 c$ can be neglected, which means the intensity is simply calculated by $I = |A|^2$.

2.2 Interference

The superposition of two or more waves in space is named *interference*. If each single wave described by $\vec{E}_i(\vec{r},t)$ is a solution of the wave equation, also the superposition

$$\vec{E}(\vec{r},t) = \sum_i \vec{E}_i(\vec{r},t) \qquad i = 1,2,... \tag{2.18}$$

is a solution, too. This is because the wave equation is a linear differential equation.

In the following interference of two monochromatic waves with equal frequencies and wavelengths is considered. The waves must have the same polarization directions, i. e. the scalar formalism can be used. The complex amplitudes of the waves are

$$A_1(x,y,z) = a_1 \exp(i\varphi_1) \tag{2.19}$$

$$A_2(x,y,z) = a_2 \exp(i\varphi_2) \tag{2.20}$$

The resulting complex amplitude is then calculated by the sum of the individual amplitudes:

$$A = A_1 + A_2 \tag{2.21}$$

According to Eq. (2.17) the intensity becomes

$$\begin{aligned} I &= |A_1 + A_2|^2 = (A_1 + A_2)(A_1 + A_2)^* \\ &= a_1^{\,2} + a_2^{\,2} + 2a_1a_2\cos(\varphi_1 - \varphi_2) \\ &= I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\Delta\varphi \end{aligned} \tag{2.22}$$

I_1, I_2 being the individual intensities and

$$\Delta\varphi = \varphi_1 - \varphi_2 \tag{2.23}$$

The resulting intensity is the sum of the individual intensities *plus* the interference term $2\sqrt{I_1 I_2}\cos\Delta\varphi$, which depends on the phase difference between the waves. The intensity reaches its maximum in all points to which applies

$$\Delta\varphi = 2n\pi \qquad \text{for } n=0,1,2,... \tag{2.24}$$

This is called *constructive interference*. The intensity reaches its minimum where

$$\Delta\varphi = (2n+1)\pi \qquad \text{for } n=0,1,2,... \tag{2.25}$$

This is named *destructive interference*. The integer n is the interference order. An interference pattern consists of dark and bright fringes as a result of constructive and destructive interference. The scalar theory applied here can also be used for waves with different polarization directions, if components of the electric field vector are considered.

The superposition of two plane waves which intersect under an angle θ with respect to each other result in an interference pattern with equidistant spacing, figure 2.1. The fringe spacing d is the distance from one interference maximum to the next and can be calculated by geometrical considerations. Figure 2.1 shows evidently that

$$\sin\theta_1 = \frac{\Delta l_1}{d} \quad ; \quad \sin\theta_2 = \frac{\Delta l_2}{d} \tag{2.26}$$

The quantities θ_1 and θ_2 are the angles between the propagation directions of the wavefronts and the vertical of the screen. The length Δl_2 is the path difference of wavefront W2 with respect to wavefront W1 at the position of the interference maximum P1 (W2 has to travel a longer way to P1 than W1). At the neighboring maximum P2 the conditions are exchanged: Now W1 has to travel a longer way; the path difference of W2 with respect to W1 is $-\Delta l_1$. The variation between the path differences at neighboring maxima is therefore $\Delta l_1 + \Delta l_2$. This difference is equal to one wavelength. Thus the interference condition is:

$$\Delta l_1 + \Delta l_2 = \lambda \tag{2.27}$$

Combining Eq. (2.26) with (2.27) results to:

$$d = \frac{\lambda}{\sin\theta_1 + \sin\theta_2} = \frac{\lambda}{2\sin\dfrac{\theta_1 + \theta_2}{2}\cos\dfrac{\theta_1 - \theta_2}{2}} \tag{2.28}$$

The approximation $\cos(\theta_1 - \theta_2)/2 \approx 1$ and $\theta = \theta_1 + \theta_2$ gives:

$$d = \frac{\lambda}{2\sin\dfrac{\theta}{2}} \tag{2.29}$$

Instead of the fringe spacing d the fringe pattern can also be described by the spatial frequency f, which is the reciprocal of d:

$$f = d^{-1} = \frac{2}{\lambda}\sin\frac{\theta}{2} \tag{2.30}$$

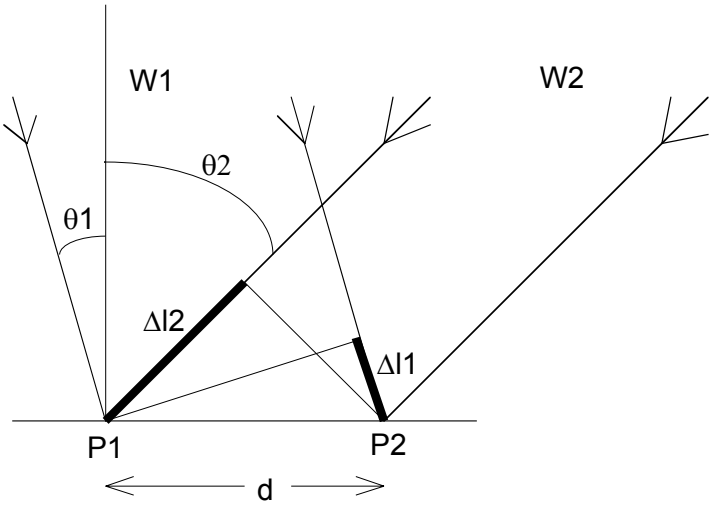


Fig. 2.1. Interference of two plane waves

2.3 Coherence

2.3.1 General

Generally the resulting intensity of two different sources, e. g. two electric light bulbs directed on a screen, is additive. Instead of dark and bright fringes as expected by Eq. (2.22) only a uniform brightness according to the sum of the individual intensities becomes visible.

In order to generate interfere fringes the phases of the individual waves must be correlated in a special way . This correlation property is named *coherence* and is investigated in this chapter. Coherence is the ability of light to interfere. The two aspects of coherence are the temporal and the spatial coherence. Temporal coherence describes the correlation of a wave with itself at different instants [71]. Spatial coherence depicts the mutual correlation of different parts of the same wavefront.

2.3.2 Temporal Coherence

The prototype of a two beam interferometer is the Michelson-interferometer, see figure 2.2. Light emitted by the light source S is split into two partial waves by the beam splitter BS. The partial waves travel to the mirrors M1 respectively M2, and are reflected back into the incident directions. After passing the beam splitter again they are superimposed at a screen. Usually the superimposed partial waves

are not exactly parallel, but are interfering with a small angle. As a result a two-dimensional interference pattern becomes visible.

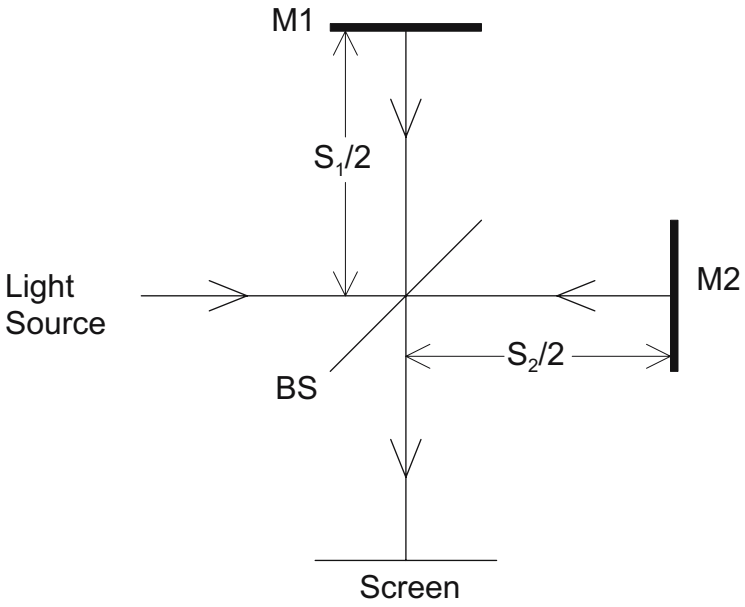


Fig. 2.2. Michelson's interferometer

The optical path length from BS to M1 and back to BS is s_1, the optical path length from BS to M2 and back to BS is s_2. Experiments prove that interference can only occur if the optical path difference $s_1 - s_2$ does not exceed a certain length L. If the optical path difference exceeds this limit, the interference fringes vanish and just an uniform brightness becomes visible on the screen. The qualitative explanation for this phenomenon is as follows: Interference fringes can only develop if the superimposed waves have a well defined (constant) phase relation. The phase difference between waves emitted by different sources vary randomly and thus the waves do not interfere. The atoms of the light source emit wave trains with finite length L. If the optical path difference exceeds this wave train length, the partial waves belonging together do not overlap after passing the different ways and interference is not possible.

The critical path length difference or, equivalently, the length of a wave train is named *coherence length L.* The corresponding emission time for the wave train

$$\tau = \frac{L}{c} \tag{2.31}$$

is called *coherence time*.

According to the laws of Fourier analysis a wave train with finite length L corresponds to light with finite spectral width Δf:

$$L = \frac{c}{\Delta f} \tag{2.32}$$

Light with long coherence length is called highly monochromatic. The coherence length is therefore a measure for the spectral width.

