ZÁPADOČESKÁ
UNIVERZITA
V PLZNI

University of West Bohemia in Pilsen
Department of Computer Science and Engineering
Univerzitni 8
30614 Pilsen
Czech Republic

Holography Principles

Technical report

Martin Janda, Ivo Hanák, and Václav Skala

Technical Report No. DCSE/TR-2006-08
December, 2006

Distribution: public

Technical Report No. DCSE/TR-2006-08
December 2006

Holography Principles

Martin Janda, Ivo Hanák, and Václav Skala

Abstract

This report summarise mathematical background of the holography phenomenon along with the overview of techniques and application of the optical holography. Moreover, specialties of digital holography are presented. This report should serve as a introductory document for whomever wish to understand to holography.

This work has been partialy supported by the Ministry of Education, Youth and Sports of the Czech Republic under the research program LC-06008 (Center for Computer Graphics). This work has been partialy supported by the EU project EU within FP6 under Grant 511568 with the acronym 3DTV.

Copies of this report are available on
`http://www.kiv.zcu.cz/publications/`
or by surface mail on request sent to the following address:

University of West Bohemia in Pilsen
Department of Computer Science and Engineering
Univerzitni 8
30614 Pilsen
Czech Republic

Contents

Chapter 1

Introduction

Holography is quite old scientific field. As an inventor of holography could be claimed prof. D. Gabor [Gab49]. He proposed holographic imaging when working on enhancing resolution of electron microscopy. However, the first holograms produced poor quality images and the development of holography stagnated for a while. The technology was greatly improved after introduction of the off-axis holograms and invention of the LASER in sixties. Since then, holography found many applications including 3D imaging and interferometry. Optical holography is introduced in the Section 3.

The efforts for bringing holography to the digital world of computers are not new either. The first attempts were already done in 1967, however, the first useful results had to wait for sufficient computational power which has been achieved not until nineties of the 20th century. For example, the first digital holograms computed at interactive rates were described in 1994 in [Luc94]. Digital holography is introduced in the Section 4.

Holography is built on quite complex physical laws of optics. These laws are referenced many times in the text of this thesis and therefore basics of the wave optics are provided in the Section 2 of this chapter. The described topics are wave equation, interference, coherence and diffraction.

Chapter 2

Holography physics

The whole holography relies heavily on quite complex rules and laws of wave optics. Wave optics considers light as an electromagnetic wave of an arbitrary wavelength in general. However, the most interesting, in the context of this thesis, is the interval ranging from 300 to 700 nm because this interval constitutes the visible light. Visible light, just light from now on, interacts with its surroundings and with itself at microscopic levels and in a quite complex manner. This interaction is referred as interference. The interference in particular is described in the section 2.2.

The principle of the interference phenomenon is based on the wave nature of the light and therefore a short introduction into the mathematics of waves is provided in the section 2.1. The relation of the wave calculus to the physics of light is also presented there.

From interference the more complex phenomenon of diffraction is derived. The diffraction is responsible for forming the object beam on a photographic plate, see the Section 3 for reference. The exact mathematical model of the diffraction was not found yet. However, there are some approximated models but, though approximated, they provide sufficiently accurate results. The diffraction models are introduced in the Section 2.5.

2.1 Wave optics

The light is, in general, an electromagnetic wave of some wavelength spectrum. The visible light is on the interval ranging from 300 nm to 700 nm. The light with wavelength longer than 700 nm is called infrared and light with wavelength shorter than 300 nm is called ultraviolet. The visible band is, of course, the most interesting one since it is visible by a human observer.

The electromagnetic wave consists of the time varying electric and magnetic fields which are tightly coupled as it is evident from Maxwell's equations. These simplified Maxwell's equations Equation (2.1), Equation (2.2) applies for vacuum:

$$\nabla \cdot \mathbf{E} = 0 \tag{2.1}$$

$$\nabla \cdot \mathbf{H} = 0 \tag{2.2}$$

$$\nabla \times \mathbf{E} = -\mu_0 \frac{\partial \mathbf{H}}{\partial t} \tag{2.3}$$

$$\nabla \times \mathbf{H} = \epsilon_0 \frac{\partial \mathbf{E}}{\partial t} \tag{2.4}$$

The notation in the Maxwell's equations is following: the $\mathbf{E}$ denotes electric field, $\mathbf{H}$ denotes magnetic field, μ_0 denotes permeability of vacuum, ϵ_0 denotes permittivity of vacuum, $\nabla\cdot$ denotes divergence operator and $\nabla\times$ denotes curl operator. The solution of those Maxwell equations are two sinusoidal plane waves, with the electric and magnetic field directions orthogonal to one other and the direction of travel, and with two fields in phase, travelling at the speed of light in vacuum.

By rewriting the Maxwell's simplified equations, one obtains the equations:

$$\nabla^2\mathbf{E} = \mu_0\epsilon_0\frac{\partial^2}{\partial t^2}\mathbf{E} \tag{2.5}$$

$$\nabla^2\mathbf{B} = \mu_0\epsilon_0\frac{\partial^2}{\partial t^2}\mathbf{B} \tag{2.6}$$

The equations Equation (2.5) and Equation (2.6) are vector equations, but under some circumstances, all components of the vectors behaves exactly the same and a single scalar equation can be used to describe the behavior of the electromagnetic disturbance. The scalar equation can be written in this form:

$$\nabla^2 u(\mathbf{p},t) - \frac{n^2}{c^2}\frac{\partial^2 u(\mathbf{p},t)}{\partial t^2} = 0, \tag{2.7}$$

where $u(\mathbf{p},t)$ represents a scalar field component at the given position $\mathbf{p}$ and time t examined in a material of refractive index n. The light travels through this material at a speed of c/n, where c is a speed of light.

The Equation (2.7) describes the behavior of a wave in a linear, uniform, isotropic, homogeneous, and non-dispersive material and it is a base for the scalar wave theory that serves as base upon which all assumptions in this thesis are build on. Even thought the scalar wave theory is an approximation rather than an exact description it is satisfactorily as it describes the behavior of the light wave in concordance with physical experiments. The error introduced by the approximation is small and it is recognisable only at the distance of few wavelengths from the aperture's boundary.

