Chapter 2
Fundamental Principles of Holography

2.1 Light Waves

The behaviour of light can be modelled either as a propagating electromagnetic (e-m) wave or as a stream of massless particles known as photons. Although the models are seemingly contradictory both are necessary to fully describe the full gamut of light phenomena. Whichever model is most appropriate depends on the phenomenon to be described or the experiment under investigation. For example, interaction of light with the atomic structure of matter is best described by the photon model: the theory of photon behaviour and its interactions is known as quantum optics. The phenomenon of refraction, diffraction and interference, however, are best described in terms of the wave model i.e. classical electromagnetism.

Interference and diffraction form the basis of holography An e-m wave is described in terms of the propagation through space of mutually perpendicular electric and magnetic fields. These fields oscillate in a plane that is perpendicular to the direction of travel i.e. they are described as transverse waves, as depicted in Fig. 2.1. Light waves can be described either by the electrical or by the magnetic field, but in optics convention is to describe the e-m wave in terms of the electric vector.

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

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

U. Schnars et al., *Digital Holography and Wavefront Sensing*,
DOI 10.1007/978-3-662-44693-5_2

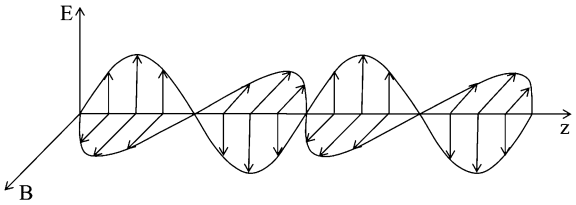

Fig. 2.1 Electromagnetic wave propagating in z-direction

and c is the speed of light in vacuum:

$$c = 2.9979 \times 10^8 \text{ m/s} \tag{2.3}$$

The electrical field $\vec{E}$ is a vector quantity and can vibrate in any direction perpendicular to the direction of 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

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

It can be easily verified that a linearly polarized, harmonic plane wave with amplitude

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

is a solution of the above 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 the *amplitude* of the wave. 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 the propagation direction. Points of equal phase are located on parallel planes that are perpendicular to the propagation direction. The modulus of $\vec{k}$ is the *wave number* and is described 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 through 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 the *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 7.5×10^{14} Hz to 3.8×10^{14} Hz. Light sensors such as the human eye, photodiodes, photographic film 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:

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

$\left\langle E^2 \right\rangle_t$ denotes the time average over many light periods. The constant factor $\varepsilon_0 c$ results if the intensity is formally derived from the Maxwell equations. The constant ε_0 is the vacuum permittivity. Note: we are using the term intensity here. In photometry and radiometry *intensity* has a different meaning (radiant power per solid angle, unit W sr^{-1}).

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 proportional to 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 the real part 'Re' can be omitted (in accordance with the superposition principle). 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 and Eq. (2.14) can be written as:

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

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, its *complex amplitude*, need be considered:

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

Equations (2.15) and (2.16) are not just valid for plane waves, but apply in general to three-dimensional waves whose amplitude, a, and phase, φ, are 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 complex conjugation. In many practical calculations where the absolute value of I is not of interest the factor $\frac{1}{2}\varepsilon_0 c$ can be neglected, and the intensity simplifies to $I = |A|^2$.

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, the superposition

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

is also a solution. 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 shall have the same polarization directions, i.e. scalar formalism can be used. The complex amplitudes of the respective waves are represented by;

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

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

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 can be written as

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

where I_1, I_2 are the individual intensities and the phase difference between the two waves is

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

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 when the phase difference between consecutive points is a multiple of 2π

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

This is known as *constructive interference*. The intensity reaches its minimum when

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

And this is known as *destructive interference*. The integer n is the interference order. An interference pattern therefore consists of a series of dark and light lines, "fringes", across the field-of-view as a result of this constructive and destructive interference. Scalar theory can also be applied to waves with different polarization directions, if the components of the electric field vector are considered.

The superposition of two plane waves which intersect at an angle θ with respect to each other results in an interference pattern with equidistant spacing, as seen in Fig. 2.2. The fringe spacing d is the distance from one interference maximum to the next and can be calculated from geometrical considerations. Figure 2.2 shows that

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

The quantities θ_1 and θ_2 are the angles between the propagation directions of the wavefronts and the vertical direction of the screen. The length Δl_2 is the path difference between wavefront W2 and wavefront W1 at the position of the interference maximum P1 (W2 has to travel a longer path to P1 than W1). At the neighboring maximum P2 the conditions are exchanged: now W1 has to travel a longer path; the path difference of W2 with respect to W1 is $-\Delta l_1$. The variation

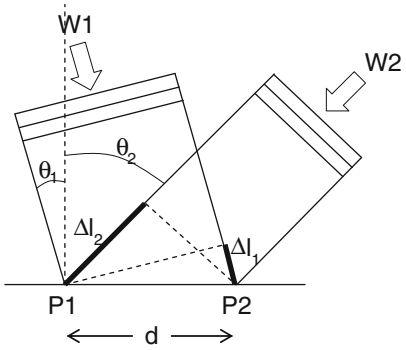

Fig. 2.2 Interference of two plane waves W1 and W2. The marginal rays are sketched. θ_1 is the angle between W1 and the vertical, θ_2 is the angle between W2 and the vertical. P1 and P2 are adjacent interference maxima

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 Eq. (2.27) gives the fringe spacing as:

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

The approximation $\cos(\theta_1 - \theta_2)/2 \approx 1$ and $\theta = \theta_1 + \theta_2$ can be applied to give

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

Instead of the fringe spacing d, the fringe pattern can also be described in terms of the spatial frequency f, which is just the reciprocal of d, i.e.