Typical coherence lengths of light radiated from thermal sources, e. g. ordinary electric light bulbs, are in the range of some micrometers. That means, interference can only be observed if the arms of the interferometer have nearly equal lengths. On the other hand lasers have coherence lengths from a few millimetres (e. g. a multi-mode diode laser) to several hundred meters (e. g. a stabilized single mode Nd:YAG-laser) up to several hundred kilometres for specially stabilized gas lasers used for research purposes.

The visibility

$$V = \frac{I_{\max} - I_{\min}}{I_{\max} + I_{\min}} \tag{2.33}$$

is a measure for the contrast of an interference pattern. I_{max} and I_{min} are two neighbouring intensity maxima and minima. They are calculated by inserting $\Delta\phi = 0$, respectively $\Delta\phi = \pi$ into Eq. (2.22). In the ideal case of infinite coherence length the visibility is therefore

$$V = \frac{2\sqrt{I_1 I_2}}{I_1 + I_2} \tag{2.34}$$

To consider the effect of finite coherence length the *complex self coherence* $\Gamma(\tau)$ is introduced:

$$\begin{aligned} \Gamma(\tau) &= \left\langle E(t+\tau)E^*(t)\right\rangle \\ &= \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} E(t+\tau)E^*(t)dt \end{aligned} \tag{2.35}$$

$E(t)$ is the electrical field (to be precise: the complex analytical signal) of one partial wave while $E(t+\tau)$ is the electrical field of the other partial wave. The latter is delayed in time by τ. Eq. (2.35) is the autocorrelation of E. The normalized quantity

$$\gamma(\tau) = \frac{\Gamma(\tau)}{\Gamma(0)} \tag{2.36}$$

defines the degree of coherence.

With finite coherence length the interference equation (2.22) has to be replaced by

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2}\,|\gamma|\cos\Delta\varphi \tag{2.37}$$

Maximum and minimum intensity are now calculated by

$$\begin{aligned} I_{\max} &= I_1 + I_2 + 2\sqrt{I_1 I_2}\,|\gamma| \\ I_{\min} &= I_1 + I_2 - 2\sqrt{I_1 I_2}\,|\gamma| \end{aligned} \tag{2.38}$$

Inserting these quantities into Eq. (2.33) gives

$$V = \frac{2\sqrt{I_1 I_2}}{I_1 + I_2}|\gamma| \tag{2.39}$$

For two partial waves with the same intensity, $I_1 = I_2$, Eq. (2.39) becomes

$$V = |\gamma| \tag{2.40}$$

$|\gamma|$ is equal to the visibility and therefore a measure of the ability of the two wave fields to interfere. $|\gamma| = 1$ describes ideally monochromatic light or, likewise, light with infinite coherence length. $|\gamma| = 0$ is true for completely incoherent light. Partially coherent light is described by $0 < |\gamma| < 1$.

2.3.3 Spatial Coherence

Spatial coherence describes the mutual correlation of different parts of the same wavefront. This property is measured with the Young interferometer, figure 2.3. An extended light source emits light from different points. Possible interferences are observed on a screen. An aperture with two transparent holes is mounted between light source and screen. Under certain conditions, which will be derived in this chapter, interferences are visible on the screen. The fringes result from light rays which travelled on different ways to the screen, either via the upper or via the lower hole in the aperture [164]. The interference pattern vanishes if the distance between the holes a exceeds the critical limit a_k. This limit is named *coherence distance*. The phenomenon is not related to the spectral width of the light source, but has following cause: The waves emitted by different source points of the extended light source are superimposed on the screen. It may happen, that a special source point generates an interference maximum at a certain point on the screen, while another source point generates a minimum at the same screen point. This is because the optical path length is different for light rays emerging from different source points. In general the contributions from all source points compensate themselves and the contrast vanishes. This compensation is avoided if following condition is fulfilled for every point of the light source:

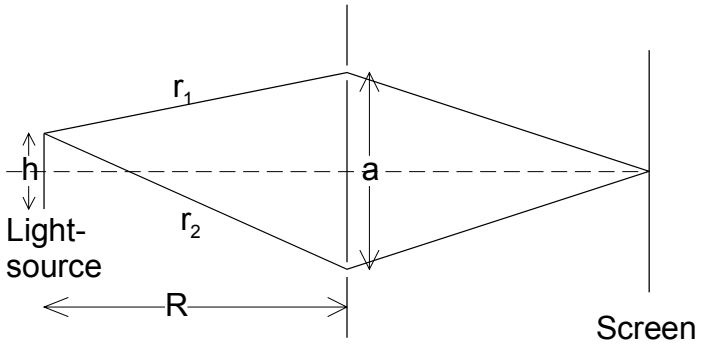


Fig. 2.3. Young's interferometer

$$r_2 - r_1 < \frac{\lambda}{2} \tag{2.41}$$

This condition is definitely fulfilled, if it is kept for the edges of the light source. The following relations are valid for points at the edges:

$$r_1^{\,2} = R^2 + \left(\frac{a-h}{2}\right)^2 \quad ; \quad r_2^{\,2} = R^2 + \left(\frac{a+h}{2}\right)^2 \tag{2.42}$$

h is the width of the light source. Using of the assumptions $a << R$ and $h << R$ results to

$$r_2 - r_1 \approx \frac{ah}{2R} \tag{2.43}$$

Combining Eq. (2.41) and (2.43) leads to following expression:

$$\frac{ah}{2R} < \frac{\lambda}{2} \tag{2.44}$$

The coherence distance is therefore:

$$\frac{a_k h}{2R} = \frac{\lambda}{2} \tag{2.45}$$

In contrast to temporal coherence the spatial coherence depends not only on properties of the light source, but also on the geometry of the interferometer. A light source may initially generate interference, which means Eq. (2.44) is fulfilled. If the distance between the holes increases or the distance between the light source and the aperture decreases, Eq. (2.44) becomes violated and the interference figure vanishes.

To consider spatial coherence the autocorrelation function defined in Eq. (2.35) is extended:

$$\Gamma(\vec{r}_1,\vec{r}_2,\tau) = \left\langle E(\vec{r}_1,t+\tau)E^*(\vec{r}_2,t)\right\rangle \tag{2.46}$$
$$= \lim_{T\to\infty}\frac{1}{2T}\int_{-T}^{T} E(\vec{r}_1,t+\tau)E^*(\vec{r}_2,t)dt$$

$\vec{r}_1$, $\vec{r}_2$ are the spatial vectors of the holes in the aperture of the Young interferometer. This function is named cross correlation function. The normalized function is

$$\gamma(\vec{r}_1,\vec{r}_2,\tau) = \frac{\Gamma(\vec{r}_1,\vec{r}_2,\tau)}{\sqrt{\Gamma(\vec{r}_1,\vec{r}_1,0)\Gamma(\vec{r}_2,\vec{r}_2,0)}} \tag{2.47}$$

where $\Gamma(\vec{r}_1,\vec{r}_1,0)$ is the intensity at $\vec{r}_1$ and $\Gamma(\vec{r}_2,\vec{r}_2,0)$ is the intensity at $\vec{r}_2$. Eq. (2.47) describes the degree of correlation between the light field at $\vec{r}_1$ at time $t+\tau$ with the light field at $\vec{r}_2$ at time *t*. The special function $\gamma(\vec{r}_1,\vec{r}_2,\tau=0)$ is a measure for the correlation between the field amplitudes at $\vec{r}_1$ and $\vec{r}_2$ at the same time and is named *complex degree of coherence*. The modulus of the normalized coherence function $|\gamma(\vec{r}_1,\vec{r}_2,\tau)|$ is measured with the Young interferometer.

2.4 Diffraction

A light wave which hits an obstacle is considered. This might be an opaque screen with some transparent holes, or vice versa, a transparent medium with opaque structures. From geometrical optics it is known that the shadow becomes visible on a screen behind the obstacle. By closer examination, one finds that this is not strictly correct. If the dimensions of the obstacle (e. g. diameter of holes in an opaque screen or size of opaque particles in a transparent volume) are in the range of the wavelength, the light distribution is not sharply bounded, but forms a pattern of dark and bright regions. This phenomenon is named *diffraction*.

Diffraction can be explained qualitatively with the *Huygens' principle*:

Every point of a wavefront can be considered as a source point for secondary spherical waves. The wavefront at any other place is the coherent superposition of these secondary waves.

Huygens' principle is graphically explained in figure 2.4.