A time-varying scalar field for a monochromatic wave in the scalar wave theory at position $\mathbf{P}$ is:

$$u(\mathbf{p},t) = A(\mathbf{p})\cos[2\pi\nu t - \varphi(\mathbf{p})], \tag{2.8}$$

where ν is an optical frequency of the wave in [Hz], $A(\mathbf{p})$ and $\varphi(\mathbf{p})$ defines amplitude and phase respectively of the wave at the position $\mathbf{p}$. This equation describes the wave properly yet the more convenient notation is:

$$\begin{aligned} u(\mathbf{p},t) &= \Re\left\{\tilde{u}(\mathbf{p})\exp(-\mathrm{i}2\pi\nu t)\right\} = \Re\left\{\tilde{u}(\mathbf{p})\exp(-\mathrm{i}\omega t)\right\}, \\ \tilde{u}(\mathbf{p},t) &= \tilde{u}(\mathbf{p})\exp(-\mathrm{i}2\pi\nu t), \end{aligned} \tag{2.9}$$

where ω is an angular speed and $\tilde{u}(\mathbf{p})$ is a complex amplitude defined as:

$$\tilde{u}(\mathbf{p}) = A(\mathbf{p})\exp[\mathrm{i}\varphi(\mathbf{p})]. \tag{2.10}$$

The function $u(\mathbf{p},t)$ is known as the wavefunction. A term diffraction pattern refers to an array a complex notations for a wavefunction $u(\mathbf{p},t)$ defined by the Equation (2.9). The wavelength of a light is defined as $\lambda = c/n\nu = \lambda_0/n$, where λ_0 is a wavelength in a vacuum. The following text assumes the propagation is done in a vacuum, if not noted otherwise.

A wave defined by the Equation (2.9) has to satisfy the scalar wave theory, i.e. Equation (2.7). If the Equation (2.9) is substituted into the scalar wave theory a relation known as the **H**elmholtz equation is obtained:

$$(\nabla^2 + k^2)\tilde{u}(\mathbf{p}) = 0, \tag{2.11}$$

where $k = 2\pi/\lambda$ is known as a **w**avenumber. A solution to the Helmholtz equation defines waves of various forms including basic ones such as planar wave and spherical wave that are described further in the text. Note, that the Helmholtz equation describes only a spatial part of a complete solution thanks to the fact that $\tilde{u}(\mathbf{p}, t)$ is separable, i.e. $\tilde{u}(\mathbf{p}, t) = \tilde{u}(\mathbf{p})\tilde{t}(t)$. The temporal part of the solution is a linear combination of sine and cosine function and thus it is not considered in following sections.

The relation between the ray and the wave optics is straightforward. A relation is clearly visible from the specification of a wavefront. **W**avefront is an iso-surface that consist of points with wave function samples of the same phase, i.e. $\varphi(\mathbf{p}) = 2\pi q, q \in \langle 0; 1\rangle$. A gradient of the phase φ is a normal of the wavefront's surface at a given point $\mathbf{p}$. Besides that it gives a direction of a local propagation for the wave and thus it gives a direction of a ray in a ray optics [Kra04].

Another important feature of the light is its optical power. This is important when creating of a final image because it defines an amount of energy delivered to a photographic material and/or sensor. **O**ptical intensity is defined as a time average of an amount of energy that crosses an unit perpendicular to the energy flow during a unit of time. If the time period is short enough the intensity of wave $\tilde{u}(\mathbf{p})$ is equal to $|\tilde{u}|^2$, i.e. it is a complex multiplication of $\tilde{u}$ with its complex conjugate [Har96]:

$$I = \tilde{u}(\mathbf{p})\tilde{u}^*(\mathbf{p}) = |\tilde{u}(\mathbf{p})|^2. \tag{2.12}$$

For computing a hologram, one has to compute the light field at the hologram plane first and then the optical intensity is evaluated to create the actual interference pattern or fringe pattern.

2.2 Interference

In the Section 2.1 the light was described as a wave. However, there is rarely just one wave present in a space. There are usually many waves and each one can interact with the other. This interaction is called interference.

The simplest situation is if two waves travel in the same direction. According to the phase of each wave the resulting electrical intensity will increase or decrease. If phases at some point in a space are the same or near the same, the constructive interference occurs at that point. If the phases are opposite or almost opposite the destructive interference occurs, see FigureFigure 2.1.

As a result the optical intensity due to two interfering waves is increased in the case of constructive interference and decreased in the case of destructive interference. This is quite surprising result that by adding two lights one can obtain less or none light. However this is true for the coherent light only.

The coherence is described in the Section 2.3 but it basically determines the stability of the interference effect in time. The coherent light produces stable interference pattern, i.e.

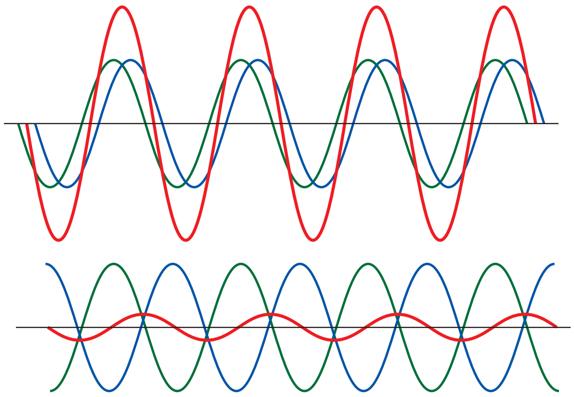

Figure 2.1: [v0.1]Two results of interference of two waves (green and blue). Constructive interference in upper and destructive interference in the lower.

just constructive or just destructive, on the contrary the incoherent light creates instable interference pattern, i.e. constructive interference changes quickly into the destructive interference. This flickering is so fast that the human eye is incapable of registering it because the human eye integrates the incoming intensity and the average value is obtained.