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

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 is visible.

In order to observe interference fringes, the phases of the individual waves have to be correlated. The ability of light to form interference patterns is called *coherence* and is investigated in this chapter. The two aspects of coherence are temporal and spatial coherence. Temporal coherence depends on the correlation of a wave with itself at different instants in time [121], whereas spatial coherence is based on the mutual correlation of different parts of the same wavefield in space.

2.3.2 Temporal Coherence

The phenomenon of interference between two coherent beams of light can be described in terms of a two beam interferometer such as the Michelson-interferometer, as shown in Fig. 2.3. Light emitted by the source S is split into two waves of reduced amplitude by the beam splitter BS. These waves travel to the mirrors M1 and M2 respectively, and are reflected back into their incident directions. After passing the beam splitter again they are superimposed at a screen. Usually the superimposed waves are not exactly parallel, but are incident at a small angle. As a result a two-dimensional interference pattern becomes visible.

The optical path length from BS to M1 and back to BS is s_1, and the optical path length from BS to M2 and back to BS is s_2. Experiments show 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 a uniform brightness is visible on the screen. The qualitative explanation for this phenomenon is that interference fringes can only develop if the superimposed waves have a well defined (constant) phase relationship between them. The phase difference between waves emitted by different sources varies randomly and thus the waves do not interfere. The atoms within the light source emit wave trains with a finite length L. If the optical path difference exceeds L, the recombined waves do not overlap after passing the different ways and interference is not observed.

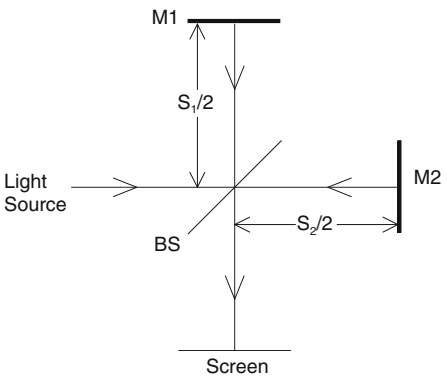

Fig. 2.3 Michelson's interferometer

The critical path length difference or, equivalently, the length of a wave train is the *coherence length* L of the wave. The corresponding time over which the wave train is emitted is its *coherence time,*

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

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

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

The coherence length is therefore a measure for the spectral linewidth of the source at a specific frequency, f. Light with a long coherence length accordingly has a correspondingly small linewidth and is therefore highly monochromatic.

Typical coherence lengths of light radiated from thermal sources, e.g. conventional 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 path lengths. On the other hand lasers have coherence lengths from a few millimetres (e.g. a multi-mode diode laser) to several 100 m (e.g. a stabilized single mode Nd:YAG-laser) up to several hundred kilometres for specially stabilized gas lasers used for research purposes.

The fringe visibility

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

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

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

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

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

$E(t)$ is the electrical field (to be precise: the complex analytical signal) of one interfering wave while $E(t + \tau)$ is the electrical field of the other wave. The latter is delayed in time by τ. Equation (2.35) represents the autocorrelation of the corresponding electric field amplitudes. The quantity

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

is the normalized self-coherence function; the absolute value of γ defines the degree of coherence.

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

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

The 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) yields

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

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 is therefore a measure of the ability of the two wave fields to interfere. When $|\gamma| = 1$ we have ideally monochromatic light or, likewise, light with infinite coherence length; when $|\gamma| = 0$ for the light is completely incoherent. Partially coherent light therefore lies in the range $0 < |\gamma| < 1$.

2.3.3 Spatial Coherence

Spatial coherence describes the mutual correlation of spatially separated parts of the same wavefield. This property can be measured using, for example, a Young's interferometer, Fig. 2.4. Here, an extended light source emits light from a large number of elementary point sources. An aperture with two transparent holes is mounted between the light source and the screen. The aim of the experiment is to determine the mutual correlation (degree of coherence) of the light incident on the aperture at the spatially separated positions given by the holes. If the light at these

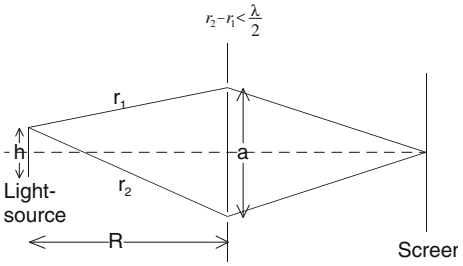

Fig. 2.4 Young's interferometer

positions is correlated, interference fringes are visible on the screen. In the following we will discuss the geometrical relations under which interference can be observed for the simple case of an extended light source. The fringes result from the different light paths traversed to the screen, either via the upper or via the lower hole in the aperture [250]. 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 is due to the waves emitted by different points of the extended light source being superimposed on the screen. It may happen that a particular source point generates an interference maximum at a certain point on the screen, while another source point generates a minimum at the same point. This occurs because the optical path difference is different for each source point. In general the contributions from all source points cancel and the contrast vanishes. This cancellation is avoided if the following condition is fulfilled for every point of the light source:

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

This condition is fulfilled if it is restricted to rays emanating from the edges of the light source. The following relations are valid for points at the edges:

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

where h is the width of the light source. Applying the assumptions $a \ll R$ and $h \ll R$ gives,

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

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

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

The coherence distance a_k is therefore given from,

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

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, but if the distance between the holes increases or the distance between the light source and the aperture decreases, Eq. (2.44) is violated and the interference vanishes.