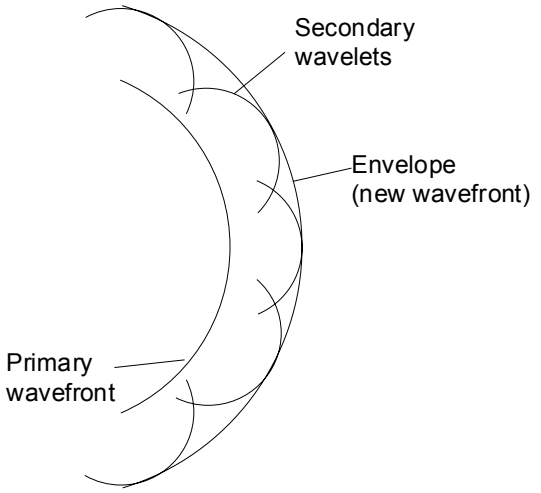


Fig. 2.4. Huygens' principle

The Fresnel-Kirchhoff integral describes diffraction quantitatively [69]:

$$\Gamma(\xi',\eta') = \frac{i}{\lambda}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} A(x,y)\frac{\exp\left(-i\frac{2\pi}{\lambda}\rho'\right)}{\rho'}Q\,dx\,dy \tag{2.48}$$

with

$$\rho' = \sqrt{(x-\xi')^2 + (y-\eta')^2 + d^2} \tag{2.49}$$

and

$$Q = \frac{1}{2}(\cos\theta + \cos\theta') \tag{2.50}$$

$A(x,y)$ is the complex amplitude in the plane of the bending aperture, see coordinate system defined in figure 2.5. $\Gamma(\xi',\eta')$ is the field in the observation plane. ρ' stands for the distance between a point in the aperture plane and a point in the observation plane. Eq. (2.48) can be understood as the mathematical formulation of Huygens' principle:

The light source S lying in the source plane with coordinates (ξ,η) radiates spherical waves. $A(x,y)$ is the complex amplitude of such a wave in the aperture plane. At first an opaque aperture with only one hole at the position (x,y) is considered. Such a hole is now the source for secondary waves. The field at the position (ξ',η') of the diffraction plane is proportional to the field at the entrance side of the aperture $A(x,y)$ and to the field of the secondary spherical wave emerging

from (*x,y*), described by $\exp(-i2\pi/\lambda\rho')/\rho'$. Now the entire aperture as a plane consisting of many sources for secondary waves is considered. The entire resulting field in the diffraction plane is therefore the integral over all secondary spherical waves, emerging from the aperture plane.

The Huygens' principle would allow that the secondary waves not only propagate in the forward direction, but also back into the direction of the source. Yet, the experiment demonstrates that the wavefronts always propagate in one direction. To exclude this unphysical situation formally the inclination factor *Q* defined in Eq. (2.50) is introduced in the Fresnel-Kirchhoff integral. *Q* depends on the angle θ between the incident ray from the source and the unit vector $\vec{n}$ perpendicular to the aperture plane, and on the angle θ' between the bended ray and $\vec{n}$, see figure 2.6. *Q* is approximately zero for $\theta \approx 0$ and $\theta' \approx \pi$. This prevents waves travelling into the backward direction. In most practical situations both θ and θ' are very small and $Q \approx 1$. The inclination factor can be considered as an ad hoc correction to the diffraction integral, as done here, or be derived in the formal diffraction theory [69, 41].

Other authors use a "+" sign in the argument of the exponential function of the Fresnel-Kirchhoff integral ($\Gamma(\xi,\eta) = ...A(x,y)\exp(+i2\pi/\lambda\rho')/\rho'...$) instead of the "-" sign used here. This depends on the definition of a harmonic wave in Eq. (2.14), which can be defined either as $\exp(+i\varphi)$ or $\exp(-i\varphi)$. However, using the "+" sign in Eq. (2.48) leads to the same expressions for all measurable quantities, as e.g. the intensity and the magnitude of the interference phase used in Digital Holographic Interferometry.

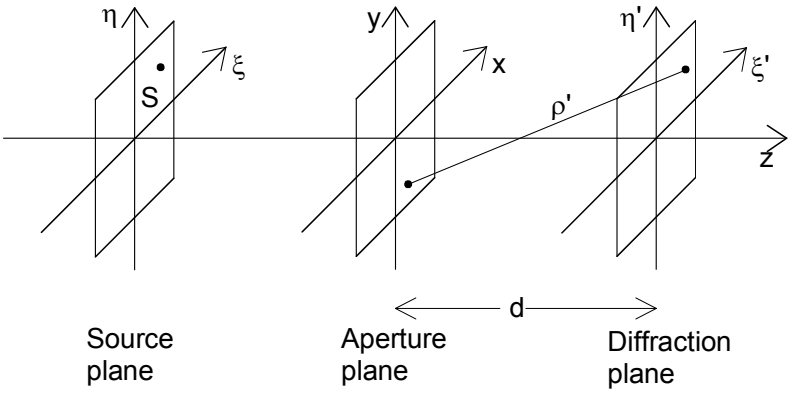


Fig. 2.5. Coordinate system

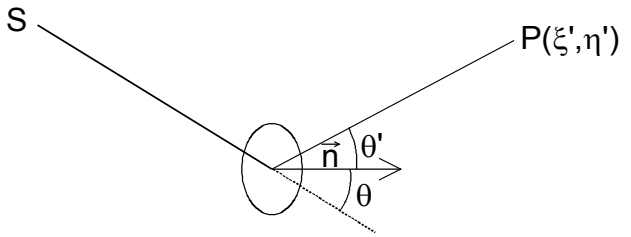


Fig. 2.6. Angles

2.5 Speckles

A rough surface illuminated with coherent light appears grainy for an observer. The intensity of the light scattered by the surface fluctuates randomly in space, dark and bright spots appear. These spots are named *speckles*, forming the entire image called speckle pattern, figure 2.7. A speckle pattern develops if the height variations of the rough surface are larger than the wavelength of the light.

Speckles result from interference of light scattered by the surface points. The phases of the waves scattered by different surface points fluctuate statistically due to the height variations. If these waves interfere with each other, a stationary speckle pattern develops.

It can be shown that the probability density function for the intensity in a speckle pattern obeys negative exponential statistics [40]:

$$P(I)dI = \frac{1}{\langle I \rangle} \exp\left(-\frac{I}{\langle I \rangle} \right) \tag{2.51}$$

$P(I)dI$ is the probability that the intensity at a certain point is lying between I and $I + dI$. $\langle I \rangle$ is the mean intensity of the entire speckle field. The most probable intensity value of a speckle is therefore zero, i. e. most speckles are black. The standard deviation σ_I is calculated by

$$\sigma_I = \langle I \rangle \tag{2.52}$$

That means the intensity variations are in the same order as the mean value. A usual definition of the contrast is

$$V = \frac{\sigma_I}{\langle I \rangle} \tag{2.53}$$

The contrast of a speckle pattern is therefore always unity.

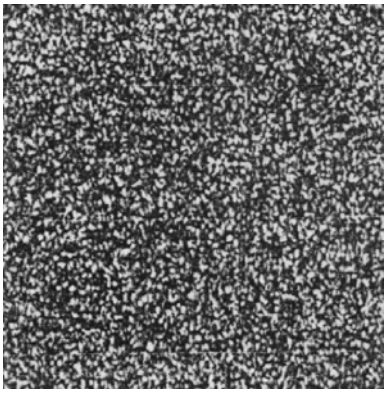


Fig. 2.7. Speckle pattern

One can distinguish between objective and subjective speckle formation. An objective speckle pattern develops on a screen, located in a distance z from the illuminated surface, figure 2.8. There is no imaging system between surface and screen. The size of a speckle in an objective speckle pattern can be estimated using the spatial frequency formula Eq. (2.30). The two edge points of the illuminated surface form the highest spatial frequency:

$$f_{\max} = \frac{2}{\lambda} \sin \frac{\theta_{\max}}{2} \approx \frac{L}{\lambda z} \tag{2.54}$$

The reciprocal of f_{max} is a measure for the speckle size:

$$d_{Sp} = \frac{\lambda z}{L} \tag{2.55}$$

A subjective speckle pattern develops if the illuminated surface is focussed with an imaging system, e. g. a camera lens or the human eye, figure 2.9. In this case the speckle diameter depends on the aperture diameter a of the imaging system. The size of a speckle in a subjective speckle pattern can be estimated again using the spatial frequency:

$$f_{\max} = \frac{2}{\lambda} \sin\left(\frac{\theta_{\max}}{2} \right) \approx \frac{a}{\lambda b} \tag{2.56}$$

b is the image distance of the imaging system. It follows for the speckle size:

$$d_{Sp} = \frac{\lambda b}{a} \tag{2.57}$$

The speckle size can be increased by closing the aperture of the imaging system.