The coherence is also essential in holography because during the process of creating optical hologram a photosensitive material captures the interference pattern formed by the interference. Such pattern has to be stable for successful capture. This is the reason why coherent light is necessary for holography purposes and why lasers are usually used since they are sources of very coherent light.

The interference can be described mathematically utilizing the complex algebra. The advantage of the complex notation of the wave equation now emerges. The Equation (2.13) demonstrates that interference can be written as a summation of the complex amplitudes of each interfering wave assuming that monochromatic or in other words coherent wave is considered.

$$\tilde{u} = \tilde{u}_1 + \tilde{u}_2 + \cdots + \tilde{u}_n \tag{2.13}$$

The optical intensity due to the interference of two waves is therefore computed according to the Equation (2.12) as

$$\begin{aligned} I &= \left|\tilde{u}_1 + \tilde{u}_2\right|^2, \\ &= \left|\tilde{u}_1\right|^2 + \left|\tilde{u}_2\right|^2 + \tilde{u}_1\tilde{u}_2^* + \tilde{u}_1^*\tilde{u}_2, \\ &= I_1 + I_2 + 2\sqrt{I_1 I_2}\cos\left(\phi_1 - \phi_2\right). \end{aligned} \tag{2.14}$$

The Equation (2.14) is very important for holography. It describes the computation of intensity of interference of two waves with complex amplitudes $\tilde{u}_1$ and $\tilde{u}_2$ constituting the scene wave and the reference wave respectively, see Section 3.1. The intensity is a result of adding the intensities of both waves and the variation term $\cos\left(\phi_1 - \phi_2\right)$, which represents the interference phenomenon. The angles ϕ_1 and ϕ_2 are starting phases of the waves. The optical intensity therefore depends only on the phase difference. Sometimes the cosine term is called the bi-polar intensity, see [Luc94].

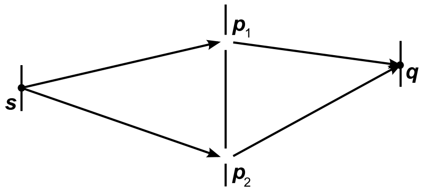

Figure 2.2: A configuration of exploring the coherence [Har96].

2.3 Coherence

The coherence is quite important property a light should exhibit to be useful for holography purposes. It is very important in a relation to the interference which is described in the previous section. It is therefore appropriate to explain what coherence is.

In general, coherence quantifies the ability of the light to form a visible diffraction pattern. It directly influences the quality of the visibility of the interference pattern which consists of areas with different degree of constructive or destructive interference. The areas are usually referenced as fringes.

The fringes are more visible if two interfering waves are more coherent and they are less visible if waves are less coherent – the degree of coherence. In other words, coherence determines the ability of two interfering waves to create total destructive interference. While perfectly coherent waves create a clean visible interference pattern that has not boundary by any spatial or temporal constraints, the incoherent ones won't create visible interference fringe at all.

Let us consider a point source that emits a monochromatic wave for an infinite time period. In such case, the configuration depicted in the Figure 2.2 forms two secondary sources $\mathbf{p}_1$ and $\mathbf{p}_2$. Waves generated by these two sources are coherent without limitations described below.

Yet, in the reality light sources are not ideal because they are not strictly monochromatic. A linearly polarized quasi-monochromatic at a given point can be represented by an analytic signal [Har96]:

$$\tilde{v}(\mathbf{p}, t) = \int_0^\infty \tilde{u}_\omega(\mathbf{p}, t)\,\mathrm{d}\omega, \tag{2.15}$$

where $\tilde{u}_\omega(\mathbf{p}, t)$ describes a wave of angular frequency ω.

A complex coherence of waves generated by two secondary light sources $\mathbf{p}_1$ and $\mathbf{p}_2$ formed according to the Figure 2.2 is defined as an normalized cross-correlation of two stationary random functions. The cross-correlation of two stationary time-dependent functions $g(t)$ and $h(t)$ is defined as follows [Har96]:

$$R(\tau) = \frac{1}{2T}\int_{-T}^{T} g^*(t)h(t+\tau)\,\mathrm{d}t = \langle g^*(t)h(t+\tau)\rangle. \tag{2.16}$$

The complex coherence depends on a time delay τ. For the viewing point $\mathbf{q}$ in the Figure 2.2 the time delay represents a difference between transit times for paths $\mathbf{p}_1\mathbf{q}$ and $\mathbf{p}_2\mathbf{q}$. Based on the Equations Equation (2.15) and Equation (2.16), the complex coherence, also known as the complex degree of coherence, of two light waves is a normalized cross-correlation between $\tilde{v}_1$ and $\tilde{v}_2$ [Har96]:

$$\tilde{\gamma}_{12}(\tau) = \frac{\langle \tilde{v}_1(t+\tau)\tilde{v}_2^*(t)\rangle}{[\langle \tilde{v}_1(t)\tilde{v}_1^*(t)\rangle\langle \tilde{v}_2(t)\tilde{v}_2^*(t)\rangle]^{1/2}} = \frac{\langle \tilde{v}_1(t+\tau)\tilde{v}_2^*(t)\rangle}{(I_1 I_2)^{1/2}} \tag{2.17}$$

The amplitude $|\tilde{\gamma}_{12}(\tau)|$ of complex coherence that describes a light in terms of coherency. If $|\tilde{\gamma}_{12}(\tau)| = 1$ then the light is considered as a coherent one, if $|\tilde{\gamma}_{12}(\tau)| = 0$ then the light is incoherent. For other values between these two extremes, the light is said to be partially coherent.

According to the configuration of secondary point sources $\mathbf{p}_1$ and $\mathbf{p}_2$ and their distance the source $\mathbf{s}$ it is possible to distinguish between two cases of coherence: a spatial and a temporal coherence. Both meanings for the coherence explores conditions for which the interference pattern, i.e. fringes, becomes invisible and thus useless for the purposes of the holography.