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

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

where $\vec{r}_1$, $\vec{r}_2$ are the spatial vectors of the holes in the aperture of the Young interferometer. This cross correlation function is the *mutual coherence 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 $\vec{r}_2$. Equation (2.47) describes the degree of correlation between the lightfield at $\vec{r}_1$ at a 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 defined as the *complex degree of coherence*. The modulus of the mutual coherence function $|\gamma(\vec{r}_1, \vec{r}_2, \tau)|$ is measured with the Young interferometer.

2.4 Diffraction

Consider a light wave incident on an obstacle such as an opaque screen with some holes, or *vice versa*, a transparent medium with opaque obstructions. From geometrical optics it is known that a shadow is visible on a screen behind the obstacle. On closer examination, we see that 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 of the order of the wavelength of the incident light, then the light distribution is not sharply bounded, but forms a pattern of dark and bright regions. This is the phenomenon *diffraction*, see Fig. 2.5.

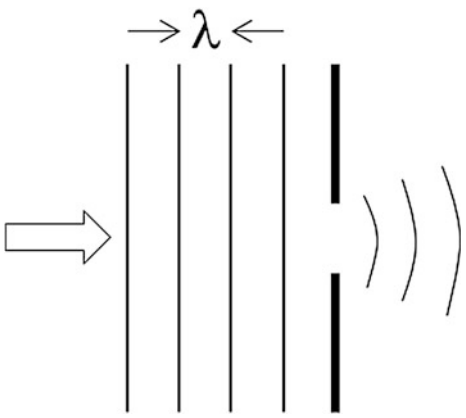

Fig. 2.5 Diffraction of a plane wave at an opaque screen with a small hole

Diffraction can be explained qualitatively with the *Huygens' principle*: *Every point of a wave front can be considered as a source point for secondary spherical waves. The wave field at any other place is the coherent superposition of these secondary waves.*

Huygens' principle is illustrated in Fig. 2.6.

The Fresnel-Kirchhoff integral describes diffraction quantitatively [116] as,

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

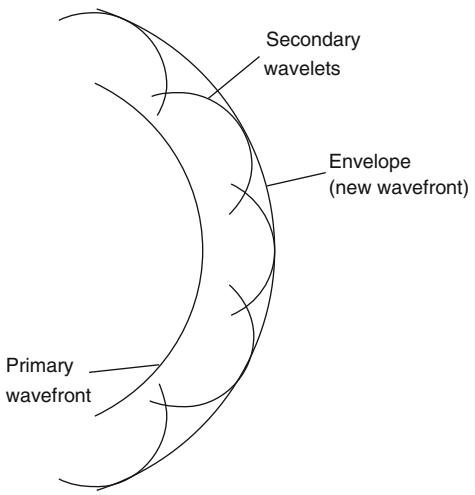

Fig. 2.6 Huygens' principle

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 diffracting aperture, see the coordinate system defined in Fig. 2.7. $\Gamma(\xi', \eta')$ is the complex amplitude in the observation plane. The term ρ' is the distance between a point in the aperture plane and a point in the observation plane.

Equation (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.

From the Huygens' principle it follows that the secondary waves not only propagate in the forward direction, but also back towards the source. Yet, experiment demonstrates that the wavefronts always propagate in one direction. To exclude this unrealistic situation the inclination factor Q defined in Eq. (2.50) is formally introduced into the Fresnel-Kirchhoff integral. Q depends on the angle θ between the incident light from the source and the unit vector $\vec{n}$ perpendicular to the aperture plane, and on the angle θ' between the diffracted light and $\vec{n}$, see Fig. 2.8. Q is approximately zero for $\theta \approx 0$ and $\theta' \approx \pi$. This excludes the concept of waves travelling in the backward direction. In most practical situations both θ and θ' are

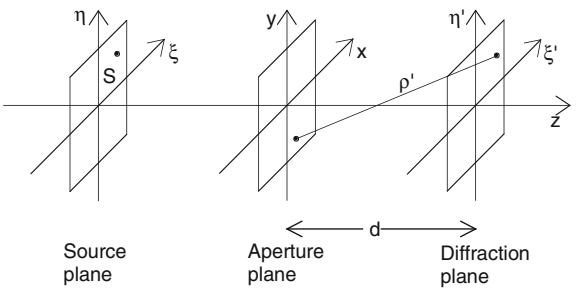

Fig. 2.7 Coordinate system

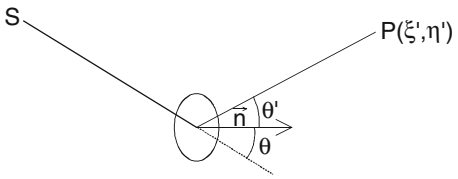

Fig. 2.8 Propagation geometry

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 [73, 116].

Some authors use a "+" sign in the argument of the exponential function of the Fresnel-Kirchhoff integral $[\Gamma(\xi,\eta) = \ldots A(x,y)\exp(+i2\pi/\lambda\rho')/\rho' \ldots]$ instead of the "−" sign used here. This is dependent on whether we define the harmonic wave in Eq. (2.14), as either $\exp(+i\varphi)$ 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.

2.5 Speckle

A rough surface illuminated with coherent light always appears "grainy" or "speckly" to an observer. This is due to the random fluctuations in intensity of the light scattered from the surface and gives rise to a series of and dark and bright spots or known as *speckle*, and forms a speckle pattern across the surface (Fig. 2.9). A speckle pattern develops if the height variations of the rough surface are larger than the wavelength of the light.