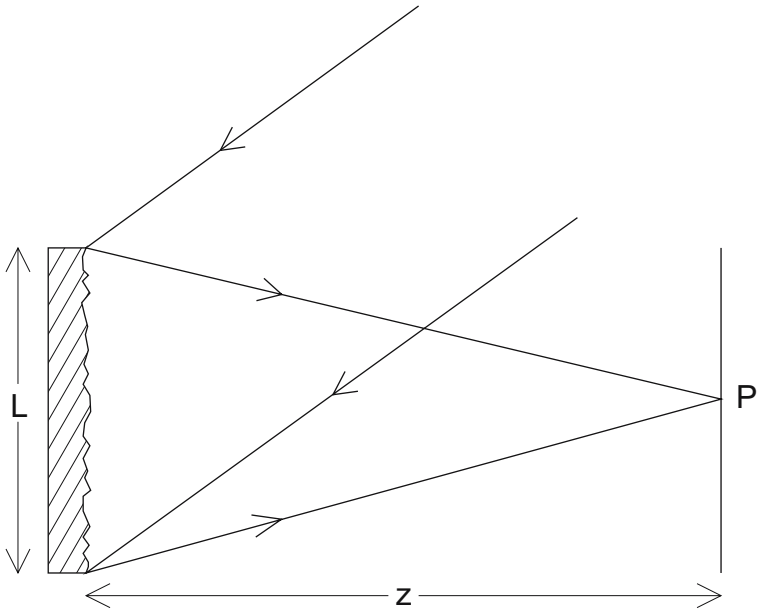


Fig. 2.8. Objective speckle formation

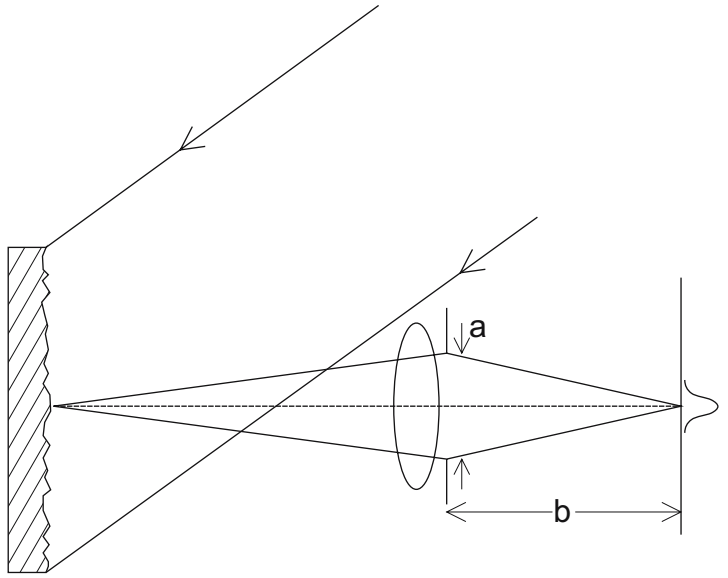


Fig. 2.9. Subjective speckle formation

2.6 Holography

2.6.1 Hologram Recording and Reconstruction

Holograms are usually recorded with an optical set-up consisting of a light source (laser), mirrors and lenses for beam guiding and a recording device, e. g. a photographic plate. A typical set-up is shown in figure 2.10 [47, 71]. Light with sufficient coherence length is split into two partial waves by a beam splitter (BS). The first wave illuminates the object. It is scattered at the object surface and reflected to the recording medium. The second wave - named reference wave - illuminates the light sensitive medium directly. Both waves interfere. The interference pattern is recorded, e.g. by chemical development of the photographic plate. The recorded interference pattern is named hologram.

The original object wave is reconstructed by illuminating the hologram with the reference wave, figure 2.11. An observer sees a virtual image, which is indistinguishable from the original object. The reconstructed image exhibits all effects of perspective and depth of focus.

The holographic process is described mathematically using the formalism of chapter 2.2. The complex amplitude of the object wave is described by

$$E_O(x,y) = a_O(x,y)\exp(i\varphi_O(x,y)) \tag{2.58}$$

with real amplitude a_O and phase φ_O.

$$E_R(x,y) = a_R(x,y)\exp(i\varphi_R(x,y)) \tag{2.59}$$

is the complex amplitude of the reference wave with real amplitude a_R and phase φ_R.

Both waves interfere at the surface of the recording medium. The intensity is calculated by

$$\begin{aligned} I(x,y) &= |E_O(x,y) + E_R(x,y)|^2 \\ &= (E_O(x,y) + E_R(x,y))(E_O(x,y) + E_R(x,y))^* \\ &= E_R(x,y)E_R{}^*(x,y) + E_O(x,y)E_O{}^*(x,y) + E_O(x,y)E_R{}^*(x,y) + E_R(x,y)E_O{}^*(x,y) \end{aligned} \tag{2.60}$$

The amplitude transmission $h(x,y)$ of the developed photographic plate (or of other recording media) is proportional to *I(x,y)*:

$$h(x,y) = h_0 + \beta\tau I(x,y) \tag{2.61}$$

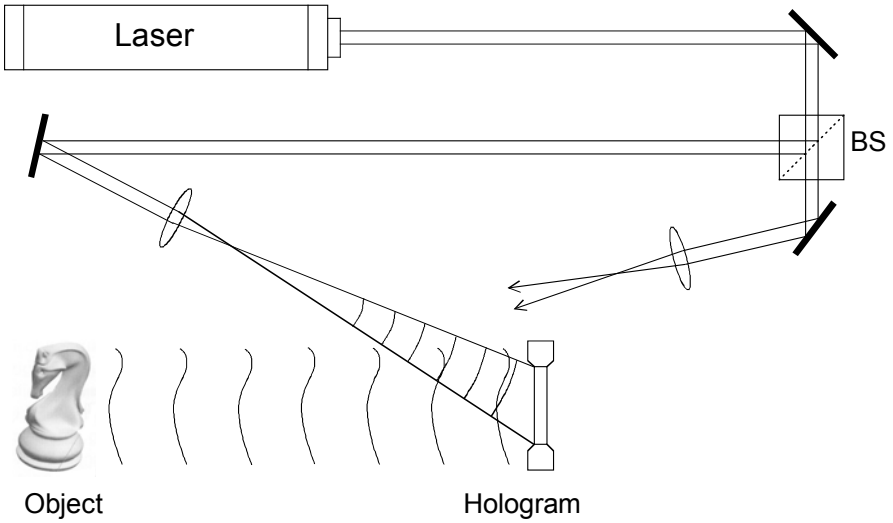


Fig. 2.10. Hologram recording

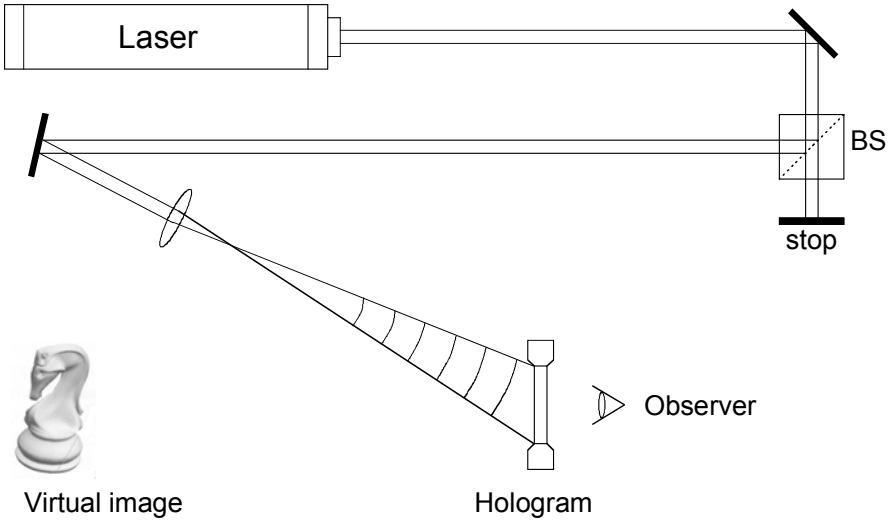


Fig. 2.11. Hologram reconstruction

The constant β is the slope of the amplitude transmittance versus exposure characteristic of the light sensitive material. For photographic emulsions β is negative. τ is the exposure time and h_0 is the amplitude transmission of the unexposed

plate. *h(x,y)* is named hologram function. In Digital Holography using CCD's as recording medium h_0 can be neglected.

For hologram reconstruction the amplitude transmission has to be multiplied with the complex amplitude of the reconstruction (reference) wave:

$$E_R(x,y)h(x,y) = \left[h_0 + \beta\tau\left(a_R{}^2 + a_O{}^2\right)\right]E_R(x,y) + \beta\tau a_R{}^2 E_O(x,y) + \beta\tau E_R{}^2(x,y)E_O{}^*(x,y) \tag{2.62}$$

The first term on the right side of this equation is the reference wave, multiplied by a factor. It represents the undiffracted wave passing the hologram (zero diffraction order). The second term is the reconstructed object wave, forming the virtual image. The real factor $\beta\tau a_R{}^2$ only influences the brightness of the image. The third term generates a distorted real image of the object. For off-axis holography the virtual image, the real image and the undiffracted wave are spatially separated.

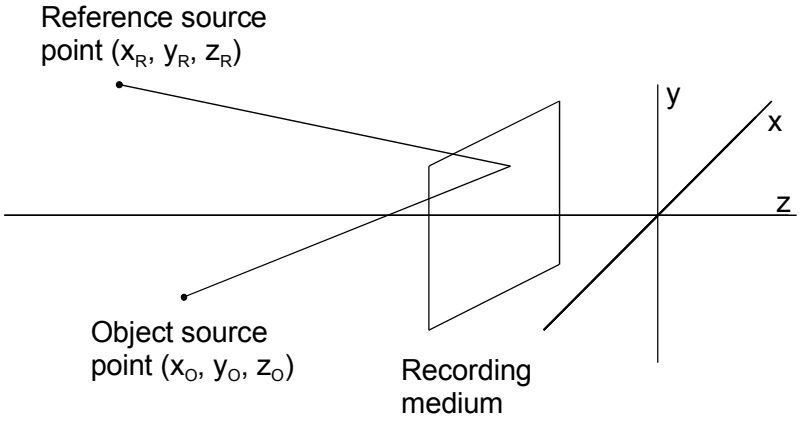
The reason for the distortion of the real image is the spatially varying complex factor $E_R{}^2$, which modulates the image forming conjugate object wave $E_O{}^*$. An undistorted real image can be generated by using the conjugate reference beam $E_R{}^*$ for reconstruction:

$$E_R{}^*(x,y)h(x,y) = \left[h_0 + \beta\tau\left(a_R{}^2 + a_O{}^2\right)\right]E_R{}^*(x,y) + \beta\tau a_R{}^2 E_O{}^*(x,y) + \beta\tau E_R{}^{*2}(x,y)E_O(x,y) \tag{2.63}$$