If two ideal and coherent light sources of intensities I_1 and I_2 forms an interference patterns of intensity I then the visibility $\mathcal{V}$ of such pattern is [Har96]:

$$\mathcal{V} = \frac{2(I_1 I_2)^{1/2}}{I_1 + I_2}\cos(\psi), \tag{2.18}$$

where ψ is an angle between electrical vectors of both light waves and thus represents polarization[1].

For partially coherent light sources of same intensity, i.e. $I_1 = I_2$, the visibility of the interference pattern is [Har96]:

$$\mathcal{V} = |\tilde{\gamma}_{12}(\tau)|. \tag{2.19}$$

The **t**emporal coherence becomes important for a very small quasi-monochromatic light sources. For such light sources, the complex coherence depends on a difference in transit times between each of secondary sources $\mathbf{p}_1$ and $\mathbf{p}_2$ and the primary source $\mathbf{s}$. Thus, it is, in fact, a normalised autocorrelation of the function $\tilde{v}(t)$. If requirements for the Equation (2.19) are fulfilled, the amount of coherence can be determined from the visibility $\mathcal{V}$ of fringes. According to [Har96], for a radiation with a mean frequency ν_0 and bandwidth Δ_ν the visibility $\mathcal{V}$ drops to zero if difference in transit times Δ_τ fulfill following condition:

$$\Delta_\tau \Delta_\nu \approx 1, \tag{2.20}$$

where $\omega = 2\pi\nu$. The time Δ_τ is denoted as a **c**oherence time of given radiation. Another forms of this property is a **c**oherence length. If the optical path difference is smaller than the coherence length the interference pattern is visible. The coherence length Δ_l for a radiation of mean wavelength λ_0 and wavelength bandwidth Δ_λ is:

$$\Delta_l \approx c\Delta_\tau \approx c/\Delta_\nu \approx \lambda_0^2/\Delta_\lambda. \tag{2.21}$$

Spatial coherence becomes an important feature of the radiation as the difference of optical paths $\mathbf{sp}_1$ and $\mathbf{sp}_2$ is small enough for the time difference to be $\tau \approx 0$. Spatial coherence relates a range between two points and visibility of the interference pattern. If two slits $\mathbf{p}_1$ and $\mathbf{p}_2$ are separated by a distance greater than the diameter of the **c**oherence area then waves generated by these two slit do not form visible interference pattern.

[1] Note, that the visibility drops to zero if $\psi = \pi/2$, i.e. waves of polarized light do not create a visible interference pattern if polarization directions are perpendicular to each other.

It can be shown that for a extended distant source, i.e. source composed of many point sources, the visibility $\mathcal{V}$ is proportional to an absolute value of the sinc function [Wei]. The argument of the sinc function is proportional to a multiply of distance a between slits and an angle η that is a range in which the individual point sources subtends the screen. If either a or η increases then the fringe visibility decreases. Note, that for an angle $\eta = 0$ that is valid for a point source the sinc function does not depend on distance a.

The spatial complex coherence can be expressed in terms of Fourier transform as well. If the distance between two points $\mathbf{p}_1 = (0, 0, z)$ and $\mathbf{p}_2 = (x, y, z)$ is much smaller than the distance from these points to a source $\mathbf{s}$ then the complex coherence of the field follows [Har96]:

$$\tilde{\mu}_{12} = \frac{\exp(\mathrm{i}\phi_{12}) \iint_S I(\xi, \eta) \exp[\mathrm{i}k(x\xi + y\eta)] \,\mathrm{d}\xi \,\mathrm{d}\eta}{\iint_S I(\xi, \eta) \,\mathrm{d}\xi \,\mathrm{d}\eta}, \tag{2.22}$$

where $\xi = x_S/z$, $\eta = y_S/z$, plane S is a XY-plane that contains the source and $\phi_{12} = -k(x^2 + y^2)/2z$.

An important side effect of the coherence is that if the light is coherent in both spatially and temporally, it is possible to omit the temporal component $-\mathrm{i}\omega t$ of the wavefunction defined by the Equation (2.9) and leave only the wave distribution to be examined or computed. This can be interpreted as exploration of the wave distribution for an infinitely small time period. As the intensity I that serves as the physically measurable property of the light depends only on the complex amplitude $\tilde{u}(\mathbf{p})$ for the monochromatic light, no unacceptable approximation is applied by that. Thus, in the following text the complex amplitude serves as a full description of the monochromatic wave distribution, if not noted otherwise.

2.4 Elementary Waves

Elementary waves represents the simplest solution for the Helmholtz equation. There are two forms of these waves: a planar wave and a spherical wave. The **p**lane wave is a wave with wavefronts that are infinite planes. The complex amplitude of such wave is [Gra03, Kra04]:

$$\begin{aligned} \tilde{u}(\mathbf{r}) &= \tilde{a} \exp(\mathrm{i}\mathbf{k} \cdot \mathbf{r}) \\ &= \tilde{a} \exp[\mathrm{i}(k_x x + k_y y + k_z z)], \end{aligned} \tag{2.23}$$

where $\mathbf{k} = (k_x, k_y, k_z)$ is a **w**avevector and $\tilde{a}$ is a complex envelope that defines the phase and the amplitude at the origin of the wave. The vector $\mathbf{r} = (x, y, z)$ points from the origin to a sample for which a distribution is obtained. The length of the wavevector is equal to the wavenumber. Intensity of the wave is constant and it is equal to $I = |\tilde{a}|^2$. Wavefunction of the planar wave is then following:

$$u(\mathbf{r}, t) = |\tilde{a}| \cos(\arg\{\tilde{a}\} + \mathbf{k} \cdot \mathbf{r} - \omega t). \tag{2.24}$$