Speckle results from interference of light scattered by the surface points. The phase 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 is observed.

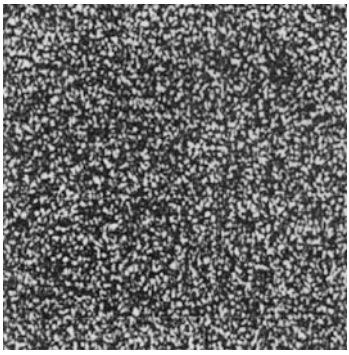

Fig. 2.9 A speckle pattern from a rough surface under coherent illumination

It can be shown that the probability density function for the intensity in a speckle pattern obeys negative exponential statistics [72]:

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

$P(I)dI$ is the probability that the intensity at a certain point is lies 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. The usual definition of the contrast of the speckle pattern is

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

The contrast of a speckle pattern is therefore always unity.

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, Fig. 2.10. There is no imaging system between the surface and the screen. The size of an individual speckle in an objective speckle pattern can be estimated using the spatial frequency formula of Eq. (2.30). The two edge points of the illuminated surface form the highest spatial frequency given as,

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

The reciprocal of $f_{\max}$ is a measure for the speckle size; and hence the diameter of the speckle is,

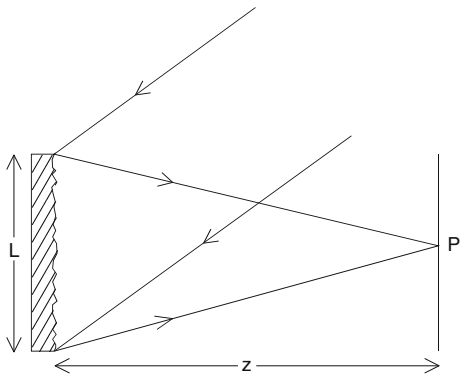

Fig. 2.10 Objective speckle formation

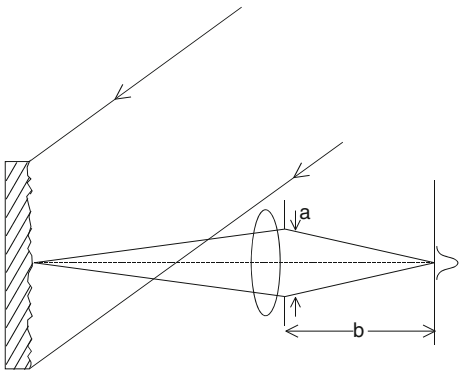

Fig. 2.11 Subjective speckle formation

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

A subjective speckle pattern develops if the illuminated surface is focused with an imaging system, e.g. a camera lens or the human eye, as in Fig. 2.11. 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:

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

where b is the image distance of the imaging system. It follows that the speckle diameter is given by

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

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

2.6 Holography

2.6.1 Hologram Recording and Reconstruction

Holograms are usually recorded with an optical set-up consisting of a light source (e.g. a laser), mirrors and lenses for beam guiding and a recording device (e.g. a photographic sensor). A typical set-up is shown in Fig. 2.12 [79, 121]. Light with sufficient coherence is split into two waves of reduced amplitude by a beam splitter

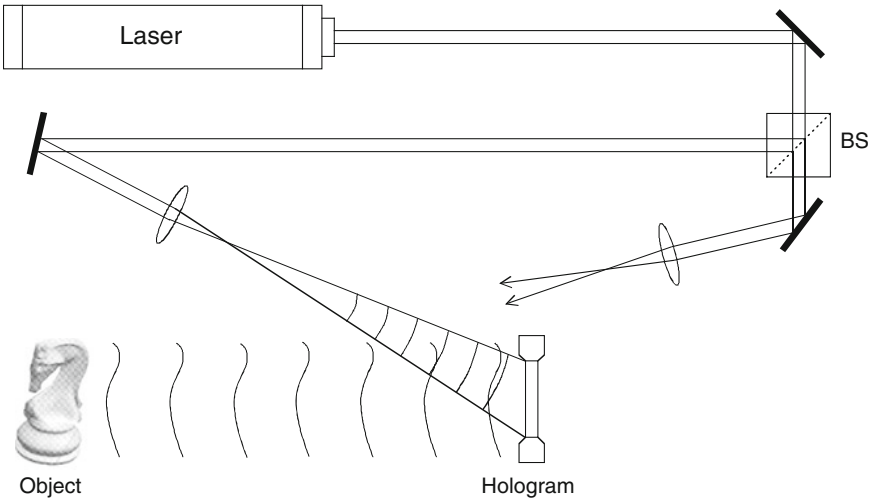

Fig. 2.12 Hologram recording

(BS). The first wave illuminates the object, is scattered at the object surface and reflected towards the recording medium. The second wave—the reference wave—directly illuminates the light sensitive medium. The waves interfere with each other to produce a characteristic interference pattern. In classical photographic holography the interference pattern is recorded on a photosensitive material such as silver halide films or plates and rendered permanent by wet chemical development of the film. In digital holography the interference pattern is recorded directly onto an electronic photosensor such as a CCD or CMOS array. The recorded interference pattern is the hologram.

The original object wave is reconstructed by illuminating the hologram with the reference wave, Fig. 2.13. An observer sees a virtual image, which is optically indistinguishable from the original object. The reconstructed image exhibits all effects of perspective, parallax and depth-of-field.

The holographic process is described mathematically using the formalism of Sect. 2.2. Across the extent of the photographic plate, 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.