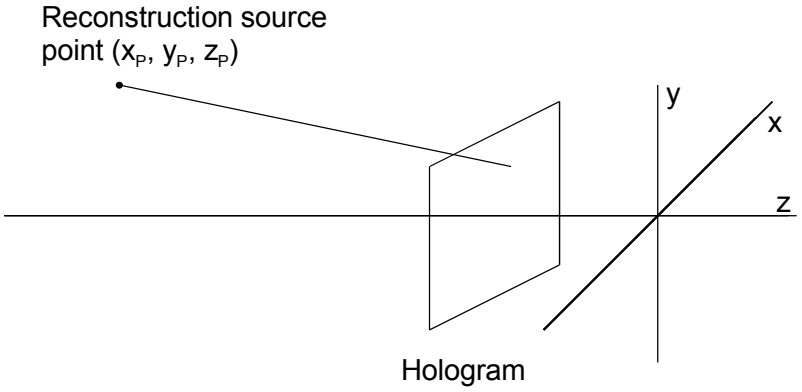
2.6.2 The Imaging Equations

The virtual image appears at the position of the original object if the hologram is reconstructed with the same parameters like those used in the recording process. However, if one changes the wavelength or the coordinates of the reconstruction wave source point with respect to the coordinates of the reference wave source point used in the recording process, the position of the reconstructed image moves. The coordinate shift is different for all points, thus the shape of the reconstructed object is distorted. The image magnification can be influenced by the reconstruction parameters, too.

The *imaging equations* relate the coordinates of an object point O with that of the corresponding point in the reconstructed image. These equations are quoted here without derivation, because they are needed to explain specific techniques such as Digital Holographic Microscopy. An exact derivation can be found in [47, 71].



(a) Hologram recording



(b) Image reconstruction

Fig. 2.12. Coordinate system used to describe holographic reconstruction

The coordinate system is shown in figure 2.12. (x_O, y_O, z_O) are the coordinates of the object point O, (x_R, y_R, z_R) are the coordinates of the source point of the reference wave used for hologram recording and (x_P, y_P, z_P) are the coordinates of the source point of the reconstruction wave. $\mu = \lambda_2 / \lambda_1$ denotes the ratio be-

tween the recording wavelength λ_1 and the reconstruction wavelength λ_2. The coordinates of that point in the reconstructed virtual image, which corresponds to the object point O, are:

$$x_1 = \frac{x_P z_O z_R + \mu x_O z_P z_R - \mu x_R z_P z_O}{z_O z_R + \mu z_P z_R - \mu z_P z_O} \tag{2.64}$$

$$y_1 = \frac{y_P z_O z_R + \mu y_O z_P z_R - \mu y_R z_P z_O}{z_O z_R + \mu z_P z_R - \mu z_P z_O} \tag{2.65}$$

$$z_1 = \frac{z_P z_O z_R}{z_O z_R + \mu z_P z_R - \mu z_P z_O} \tag{2.66}$$

The coordinates of that point in the reconstructed real image, which corresponds to the object point O, are:

$$x_2 = \frac{x_P z_O z_R - \mu x_O z_P z_R + \mu x_R z_P z_O}{z_O z_R - \mu z_P z_R + \mu z_P z_O} \tag{2.67}$$

$$y_2 = \frac{y_P z_O z_R - \mu y_O z_P z_R + \mu y_R z_P z_O}{z_O z_R - \mu z_P z_R + \mu z_P z_O} \tag{2.68}$$

$$z_2 = \frac{z_P z_O z_R}{z_O z_R - \mu z_P z_R + \mu z_P z_O} \tag{2.69}$$

An extended object can be considered to be made up of a number of point objects. The coordinates of all surface points are described by the above mentioned equations. The lateral magnification of the entire virtual image is depicted:

$$M_{lat,1} = \frac{dx_1}{dx_O} = \left[1 + z_0\left(\frac{1}{\mu z_P} - \frac{1}{z_R}\right)\right]^{-1} \tag{2.70}$$

The lateral magnification of the real image results in:

$$M_{lat,2} = \frac{dx_2}{dx_O} = \left[1 - z_0\left(\frac{1}{\mu z_P} + \frac{1}{z_R}\right)\right]^{-1} \tag{2.71}$$

The longitudinal magnification of the virtual image is given by:

$$M_{long,1} = \frac{dz_1}{dz_O} = \frac{1}{\mu} M_{lat,1}{}^2 \tag{2.72}$$

The longitudinal magnification of the real image is:

$$M_{long,2} = \frac{dz_2}{dz_O} = -\frac{1}{\mu} M_{lat,2}{}^2 \tag{2.73}$$

There is a difference between real and virtual image to be mentioned: Since the real image is formed by the conjugate object wave O^*, it has the curious property that its depth is inverted. Corresponding points of the virtual image (which coincide with the original object points) and of the real image are located at equal distances from the hologram plane, but at opposite sides of it. The background and the foreground of the real image are therefore exchanged. The real image appears with the "wrong perspective". It is called a *pseudoscopic image*, in contrast to a normal or *orthoscopic* image.

2.7 Holographic Interferometry

2.7.1 Generation of Holographic Interferograms

Holographic Interferometry (HI) is a method to measure optical path length variations, which are caused by deformations of opaque bodies or refractive index variations in transparent media, e.g. fluids or gases [113]. HI is a non-contact, non-destructive method with very high sensitivity. Optical path changes up to one hundredth of a wavelength are resolvable.

Two coherent wave fields, which are reflected in two different states of the object, interfere. This is achieved e.g. in double-exposure holography by the recording of two wave fields on a single photographic plate, figure 2.13. The first exposure represents the object in its reference state, the second exposure represents the object in its loaded (e. g. deformed) state. The hologram is reconstructed by illumination with the reference wave, figure 2.14. As a result of the superposition of two holographic recordings with *slightly* different object waves only one image superimposed by interference fringes is visible, see example in figure 2.15. From this holographic interferogram the observer can determine optical path changes due to the object deformation or other effects.

In the real time technique the hologram is replaced - after processing - in exactly the recording position. When it is illuminated with the reference wave, the reconstructed virtual image coincides with the object. Interference patterns caused by phase changes between the holographically reconstructed reference object wave and the actual object wave are observable in real time.

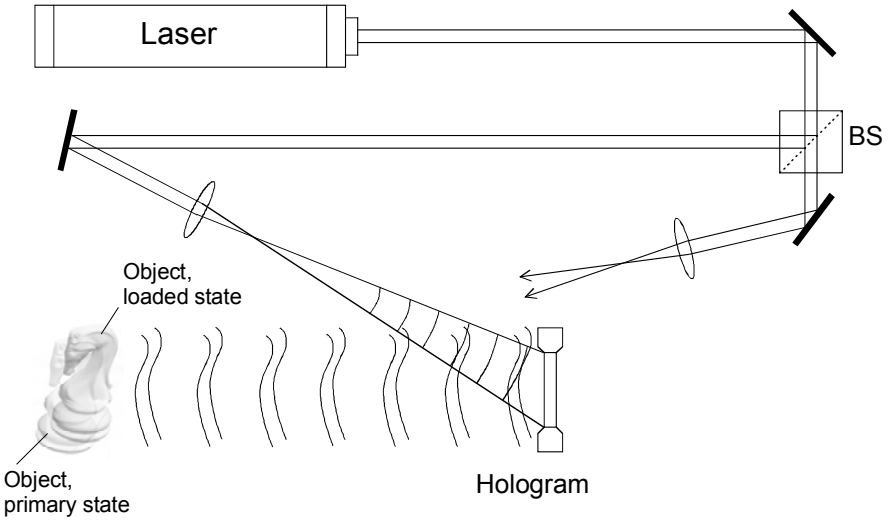


Fig. 2.13. Recording of a double exposed hologram

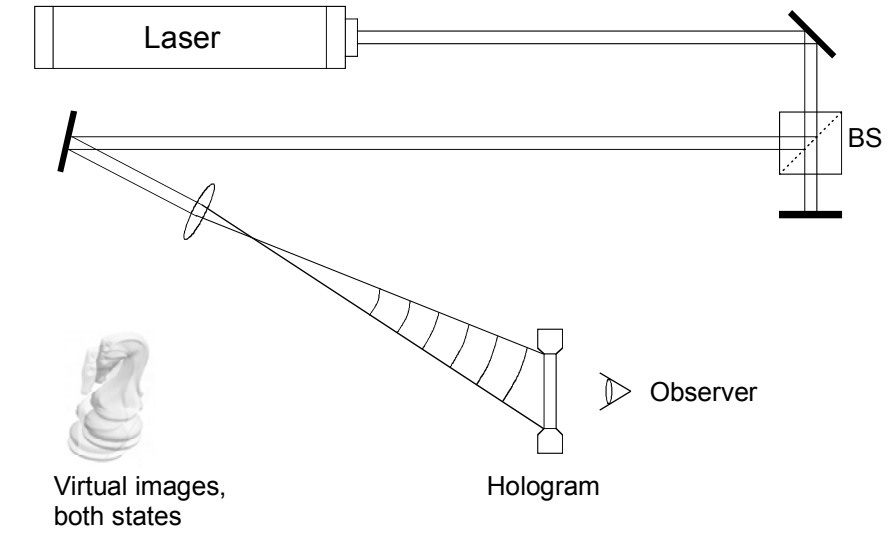


Fig. 2.14. Reconstruction

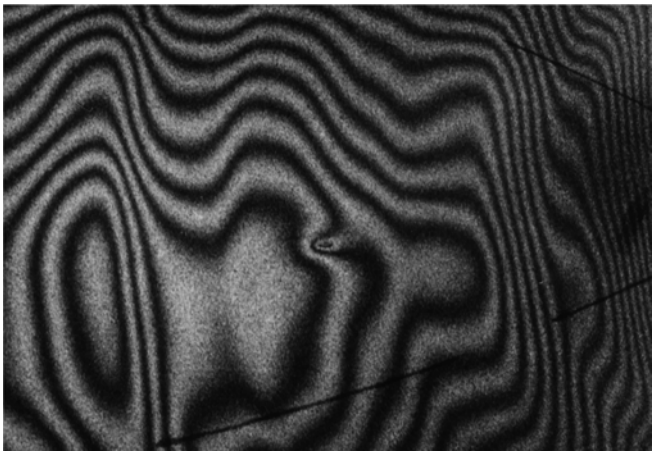


Fig. 2.15. Holographic interferogram

The following mathematical description is valid for the double exposure and for the real time technique. The complex amplitude of the object wave in the initial state is:

$$E_1(x,y) = a(x,y)\exp[i\varphi(x,y)] \tag{2.74}$$

$a(x,y)$ is the real amplitude and $\varphi(x,y)$ the phase of the object wave.