The **s**pherical wave is a wave where wavefronts have a form of concentric spherical surfaces centered at the point source. The complex amplitude of the spherical wave with an origin identical to the source of the field is:

$$\tilde{u}(\mathbf{r}) = \frac{\tilde{a}}{r} \exp(\mathrm{i}kr), \tag{2.25}$$

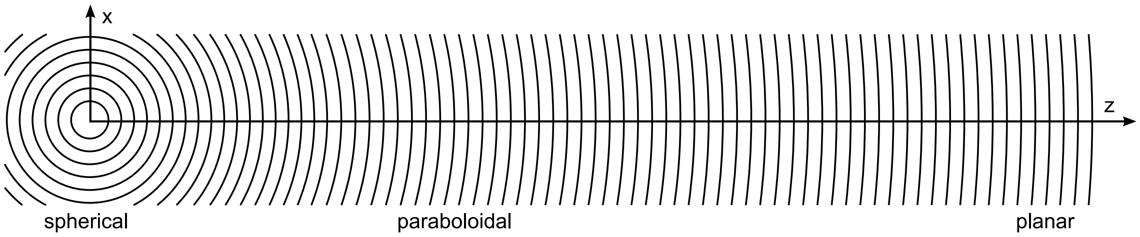

Figure 2.3: Relation between spherical and planar wave [Kra04].

where $r = |\mathbf{r}|$, i.e. it is a distance from the source. The fraction $\tilde{a}/r$ compensates that fact that the surface of the spherical surface grows quadratically. If the intensity, i.e. $|\tilde{a}|^2$, was not modified at all then it means that the intensity per unit grows quadratically as well. Yet, a wave cannot increase its energy on its own. Thus, the complex amplitude $\tilde{a}$ has to be modified by $1/r$ to compensate the quadratical growth of the surface[2] in the Equation (2.25).

The relation between the planar and the spherical wave is more apparent when the wave propagates further from the point source. If an observation is done along the Z-axis through a window which is constant in size spherical wavefronts becomes a planar as it is depicted in the Figure 2.3. This means that if the distance is large enough in comparison to extents in X-axis and Y-axis it is possible to approximate the spherical wave with the paraboloidal wave. This is a mechanism utilized by the Fresnel approximation, see below. If the distance increases even further it is possible to approximate the spherical wave with the planar one.

2.5 Diffraction

The Diffraction is basically the same phenomenon as the interference. The difference is that the interference is referenced in a case of superposition of several light sources and diffraction is referenced in a case of superposition of many sources. In the case of holography, the interference is usually addressed when interference of the scene light field and the reference beam is evaluated and the diffraction is addressed when light field of a scene is evaluated.

The nature of the diffraction can be illustrated on the well known Huygens principle proposed by C. Huygens. This principle states that the wavefront of a disturbance in a time $t + \Delta t$ is an envelope of wavefronts of a secondary sources emanating from each point of the wavefront in a time t, see Figure 2.4 for reference. This principle was modified by A. Fresnel who stated that the secondary sources interfere with each other and the amplitude at each point of the wavefront is obtained as superposition of the amplitudes of all the secondary wavelets. This Huygens-Fresnel principle matches many optical phenomena and it was also shown by G. Kirchhoff how this principle can be deduced from Maxwell's equations.

Although this principle works in many cases its validity is in question. For example, it does not determine the direction of the wavefront propagation. It is only an intuitive choice that the wavefront diverges from the source and not converges back to the source depending on the chosen orientation of the envelope of the secondary wavelets. In this

[2] Note, that the intensity grows quadratically with the magnitude of the complex amplitude

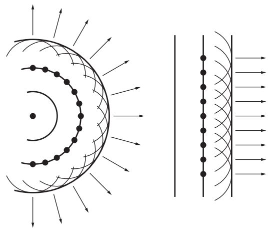

Figure 2.4: [v0.5]Huygens principle demonstrated on spherical (left) and planar (right) waves.

work, as in many others, this principle is accepted as an appropriate description of the wave behavior of the light and its inadequacies are neglected. Some of the synthesis methods are based on this principle.

The intuitive way of the diffraction understanding is covered by the Huygens principle but more formal descriptions also exists. Some of them are presented in the following material.

The mathematical description of the diffraction is quite difficult. It's due to the vectorial nature of the problem and many propagation medium properties like linearity, isotropy, homogeneity or dispersiveness that increase dimension of the problem. The basic diffraction model assumes an ideal material that is linear, isotropic, homogeneous, nondispersive and nonmagnetic. Under these conditions, the electromagnetic wave behavior can be described using only one scalar equation that described behavior of both magnetic and electric field. There is one other condition that further simplifies the diffraction model: diffraction structures that are large compared to the wavelength of the diffracted wave. All those simplifications and constraints turn the diffraction model into approximation but even thought simplifications are significant they causes only small loss of accuracy and thus they are more then appropriate in many situations.

There are two fundamental description of the diffraction: the Kirchhoff formulation and the Rayleigh-Sommerfeld formulation [Goo05, LBL02]. Both formulations describe the field in front of a screen or an aperture properly and accurately. Nevertheless, the Kirchoff one has a certain limitation as it fails when the point closes to the screen for a distance lower than few wavelengths. Also, it assumes that the field behind the aperture is zero and this is in contradiction with physical experiments. Despite its limitations, the Kirchoff formulation is widely used in practice.

The **K**irchhoff formulation of diffraction is based on the integral theorem of **H**elmholtz and Kirchoff. It states that the field at any point can be expressed in terms of wave values on any closed surface surrounding that point [Goo05]. The theorem is an application of the Green's theorem and the Helmholtz equation. While the Helmholtz equation describe behavior of waves, the **G**reen's theorem defines a relation between two complex functions $\tilde{u}(\mathbf{p})$ and $\tilde{g}(\mathbf{p})$ of position, closed volume V in which the observation is performed, arbitrary boundary surface S that encloses V, and derivatives of functions $\tilde{u}$ and $\tilde{g}$ along inward normal $\mathbf{n}$ of boundary surface S:

$$-\iint_S \frac{\partial \tilde{u}}{\partial \mathbf{n}} \tilde{g} - \tilde{u} \frac{\partial \tilde{g}}{\partial \mathbf{n}} \, \mathrm{d}s = \iiint_V \tilde{g} \nabla^2 \tilde{u} - \tilde{u} \nabla^2 \tilde{g} \, \mathrm{d}v, \tag{2.26}$$

where functions $\tilde{u}(\mathbf{p})$ and $\tilde{g}(\mathbf{p})$ provides twice continuously differentiable scalar fields mappings between V and S. If both functions satisfy the Helmholtz equation then [LBL02]:

$$-\iint_S \frac{\partial \tilde{u}}{\partial \mathbf{n}}\tilde{g} - \tilde{u}\frac{\partial \tilde{g}}{\partial \mathbf{n}}\,\mathrm{d}s = 0 \tag{2.27}$$