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

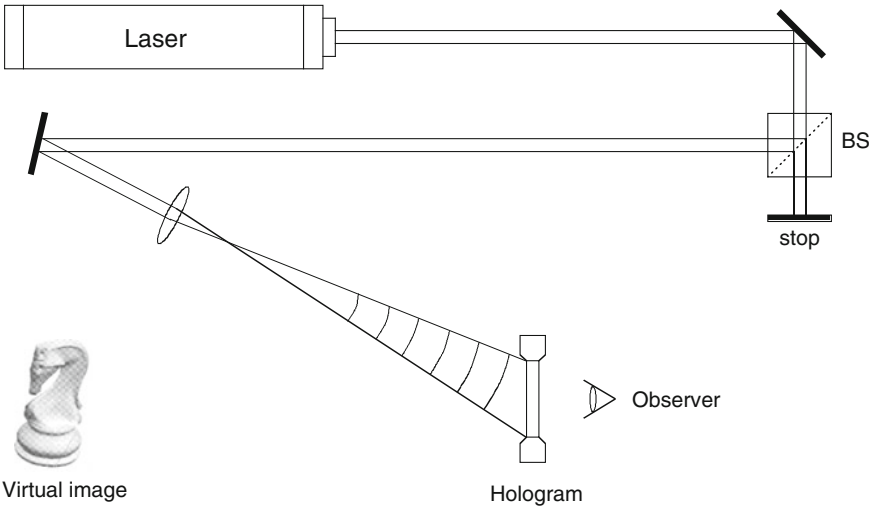

Fig. 2.13 Hologram reconstruction

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 and the resultant intensity is described 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 $I(x, y)$:

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

The constant β is the slope of the amplitude transmittance versus exposure characteristic of the light sensitive material. For photographic emulsions β is negative. The exposure duration is denoted by τ and h_0 is the amplitude transmission of the unexposed plate; $h(x,y)$ is the hologram function. In Digital Holography using CCD or CMOS arrays as the recording medium, h_0 can be neglected.

For hologram reconstruction in classical holography, the hologram is illuminated with a replica of the original reference wave in terms of wavelength and phase. This is represented mathematically as a multiplication of the amplitude transmission of the medium 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 constant factor. It represents the non-diffracted wave passing through the hologram (zero diffraction order). The second term is the reconstructed object wave and forms 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 non-diffracted 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 replaying the hologram with the complex conjugate of the reference beam E_R^*. This is mathematically represented by,

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

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 as 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 is also influenced by the reconstruction parameters.

The *imaging equations* relate the coordinates of an object point O to those of the corresponding point in the reconstructed image. These equations are quoted here without derivation but are described in some detail in other textbooks [79, 121].

The coordinate system is shown in Fig. 2.14. The coordinates of the object point O are denoted as (x_O, y_O, z_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. The ratio between the recording wavelength λ_1 and the reconstruction wavelength λ_2 is denoted by $\mu = \lambda_2/\lambda_1$. The coordinates of the point in the reconstructed virtual image, which corresponds to the object point O, are:

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

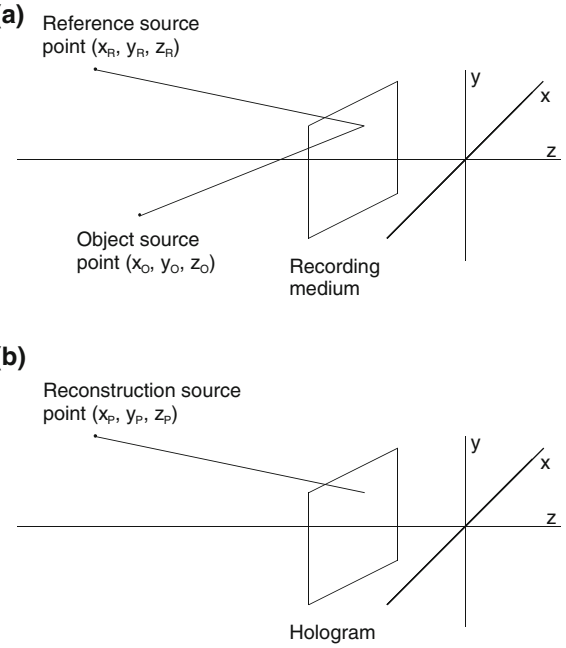

Fig. 2.14 Coordinate system used to describe holographic reconstruction. **a** Hologram recording. **b** Image reconstruction

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

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

The coordinates of the 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 equations. The lateral magnification of the entire virtual image is:

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

The lateral magnification of the real image is given by,

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

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 which should be noted: 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 of measuring optical path length variations, which are caused by deformations of opaque bodies or refractive index variations in transparent media, e.g. fluids or gases [175]. HI is a non-contact, non-destructive metrological technique with a very high measurement sensitivity. Optical path changes up to one hundredth of a wavelength are resolvable.