Optical path changes due to deformations of the object surface can be described by a variation of the phase from φ to $\varphi+\Delta\varphi$. $\Delta\varphi$ stands for the difference between the reference and the actual phase. It is called *interference phase*. The complex amplitude of the actual object wave is therefore denoted by

$$E_2(x,y) = a(x,y)\exp[i(\varphi(x,y) + \Delta\varphi(x,y))] \tag{2.75}$$

The intensity of a holographic interference pattern is described by the square of the sum of the complex amplitudes. It is calculated as follows:

$$\begin{aligned} I(x,y) &= |E_1 + E_2|^2 = (E_1 + E_2)(E_1 + E_2)^* \\ &= 2a^2(1 + \cos(\Delta\varphi)) \end{aligned} \tag{2.76}$$

The general expression for the intensity within an interference pattern is therefore:

$$I(x,y) = A(x,y) + B(x,y)\cos\Delta\varphi(x,y) \tag{2.77}$$

The parameters $A(x,y)$ and $B(x,y)$ depend on the coordinates in the interferogram.

In practice these parameters are not known due to several disturbing effects:

- The object is illuminated by an expanded laser beam having a gaussian profile. The brightness of the holographic interferogram varies accordingly.
- The interferogram is superimposed by a high frequency speckle noise.
- Dust particles in the optical path result in diffraction patterns.
- The surface of the object under investigation may have a varying reflectivity influencing the brightness and visibility of the interferogram.
- Electronic recording and transmission of holographic interferograms generates additional noise.

Eq. (2.77) describes the relation between the intensity of the interference pattern and the interference phase, which contains the information about the physical quantity to be measured (object displacement, refractive index change or object shape). In general it is not possible to calculate $\Delta\varphi$ directly from the measured intensity, because the parameters *A(x,y)* and *B(x,y)* are not known. In addition the cosine is an even function (cos30°=cos-30°) and the sign of $\Delta\varphi$ is not determined unambiguously. Therefore several techniques have been developed to determine the interference phase by recording additional information. The most common techniques are the various phase shifting methods, which are briefly discussed in chapter 2.7.5.

2.7.2 Displacement Measurement by HI

In this chapter a relation between the measured interference phase and the displacement of the object surface under investigation is derived [71, 147]. The geometric quantities are explained in figure 2.16. The vector $\vec{d}(x,y,z)$ is called displacement vector. It describes the shift of a surface point from its initial position P_1 to the new position P_2 due to deformation. $\vec{s}_1$ and $\vec{s}_2$ are unit vectors from the illumination source point S to P_1, resp. P_2. $\vec{b}_1$ and $\vec{b}_2$ are unit vectors from P_1 to the observation point B, resp. from P_2 to B. The optical path difference between a ray from S to B via P_1 and a ray from S to B via P_2 is:

$$\begin{aligned}\delta &= \overline{SP_1} + \overline{P_1B} - \left(\overline{SP_2} + \overline{P_2B}\right) \\ &= \vec{s}_1\overrightarrow{SP_1} + \vec{b}_1\overrightarrow{P_1B} - \vec{s}_2\overrightarrow{SP_2} - \vec{b}_2\overrightarrow{P_2B}\end{aligned} \tag{2.78}$$

The lengths $\overline{SP_{1/2}}$ and $\overline{P_{1/2}B}$ are in the range of meter, while $\left|\vec{d}\right|$ is in the range of several micrometers. The vectors $\vec{s}_1$ and $\vec{s}_2$ can therefore be replaced by a unit vector $\vec{s}$ pointing into the bisector of the angle spread by $\vec{s}_1$ and $\vec{s}_2$:

$$\vec{s}_1 = \vec{s}_2 = \vec{s} \tag{2.79}$$

$\vec{b}_1$ and $\vec{b}_2$ are accordingly replaced by a unit vector $\vec{b}$ pointing into the bisector of the angle spread by $\vec{b}_1$ and $\vec{b}_2$:

$$\vec{b}_1 = \vec{b}_2 = \vec{b} \tag{2.80}$$

The displacement vector $\vec{d}(x, y, z)$ is given by:

$$\vec{d} = \overrightarrow{P_1 B} - \overrightarrow{P_2 B} \tag{2.81}$$

and

$$\vec{d} = \overrightarrow{SP_2} - \overrightarrow{SP_1} \tag{2.82}$$

Inserting Eq. (2.79) to (2.82) into (2.78) gives:

$$\delta = (\vec{b} - \vec{s})\vec{d} \tag{2.83}$$

The following expression results for the interference phase:

$$\Delta\varphi(x, y) = \frac{2\pi}{\lambda}\vec{d}(x, y, z)(\vec{b} - \vec{s}) = \vec{d}(x, y, z)\vec{S} \tag{2.84}$$

The vector

$$\vec{S} = \frac{2\pi}{\lambda}(\vec{b} - \vec{s}) \tag{2.85}$$

is called *sensitivity vector*. The sensitivity vector is only defined by the geometry of the holographic arrangement. It gives the direction in which the set-up has maximum sensitivity. At each point the projection of the displacement vector onto the sensitivity vector is measured. Eq. (2.84) is the basis of all quantitative measurements of the deformation of opaque bodies.

In the general case of a three dimensional deformation field Eq. (2.84) contains the three components of $\vec{d}$ as unknown parameters. Three interferograms of the same surface with linear independent sensitivity vectors are necessary to determine the displacement. In many practical cases not the three dimensional displacement field is of interest, but the deformation perpendicular to the surface. This *out-of-plane* deformation can be measured using an optimised set-up with parallel illumination and observation directions ($\vec{S} = 2\pi/\lambda(0,0,2)$). The component d_z is then calculated from the interference phase by

$$d_z = \Delta\varphi\frac{\lambda}{4\pi} \tag{2.86}$$

A phase variation of 2π corresponds to a deformation of $\lambda/2$.

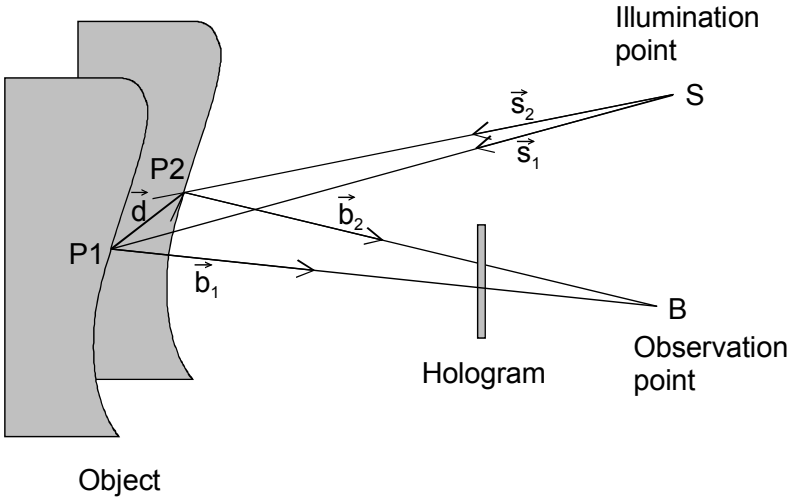


Fig. 2.16. Calculation of the interference phase

2.7.3 Holographic Contouring

Another application of HI is the generation of a fringe pattern corresponding to contours of constant elevation with respect to a reference plane. Such contour fringes can be used to determine the shape of a three-dimensional object.

Holographic contour interferograms can be generated by different methods. In the following the

- *two-wavelength method* and the
- *two-illumination-point method*

are described. A third method, *the two-refractive-index technique*, has less practical applications and is not considered here.