The goal of the Kirchoff formulation is do find a field at a point $\mathbf{p}_0$. For such purpose it uses a boundary surface on a one side of the aperture. This boundary surface consist of two parts: a plane S_p close to the aperture including its transparent portion Σ and a spherical surface S_ϵ. Next, a set of boundary condition known as **K**irchhoff boundary conditions[3] that describes behavior of field $\tilde{u}$ in close neighborhood to the screen. The influence of the spherical surface vanishes as the radius of the spherical surface increases towards infinity [BW05]. By application of there condition the Equation (2.27) is simplified to:

$$\tilde{u}(\mathbf{p}_0) = \frac{1}{4\pi}\iint_\Sigma \left(\frac{\partial \tilde{u}}{\partial \mathbf{n}}\tilde{g} - \tilde{u}\frac{\partial \tilde{g}}{\partial \mathbf{n}}\right)\mathrm{d}s, \tag{2.28}$$

where Σ is a transparent portion of the screen, $\tilde{u}$ is a complex function describing the wave distribution and $\tilde{g}$ is a complex function, see below.

A further simplification of the Equation (2.28) is based on use of a proper Green's function instead of the function $\tilde{g}$ in the Equation (2.27). Such function that satisfies the Helmholtz equation is:

$$\tilde{g}(\mathbf{p}_1) = \frac{\exp(\mathrm{i}kr_{01})}{r_{01}},$$

where $\mathbf{r}_{01} = \mathbf{p}_0 - \mathbf{p}_1$ and $r_{01} = |\mathbf{r}_{01}|$. The derivation along normal can be approximated according to an assumption on distances between observation point $\mathbf{p}_0$ in enclosed volume and point $\mathbf{p}_1$ on the surface Σ. If $r_{01} \gg \lambda$ then:

$$\begin{aligned}\frac{\partial \tilde{g}}{\partial \mathbf{n}} &= \frac{\exp(\mathrm{i}kr_{01})}{r_{01}}\left(\mathrm{i}k - \frac{1}{r_{01}}\right)\cos(\mathbf{n}, \mathbf{r}_{01}) \\ &\approx \mathrm{i}k\frac{\exp(\mathrm{i}kr_{01})}{r_{01}}\cos(\mathbf{n}, \mathbf{r}_{01}).\end{aligned}$$

Also, it is assumed that the screen or the aperture is illuminated by a spherical wave emerging from the point $\mathbf{p}_2$. Hence, the field $\tilde{u}$ at the point $\mathbf{p}_1$ is:

$$\tilde{u}(\mathbf{p}_1) = \frac{\tilde{a}\exp(\mathrm{i}kr_{21})}{r_{21}},$$

where r_{21} is distance between $\mathbf{p}_1$ and $\mathbf{p}_2$.

Application of assumptions and substitutions described above leads to a form known as the **F**resnel-Kirchhoff diffraction formula:

$$\tilde{u}(\mathbf{p}_0) = \frac{\tilde{a}}{\mathrm{i}\lambda}\iint_\Sigma \frac{\exp[\mathrm{i}k(r_{21} + r_{01})]}{r_{21}r_{01}}\left[\frac{\cos(\mathbf{n}, \mathbf{r}_{01}) - \cos(\mathbf{n}, \mathbf{r}_{21})}{2}\right]\mathrm{d}s, \tag{2.29}$$

where $\tilde{a}$ is basic amplitude/phase of the spherical wave, $\mathbf{n}$ is a normal of transparent portion Σ of the planar screen, and $\cos(\mathbf{a}, \mathbf{b})$ is a cosine of angle between vectors $\mathbf{a}$ and $\mathbf{b}$.

[3] The first assumption is that the distribution of the field $\tilde{u}$ across the surface Σ including its derivate along normal $\mathbf{n}$ is no different from the same configuration without the screen. The second assumption is that a portion of the surface close to the screen $S_p - \Sigma$ lies in a geometrical shadow an thus the function $\tilde{u}$ as well as its derivate along the normal is zero. Both assumption are not physically valid as they are never fulfilled completely but they simplifies the equation by replacing of enclosing surface just with the surface Σ. For more details, refer to [Goo05].

The vectors $\mathbf{r}_{01}$ and $\mathbf{r}_{21}$ are a vectors of lengths r_{01} and r_{21} between an observation point $\mathbf{p}_0$, point $\mathbf{p}_1$ on surface Σ, and source of the spherical wave $\mathbf{p}_2$.

A more practical form of the Fresnel-Kirchhoff formula can be obtained with a proper reorganization and substitution:

$$\tilde{u}(\mathbf{p}_0) = \frac{\tilde{a}}{\mathrm{i}\lambda} \iint_{\Sigma} \tilde{u}'(\mathbf{p}_1) \frac{\exp(\mathrm{i}kr_{01})}{r_{01}} \,\mathrm{d}s. \tag{2.30}$$

The interpretation of this equation is that the field at the point $\mathbf{p}_0$ is a superposition of infinite number of point sources on the surface Σ with a given complex amplitude $\tilde{u}'$. This is a consequence of the wave nature of the light and plays an important role in numerical reconstruction of the hologram, see below. For more details on the Kirchhoff formulation refer to [Goo05].

The Sommerfeld-Rayleigh formulation is a further enhancement of the Kirchhoff one that removes inconsistencies mentioned above. It removes the boundary condition from the function $\tilde{u}$ by assuming that either $\tilde{g}$ or $\partial\tilde{g}/\partial\mathbf{n}$ in the Equation (2.28) vanishes on a portion of the boundary surface close to the aperture according to a proper definition of alternate Green's function, see below. Unlike the Kirchhoff solution, it assumes that the screen is planar.