Two coherent wave fields, which are reflected from an object when it is in two different states of excitation, interfere. This is achieved e.g. in double-exposure

holography by the recording of two wave fields on a single photographic plate, Fig. 2.15. The first exposure represents the object in its reference state (undeformed state), the second exposure represents the object in its loaded (deformed) state. The hologram is reconstructed by illumination with the reference wave, Fig. 2.16. As a result of the superposition of two holographic recordings with *slightly* different

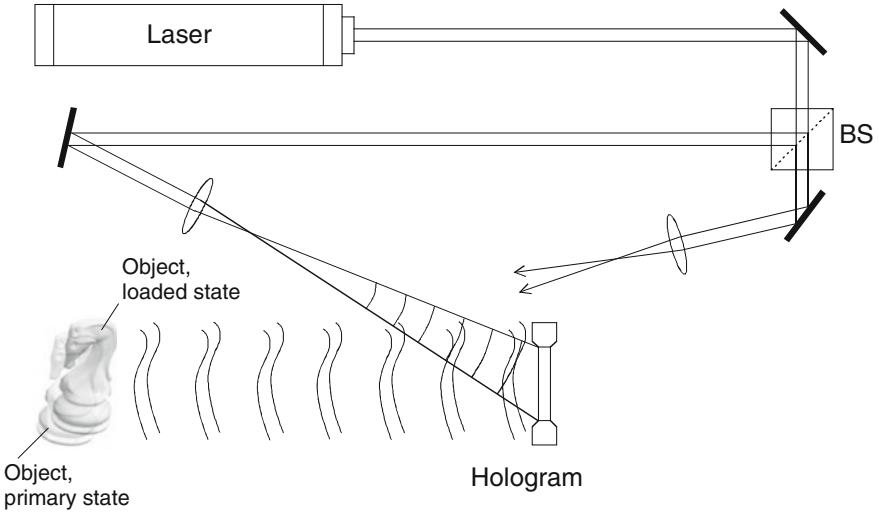

Fig. 2.15 Recording of a double exposed hologram

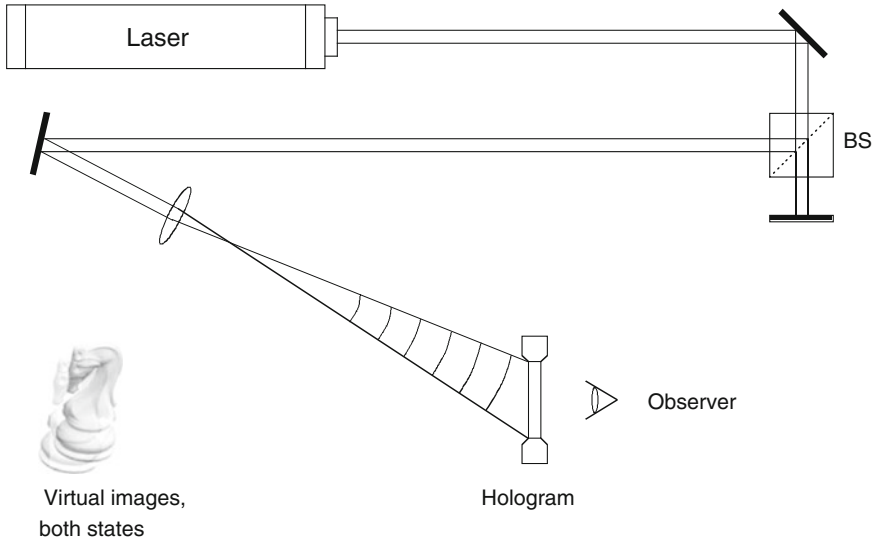

Fig. 2.16 Reconstruction of the double-exposed hologram

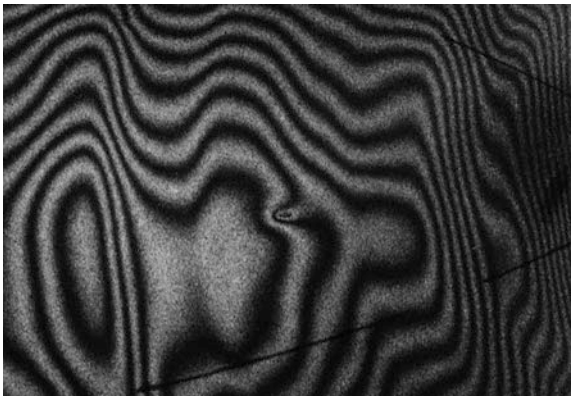

Fig. 2.17 A holographic interferogram of a pressure vessel

object waves, only one image superimposed by interference fringes is visible, see example in Fig. 2.17. From this holographic interferogram the observer can determine optical path changes due to the object deformation or other effects.

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

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

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

where $a(x,y)$ is the real amplitude and $\varphi(x,y)$ is 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 $\varphi + \Delta\varphi$. The term $\Delta\varphi$ represents the difference between the reference and the actual phase and is known as the *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, such as,

- uneven illumination of the object due to the Gaussian profile the expanded laser beam gives rise to varying brightness of the holographic interferogram.
- high frequency speckle noise is superimposed upon interferogram.
- additional superimposed diffraction patterns due to dust particles in the optical path.
- the varying reflectivity of the object under investigation may influence the brightness and visibility of the interferogram.
- electronic recording and transmission of holographic interferograms can generate additional noise.

Equation (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 ($\cos 30° = \cos -30°$) and the sign of $\Delta\varphi$ cannot be 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 Sect. 2.7.5.