The principal set-up of the two-wavelength method is shown in figure 2.17. A plane wave illuminates the object surface . The back scattered light interferes with the plane reference wave at the holographic recording medium. In the set-up of figure 2.17 the illumination wave is reflected onto the object surface via a beam splitter in order to ensure parallel illumination and observation directions. Two holograms are recorded with different wavelengths λ_1 and λ_2 on the same photographic plate. This can be done either simultaneously using two lasers with different wavelengths or in succession changing the wavelength of a tuneable laser, e. g. a dye laser. After processing the double exposed hologram is replaced and reconstructed with only one of the two wavelengths, say λ_2. Two virtual images become visible. The image recorded with λ_2 coincides with the object surface. The other image, recorded with λ_1 but reconstructed with λ_2, is slightly distorted. The z-coordinate of this image z' is calculated with the imaging equation (2.66):

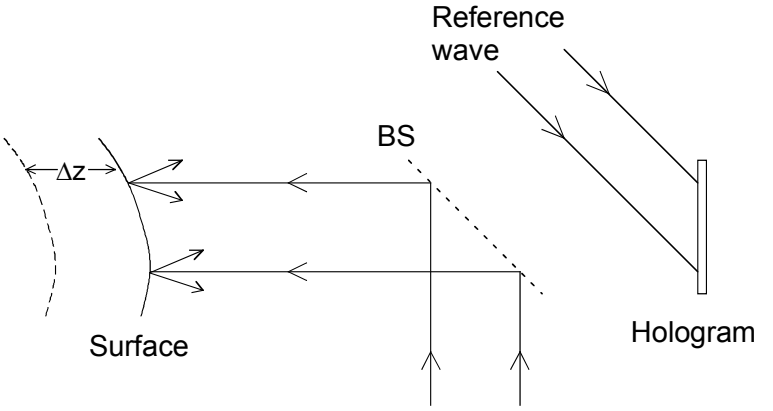


Fig. 2.17. Holographic contouring

$$z' = \frac{{z_R}^2 z}{z z_R + \frac{\lambda_2}{\lambda_1}{z_R}^2 - \frac{\lambda_2}{\lambda_1} z z_R} \approx z\frac{\lambda_1}{\lambda_2} \tag{2.87}$$

The indices "1" for virtual image ($z_1' \equiv z'$) and "O" for object ($z_O \equiv z$) are omitted and it is assumed not to change the source coordinates of the reconstruction wave with respect to those of the recording coordinates ($z_P \equiv z_R \to \infty$). The axial displacement of the image recorded with λ_1 but reconstructed with λ_2 with respect to the image recorded and reconstructed with λ_2 is therefore:

$$\Delta z = z' - z = z\frac{|\lambda_1 - \lambda_2|}{\lambda_2} \tag{2.88}$$

The path difference of the light rays on their way from the source to the surface and from the surface to the hologram is $2\Delta z$. The corresponding phase shift is:

$$\Delta\varphi(x,y) = \frac{2\pi}{\lambda_1} 2\Delta z = 4\pi z \frac{|\lambda_1 - \lambda_2|}{\lambda_1 \lambda_2} \tag{2.89}$$

The two shifted images interfere. According to Eq. (2.89) the phase shift depends on the distance z from the hologram plane. All points of the object surface having the same z-coordinate (height) are therefore connected by a contour line. As a result an image of the surface superimposed by contour fringes develops.

The height jump between adjacent fringes is:

$$\Delta H = z(\Delta\varphi = (n+1)2\pi) - z(\Delta\varphi = n2\pi) = \frac{\lambda_1\lambda_2}{2|\lambda_1 - \lambda_2|} = \frac{\Lambda}{2} \tag{2.90}$$

$\Lambda = \lambda_1\lambda_2 / |\lambda_1 - \lambda_2|$ is called *synthetic wavelength* or *equivalent wavelength*.

The object is intersected by parallel planes which have a distance of ΔH , see the principle in figure 2.18 and typical example in figure 2.19.

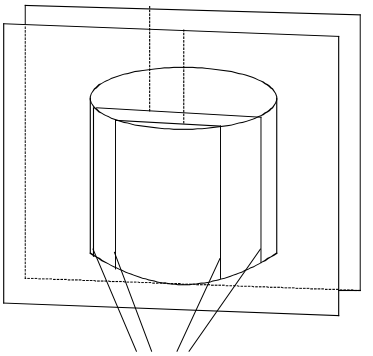


Fig. 2.18. Object intersection by contour lines

Fig. 2.19. Two-wavelength contour fringes

The equations derived in this chapter are valid only for small wavelength differences, because in addition to the axial displacement (which generates contour lines) also a lateral image displacement occurs. This lateral displacement can be neglected for small wavelength differences.

The principle of the two-illumination-point method is to make a double exposure hologram in which the point source illuminating the object is shifted slightly between the two exposures. If the illumination point S is shifted to S' between the two exposures (figure 2.20), the resulting optical path length difference δ is:

$$\begin{aligned}\delta &= \overline{SP} + \overline{PB} - \left(\overline{S'P} + \overline{PB}\right) = \overline{SP} - \overline{S'P} \\ &= \vec{s}_1 \overrightarrow{SP} - \vec{s}_2 \overrightarrow{S'P}\end{aligned} \tag{2.91}$$

The unit vectors $\vec{s}_1$ and $\vec{s}_2$ are defined as for the derivation of the interference phase due to deformation in chapter 2.7.2. The same approximation is used and these vectors are replaced by a common unit vector:

$$\vec{s}_1 = \vec{s}_2 = \vec{s} \tag{2.92}$$

Furthermore

$$\vec{p} = \overrightarrow{SP} - \overrightarrow{S'P} \tag{2.93}$$

is introduced as a vector from S to S'. The optical path difference is then given by

$$\delta = \vec{p}\,\vec{s} \tag{2.94}$$

The corresponding phase change is:

$$\Delta\varphi = \frac{2\pi}{\lambda}\vec{p}\,\vec{s} \tag{2.95}$$

The object surface is intersected by fringes which consist of a set of hyperboloids. Their common foci are the two points of illumination S and S'. If the dimensions of the object are small compared to the distances between the source points and the object, plane contouring surfaces result. A collimated illumination together with a telecentric imaging system also generates plane contouring surfaces. The distance between two neighbouring surfaces is

$$\Delta H = \frac{\lambda}{2\sin\dfrac{\theta}{2}} \tag{2.96}$$

where θ is the angle between the two illumination directions. Eq. (2.96) is analogue to the fringe spacing in an interference pattern formed by two intersecting plane waves, see Eq. (2.29) in chapter 2.2.

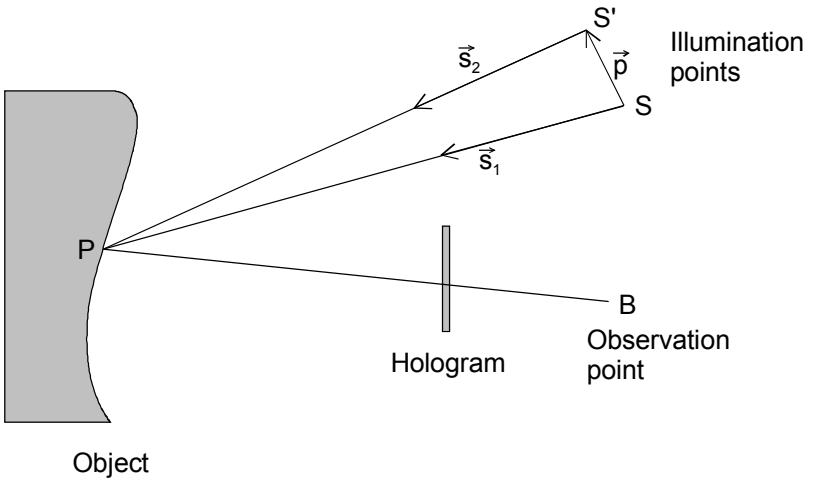


Fig. 2.20. Two-illumination point contouring

2.7.4 Refractive Index Measurement by HI

Another application of HI is the measurement of refractive index variations within transparent media. This mode of HI is used to determine temperature or concentration variations in fluid or gaseous media.

A refractive index change in a transparent medium causes a change of the optical path length and thereby a phase variation between two light waves passing the medium before and after the change. The interference phase due to refractive index variations is given by:

$$\Delta\varphi(x,y) = \frac{2\pi}{\lambda}\int_{l_1}^{l_2}[n(x,y,z) - n_0]dz \tag{2.97}$$

where n_0 is the refractive index of the medium under observation in its initial, unperturbed state and $n(x,y,z)$ is the final refractive index distribution. The light passes the medium in z-direction and the integration is taken along the propagation direction. Eq. (2.97) is valid for small refractive index gradients, where the light rays propagate along straight lines. The simplest case is that of a two-dimensional phase object with no variation of refractive index in z-direction. In this case the refractive index distribution $n(x,y)$ can be calculated directly from Eq. (2.97). In the general case of a refractive index varying also in z-direction Eq. (2.97) cannot be solved without further information about the process. However, in many practical experiments only two-dimensional phase objects have to be considered.

A set-up for the recording of holograms of transparent phase objects consists of a coherent light source, the transparent medium under investigation and optical components, see figure 2.21.