In order to fulfill this assumption it uses a second point $\mathbf{p}_0{}'$ that is mirror image of the point $\mathbf{p}_0$. At that second point a point source of the same wavelength as the first one is positioned. Both wave sources are generated with π phase difference. Note, that the behavior required is also fulfilled if both wave sources are oscillating in phase, i.e. with zero angle difference. Yet, for the π phase difference, a formula that describes the field commonly known as the first Rayleigh-Sommerfeld solution is:

$$\tilde{u}(\mathbf{p}_0) = \frac{-1}{4\pi} \iint_{\Sigma} \tilde{u} \frac{\partial \tilde{g}_-}{\partial \mathbf{n}} \,\mathrm{d}s, \tag{2.31}$$

where $\tilde{g}_- = [\exp(\mathrm{i}jkr_{01})/r_{01}] - [\exp(\mathrm{i}jkr'_{01})/r'_{01}]$, i.e. it is a field constructed as a difference of fields generated by source at $\mathbf{p}_0$ and its mirror at $\mathbf{p}_0{}'$. Note, that function $\tilde{g}_-$ vanishes on the transparent portion of planar screen, i.e. surface Σ.

By applying a similar set of assumption on distances similar to that of the Kirchoff formulation, i.e. $|\mathbf{p}_0 - \mathbf{p}'_0| \gg \lambda$, it is possible to approximate normal derivate of the function $\tilde{g}$ by:

$$\begin{aligned} \frac{\partial \tilde{g}}{\partial \mathbf{n}} &= \frac{\exp(\mathrm{i}k|\mathbf{p}_0 - \mathbf{p}'_0|)}{|\mathbf{p}_0 - \mathbf{p}'_0|} \left(\mathrm{i}k - \frac{1}{|\mathbf{p}_0 - \mathbf{p}'_0|}\right) \alpha(\mathbf{n}, \mathbf{p}_0 - \mathbf{p}'_0) \\ &\approx \mathrm{i}k \frac{\exp(\mathrm{i}k|\mathbf{p}_0 - \mathbf{p}'_0|)}{|\mathbf{p}_0 - \mathbf{p}'_0|} \,\alpha(\mathbf{n}, \mathbf{p}_0 - \mathbf{p}'_0), \end{aligned}$$

where $\alpha(\mathbf{a}, \mathbf{b}) = (\mathbf{a} \cdot \mathbf{b})/(|\mathbf{a}||\mathbf{b}|)$ is a cosine of angle between vectors $\mathbf{a}$ and $\mathbf{b}$. By application of this approximation to an alternate Green's function $\tilde{g}_-$ and by the fact that $\tilde{g}_-$ vanishes on transparent portion Σ of planar screen it is possible to obtain two formulas known as the **R**ayleigh-Sommerfeld diffraction formula. For more details on derivation of these formulas, refer to [Goo05, LBL02, Mie02]. The first configuration that has a π phase difference in phases is:

$$\tilde{u}_I(\mathbf{p}_0) = \frac{\tilde{a}}{\mathrm{i}\lambda} \iint_{\Sigma} \frac{\exp[\mathrm{i}k(r_{21} + r_{01})]}{r_{21} r_{01}} \cos(\mathbf{n}, \mathbf{r}_{01}) \,\mathrm{d}s. \tag{2.32}$$

The second configuration that has zero phase difference in phases is:

$$\tilde{u}_{II}(\mathbf{p}_0) = -\frac{\tilde{a}}{\mathrm{i}\lambda} \iint_{\Sigma} \frac{\exp[\mathrm{i}k(r_{21} + r_{01})]}{r_{21} r_{01}} \cos(\mathbf{n}, \mathbf{r}_{21}) \, \mathrm{d}s. \tag{2.33}$$

Note, the both the first and the second RayLeigh-Sommerfeld resembles the Kirchhoff-Fresnel diffraction formula with the difference in sign and the last cosine-based component. It can be shown that the Kirchoff solution is an average of both the first and the second Rayleigh-Sommerfeld solution. Kirchoff and Rayleigh-Sommerfeld solutions are almost identical for small angles and larger distance but they differs for distances closer to the aperture. For more detail on comparison and discussion on consequences beyond scope of this work, refer to [Goo05].

2.6 Wave Propagation

The propagation of the wave in a free space plays an important role in the hologram reconstruction as it is required for presenting of optical field or diffraction pattern generated by a screen and/or aperture to the viewer. The propagation is driven by the Helmholtz equation and the Huygens-Fresnel principle.

Basically, the propagation of the wave is reflected in a change in the phase. Yet, it is a question whether the changes should be introduces by adding or subtracting a value from the phase. As the time dependent component $\exp(-\mathrm{i}\omega t)$ of the wavefunction in the Equation (2.9) rotates in a clockwise direction waves emitted earlier in time have phase greater than waves emitted later. The later the wave is emitted the closer it is to its source. Thus, if waves are to be propagated from the source then the phase has to be increased.

In order to minimize the confusion from signs of phases, it is assumed that the propagation of the wave is examined in direction of a positive Z-axis. This also simplifies equation for propagation in a direction parallel to the Z-axis. If other configuration is required then it is always possible to transform the scene to fit the requirement.

2.6.1 Huygens-Fresnel Principle

The Huygens-Fresnel principle[4] states that every point on a primary wavefront is a source of spherical waves with the same optical frequency and the primary wave is. The resulting field is a superposition of these secondary waves defined by their complex amplitude and/or wavefunction [Wei].