2.7.2 Displacement Measurement by HI

In this chapter a relationship between the measured interference phase and the displacement of the object surface under investigation is derived [121, 218]. The geometric quantities are explained in Fig. 2.18. The vector $\vec{d}(x,y,z)$ is the 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. The terms $\vec{s}_1$ and $\vec{s}_2$ are unit vectors from the illumination source point S to P_1, and P_2 respectively. Similarly, $\vec{b}_1$ and $\vec{b}_2$ are unit vectors from P_1 to the observation point B, and from P_2 to B, respectively. 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 therefore given by,

$$\begin{aligned}\delta &= \overline{SP_1} + \overline{P_1B} - \left(\overline{SP_2} + \overline{P_2B}\right) \\ &= \vec{s_1}\,\overrightarrow{SP_1} + \vec{b_1}\,\overrightarrow{P_1B} - \vec{s_2}\,\overrightarrow{SP_2} - \vec{b_2}\,\overrightarrow{P_2B}\end{aligned} \tag{2.78}$$

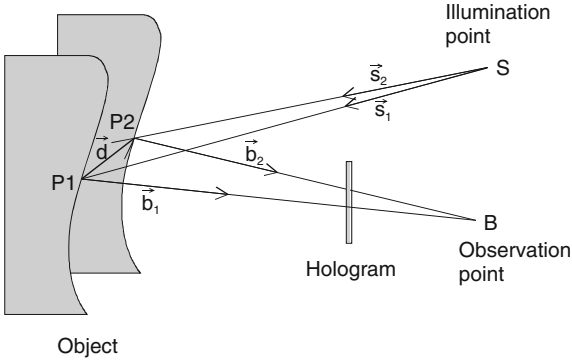

Fig. 2.18 Calculation of the interference phase

The lengths $\overline{SP_{1/2}}$ and $\overline{P_{1/2}B}$ are in the range of metres, while $\left|\vec{d}\right|$ is in the range of several micrometres. 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:

$$\vec{d} = \overrightarrow{P_1B} - \overrightarrow{P_2B} \tag{2.81}$$

and

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

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

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

The following expression results for the interference phase:

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

The vector

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

is called the *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. Equation (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 it is not the three dimensional displacement field that 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 $\lambda/2$.

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 Fig. 2.19. 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 Fig. 2.19 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

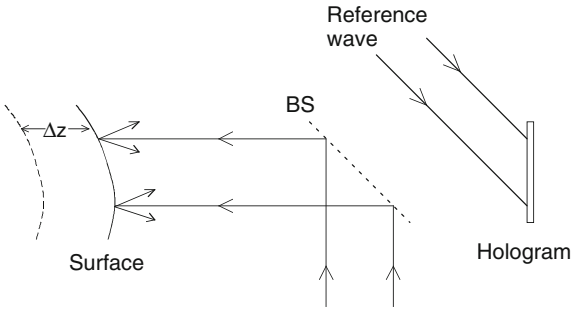

Fig. 2.19 Holographic contouring

photographic plate. This can be done either simultaneously using two lasers with different wavelengths or in succession by 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 Eq. (2.66):

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

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 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 thus,

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

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 known as the *synthetic wavelength* or *equivalent wavelength*. The object is intersected by parallel planes which have a distance of ΔH, see the principle in Fig. 2.20 and a typical example in Fig. 2.21.

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 (Fig. 2.22), 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 Sect. 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

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

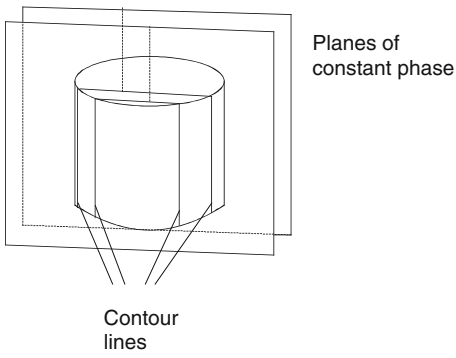

Fig. 2.20 Object intersection by contour lines

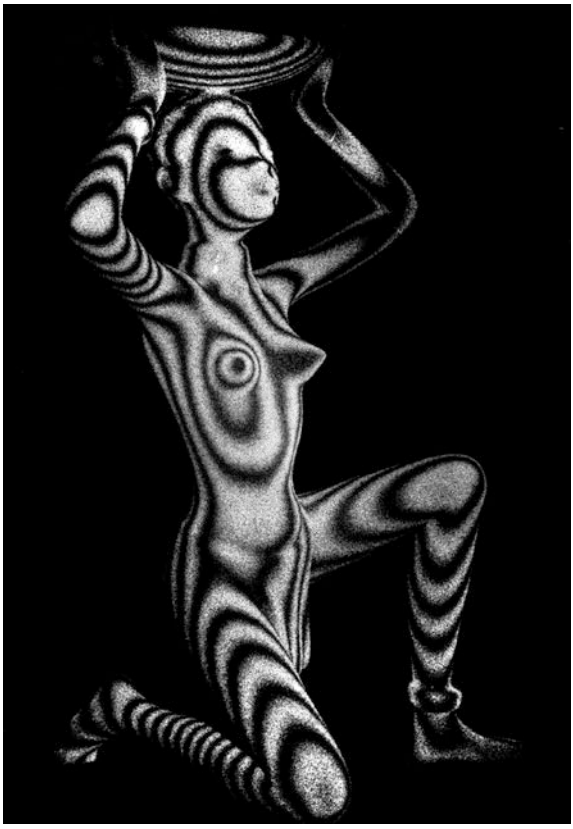

Fig. 2.21 Two-wavelength contour fringes

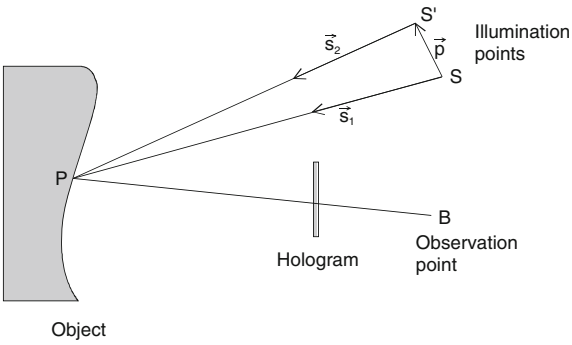

Fig. 2.22 Two-illumination point contouring

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

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

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

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 through the medium in the z-direction and integration is along the propagation direction. Equation (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. 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 the 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 as in Fig. 2.23.