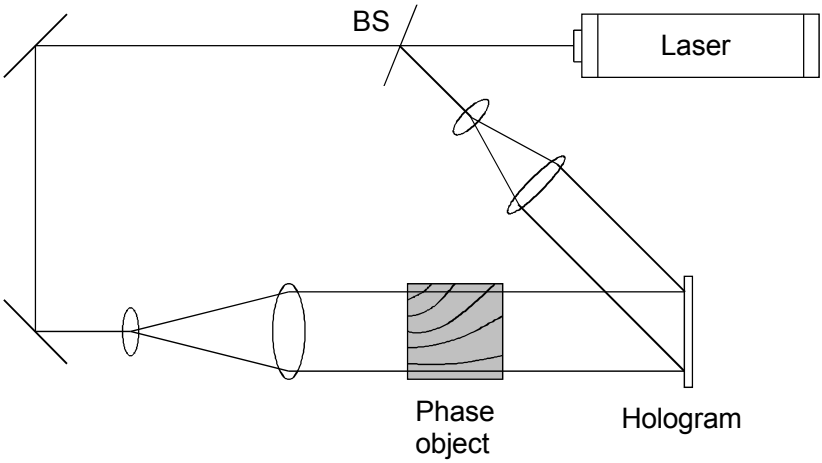


Fig. 2.21. Recording set-up for transparent phase objects

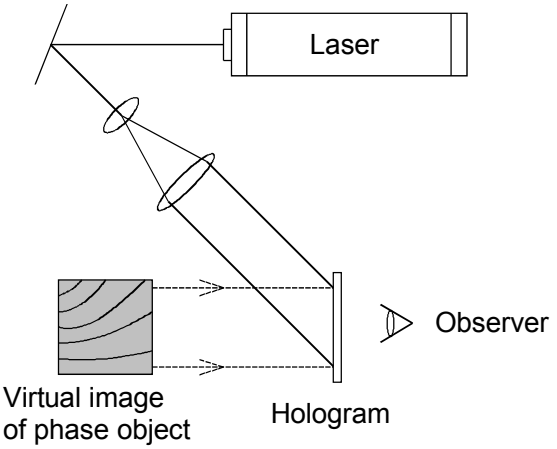


Fig. 2.22. Reconstruction of phase objects

The laser beam is split into two partial waves. One partial wave is expanded by a telescopic lens system and illuminates the medium, which is located e. g. in a test cell with transparent walls. The transmitted part, the object wave, interferes with the reference wave at the surface of the hologram plate. After processing the object wave is reconstructed by illuminating the hologram with the reference wave again, figure 2.22. Holographic Interferometry can be done either by the double exposure method or by the real-time method.

A holographic interferogram of a pure transparent object without any scattering parts consists of clear fringes, not disturbed by speckle noise. These fringes are not localized in space, because there are no object contours visible. Yet, for some

applications localized fringes are desired. In that case a diffusing screen has to be placed in front of or behind the object volume.

2.7.5 Phase Shifting HI

As already mentioned in chapter 2.7.1 it is not possible to calculate $\Delta\varphi$ unambiguously from the measured intensity, because the parameters $A(x,y)$ and $B(x,y)$ in Eq. (2.77) are not known and the sign is not determined.

Phase shifting Holographic Interferometry is a method to determine the interference phase by recording additional information [9, 20, 58, 60]. The principle is to record three or more interference patterns with mutual phase shifts. For the case of three recordings, the interference patterns are described by:

$$\begin{aligned} I_1(x,y) &= A(x,y) + B(x,y)\cos(\Delta\varphi) \\ I_2(x,y) &= A(x,y) + B(x,y)\cos(\Delta\varphi + \alpha) \\ I_3(x,y) &= A(x,y) + B(x,y)\cos(\Delta\varphi + 2\alpha) \end{aligned} \tag{2.98}$$

The equation system can be solved unambiguously for $\Delta\varphi$ if the phase angle α is known (e.g. 120°).

The phase shift can be realized in practice e. g. by a mirror mounted on a piezoelectric translator. The mirror is placed either in the object- or in the reference beam. If appropriate voltages are applied to the piezo during the hologram reconstruction, well defined path changes in the range of fractions of a wavelength can be introduced. These path changes correspond to phase differences between object- and reference wave.

Instead of using the minimum number of three reconstructions with two mutual phase shifts, Eq. (2.98), it is also possible to generate four reconstructions with three mutual phase shifts:

$$\begin{aligned} I_1(x,y) &= A(x,y) + B(x,y)\cos(\Delta\varphi) \\ I_2(x,y) &= A(x,y) + B(x,y)\cos(\Delta\varphi + \alpha) \\ I_3(x,y) &= A(x,y) + B(x,y)\cos(\Delta\varphi + 2\alpha) \\ I_4(x,y) &= A(x,y) + B(x,y)\cos(\Delta\varphi + 3\alpha) \end{aligned} \tag{2.99}$$

In that case the equation system can be solved without knowledge of the phase shift angle α, as long as it is constant. The solution for $\Delta\varphi$ is [71]:

$$\Delta\varphi = \arctan \frac{\sqrt{I_1 + I_2 - I_3 - I_4} \cdot \sqrt{3I_2 - 3I_3 - I_1 + I_4}}{I_2 + I_3 - I_1 - I_4} \tag{2.100}$$

Various HI phase shifting methods have been published, which differ in the number of recordings (at least 3), the value of α, the way to generate the phase shift (stepwise or continuously) or other details. These methods will not be dis-

cussed here in detail. The principle has been described briefly in order to prepare for a comparison of phase determination in conventional HI using photographic plates with the way to obtain phase information in Digital Holographic Interferometry (chapter 4). Finally it is remarked that phase shifting HI is not the only way to determine the phase from a fringe pattern, but it is the most applied method. Other phase evaluating techniques are the Fourier Transform Method, skeletonizing or the heterodyne technique.

2.7.6 Phase Unwrapping

Even after having determined the interference phase by a method such as HI phase shifting a problem remains: The cosine function is periodic, i. e. the interference phase distribution is indefinite to an additive integer of 2π.

$$\cos(\Delta\varphi) = \cos(\Delta\varphi + 2\pi n) \qquad n \in Z \tag{2.101}$$

Interference phase maps calculated with the arctan function or other inverse trigonometric functions therefore contain 2π jumps at those positions, where an extreme value of $\Delta\varphi$ (either $-\pi$ or π) is reached. The interference phase along a line of such a phase image looks like a saw tooth function, figure 2.23 (a). The correction of these modulo 2π jumps in order to generate a continuously phase distribution is called *demodulation*, *continuation* or *phase unwrapping*.

Several unwrapping algorithm have been developed in the last years. In the following the so called path-dependent unwrapping algorithm is described. At first a one-dimensional interference phase distribution is considered. The difference between the phase values of adjacent pixels $\Delta\varphi(n+1) - \Delta\varphi(n)$ is calculated. If this difference is less than $-\pi$ all phase values from the n+1 pixel onwards are increased by 2π. If this difference is greater than $+\pi$, 2π is subtracted from all phase values, starting at number n+1. If none of above mentioned conditions is valid the phase value remains unchanged. The practical implementation of this procedure is done by calculating first a step function, which cumulates the 2π jumps for all pixels, figure 2.23 (b). The continuous phase distribution is then calculated by adding this step function to the unwrapped phase distribution, figure 2.23 (c). Almost every pixel can be used as a starting point for this unwrapping procedure, not necessarily the pixel at the start of the line. If a central pixel is chosen as starting point the procedure has to be carried out in both directions from that point.

This one-dimensional unwrapping scheme can be transferred to two dimensions. One possibility is to unwrap first one row of the two dimensional phase map with the algorithm described above. The pixels of this unwrapped row act then as a starting points for column demodulation.

One disadvantage of the simple unwrapping procedure described here is that difficulties occur if masked regions are in the phase image. These masked areas might be caused e. g. by holes in the object surface. To avoid this and other difficulties several other, more sophisticated demodulation algorithm have been developed.

Finally it should be mentioned that the unwrapping procedure is always the same for all methods of metrology, which generate saw-tooth like images. This means the various unwrapping algorithm developed for HI and other methods can be used also for Digital Holographic Interferometry, because this technique also generates modulo 2π-images (see chapter 4).

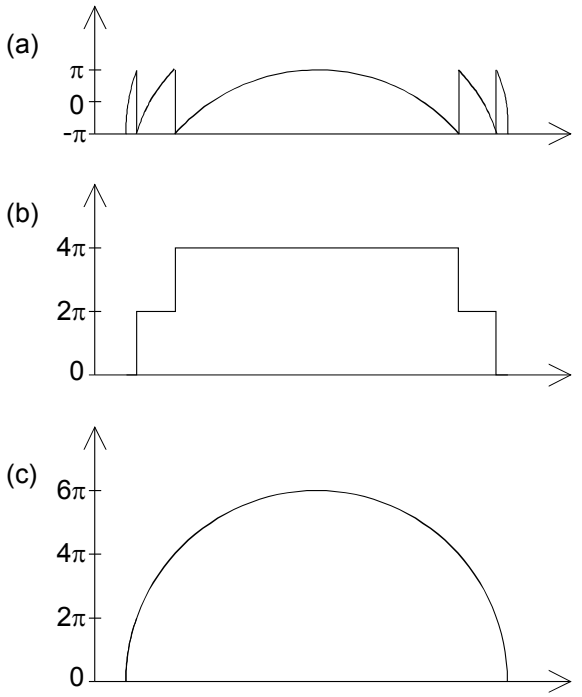


Fig. 2.23. Phase unwrapping

(a) Interference phase modulo 2π: $\Delta\varphi_{2\pi}(x)$

(b) Step function: $\Delta\varphi_{jump}(x)$

(c) unwrapped interference phase: $\Delta\varphi_{2\pi}(x)+\Delta\varphi_{jump}(x)$