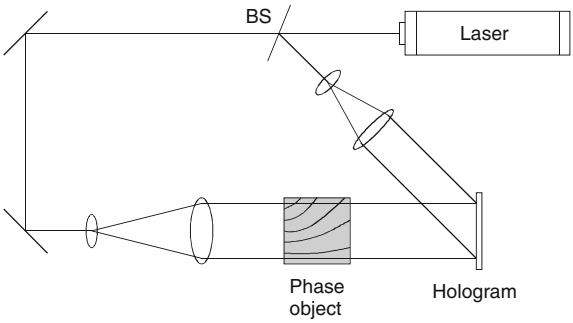

Fig. 2.23 Recording set-up for transparent phase objects

The laser beam is split into two separate waves. One wave is expanded by a telescopic lens system and illuminates the medium, which is located for example 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, Fig. 2.24. Holographic Interferometry can be carried out either by the double exposure method or by the real-time method.

A holographic interferogram of a pure transparent object without any scattering consists of clear fringes undisturbed 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 can be placed in front of or behind the object volume.

2.7.5 Phase Shifting HI

As discussed in Sect. 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 which enables us to determine the interference phase by recording additional information [17, 36, 98,

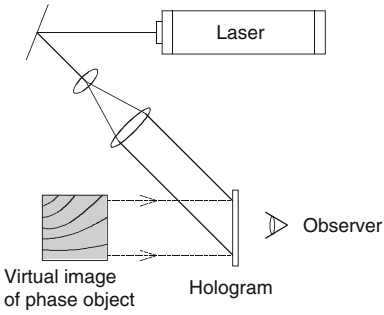

Fig. 2.24 Reconstruction of phase objects

99]. 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 Δφ if the phase angle α is known (e.g. 120°).

The phase shift can be realized in practice for example by employing a mirror mounted on a piezo-electric translator. The mirror is placed either in the object beam or in the reference beam. If appropriate voltages are applied to the piezo-electric translator 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 Δφ is [121]:

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

Various HI phase shifting methods have been published [121], which differ in the number of recordings (at least 3), the value of α, and the method of generating the phase shift (stepwise or continuously). These methods will not be discussed in detail here. The principle has been described briefly in order to prepare for a comparison of phase determination in conventional HI using photographic plates and with the techniques used to obtain phase information in Digital Holographic Interferometry (Chap. 4). Finally it is noted that phase shifting HI is not the only way to determine the phase from a fringe pattern, but it is the most commonly applied. Other phase evaluating techniques include Fourier Transform methods, skeletonizing or heterodyne techniques.

2.7.6 Phase Unwrapping

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

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

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 change along a line of such a phase image resembles a saw tooth function, Fig. 2.25a. The correction of these modulo 2π jumps in order to generate a continuous phase distribution is called *demodulation*, *continuation* or *phase unwrapping*.

Several unwrapping algorithms 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)th pixel onwards are increased by 2π. If this difference is greater than $+\pi$, 2π is subtracted from all phase values, starting from (n + 1). If none of the above mentioned conditions is valid the phase value remains unchanged. The practical implementation of this procedure is done by first calculating a step function, which cumulates the 2π jumps for all pixels, Fig. 2.25b. The continuous phase distribution is then calculated by adding

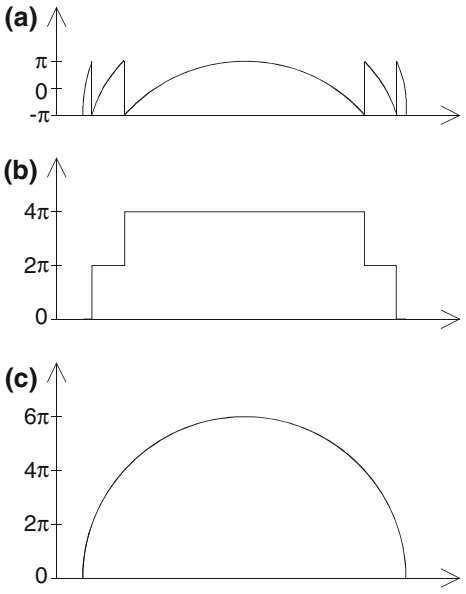

Fig. 2.25 Phase unwrapping. **a** Interference phase modulo 2π: $\Delta\varphi_{2\pi}(x)$ **b** Step function: $\Delta\varphi_{jump}(x)$ **c** unwrapped interference phase: $\Delta\varphi_{2\pi}(x) + \Delta\varphi_{jump}(x)$

this step function to the unwrapped phase distribution, Fig. 2.25c. 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 the 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 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 by e.g. holes in the object surface. To avoid this and other difficulties several other, more sophisticated demodulation algorithms have been developed [121].

Finally it should be mentioned that the unwrapping procedure is always the same for all methods of metrology that 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 Chap. 4).

http://www.springer.com/978-3-662-44692-8

Digital Holography and Wavefront Sensing
Principles, Techniques and Applications
Schnars, U.; Falldorf, C.; Watson, J.; Jüptner, W.
2015, XI, 226 p. 145 illus., 27 illus. in color., Hardcover
ISBN: 978-3-662-44692-8