The Huygens-Fresnel principle is confirmed by both Kirchoff and Rayleigh-Sommerfeld diffraction formulas. Using the first Rayleigh-Sommerfeld solution from the Equation (2.32), the Huygens-Fresnel principle is [LBL02]:

$$\tilde{u}_I(\mathbf{p}_0) = \frac{1}{\mathrm{i}\lambda} \iint_{\Sigma} \tilde{u}(\mathbf{p}_1) \frac{\exp(\mathrm{i}kr_{01})}{r_{01}} \cos\theta \, \mathrm{d}s, \tag{2.34}$$

[4] Original Huygens principle describes a new wavefront as an envelope of spherical wave sources generated on a surface of the previous wavefront. Yet, this is not physically valid as it may ignore obstacles in wave propagation and it allows a creating of back-waves.

where $\tilde{u}(\mathbf{p}_1)$ represents a secondary point source positioned at the point $\mathbf{p}_1$ within the aperture Σ. The complex amplitudes of the secondary sources are proportional to the amplitude at the point $\mathbf{p}_1$ of the original wave and the phase leads the phase of the original wave by $\pi/2$ due to factor $1/\mathrm{i}$. Note, that due to the Rayleigh-Sommerfeld solution, the aperture Σ is expected to be a plane or its portion and thus all following solutions for the wave propagation solve the problem of propagating between two parallel planes, if not noted otherwise.

The important aspect of the Equation (2.34) is that it is basically a convolution integral as it can be expressed as follows:

$$\tilde{u}(\mathbf{p}_0) = \iint_{\Sigma} \tilde{h}(\mathbf{p}_0, \mathbf{p}_1)\tilde{u}(\mathbf{p}_1)\,\mathrm{d}s, \tag{2.35}$$

where $\tilde{h}(\mathbf{p}_0, \mathbf{p}_1)$ is the impulse response function that is given explicitly by:

$$\tilde{h}(\mathbf{p}_0, \mathbf{p}_1) = \frac{1}{\mathrm{i}\lambda}\frac{\exp(\mathrm{i}kr_{01})}{r_{01}}\cos\theta.$$

2.6.2 Fresnel and Fraunhofer Approximation

As noted in previous subsections, it is possible to express the Huygens-Fresnel principle in terms of the first Rayleigh-Sommerfeld solution. The angle θ is an angle between the outward normal $\mathbf{n}$ and the vector $\mathbf{r}_{01}$. The cosine of this angle can be also expressed as following:

$$\cos\theta = \frac{z}{r_{01}},$$

and if a configuration is that of the Figure 2.5 it is possible to express the Huygens-Fresnel principle as following:

$$\tilde{u}(x,y) = \frac{z}{\mathrm{i}\lambda}\iint_{\Sigma} \tilde{u}(\xi,\eta)\frac{\exp(\mathrm{i}kr_{01})}{r_{01}^2}\,\mathrm{d}\xi\,\mathrm{d}\eta, \tag{2.36}$$

where $r_{01} = [z^2 + (x-\xi)^2 + (y-\eta)^2]^{1/2}$. Besides that, if the distance z is greater than the extent in the X-axis and Y-axis then $cos\theta \approx 1$ and thus the whole cosine term can be omitted completely. Nevertheless, at the end both approaches lead to the same equations. All following approximations simplifies the expression for r_{01} because it contains square root function that does not allows use of the Fourier transform that greatly reduces the computation complexity of the expression.

The first approximation is known as the **F**resnel approximation. It is based on a value of the Z-axis component of $\mathbf{r}_{01}$ that, if large enough, allows application of binomial expansion for the square root function. In order to apply the assumption, first, a reordering of the expression for the distance r_{01} has to be applied:

$$r_{01} = z\left[1 + \left(\frac{x-\xi}{z}\right)^2 + \left(\frac{y-\eta}{z}\right)^2\right]^{1/2}. \tag{2.37}$$

Such expression now resembles a an expression $\sqrt{1+b}$, where $|b| < 1$. This expression can be decomposed by use of a binomial expansion to a form:

$$\sqrt{1+b} = 1 + \frac{1}{2}b - \frac{1}{8}b^2 + \ldots. \tag{2.38}$$

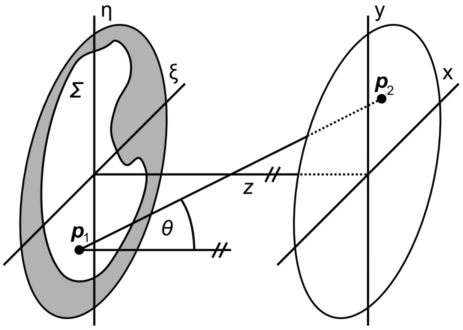

Figure 2.5: Configuration of the Fresnel/Fraunhofer approximation [Goo05].

If the extent in Z-axis is greater than the extent in both X-axis and Y-axis, i.e. if it is valid that $z \gg (x-\xi)^2 + (y-\eta)^2$, then it is possible to apply the approximation by dropping the third component of the binomial expansion in the Equation (2.38). In order to avoid the degradation of the result, the third term shall not cause a change in the distance greater than 1° if applied to the phase in the Equation (2.36) because the wave propagation is sensitive to the phase [MNF+02]. This is fulfilled if the following condition is valid:

$$z^3 \gg \frac{k}{8}[(x-\xi)^2 + (y-\eta)^2]^2_{\max}, \tag{2.39}$$

where $k = 2\pi/\lambda$ is a wavenumber. Note, that this condition requests a viewing distance of $z \gg 250$ mm for a circular aperture of diameter 10 mm is observed from a region of size 1 cm and the wavelength is $\lambda = 0.5$ μm. If z fulfills the condition the viewer is known to be in a **n**ear field or **F**resnel region.

The above mentioned condition is sufficient but it is also very strict. In fact it is overly strict. As it is shown in [Goo05] for a diffraction aperture illuminated with an uniform plane wave the Fresnel approximation is valid even for closed distances than distances forced by the Equation (2.39).

Nevertheless, by fulfilling the condition it is possible to apply the Equation (2.37) to the first two components of the binomial expansion according to the Equation (2.38) and obtain:

$$r_{01} \approx z + \frac{(x-\xi)^2 + (y-\eta)^2}{2z}. \tag{2.40}$$

Then, this approximation of the distance r_{01} is applied for the phase component in the Equation (2.36) as it is required to keep it as accurate as possible. As the distance in the phase is multiplied by a wavenumber $k \approx 10^7$ a small error in r_{01} may cause significant change in phase. The distance used in the denominator is approximated by much coarser approximation because it modifies only the amplitude. For such purpose $r_{01} \approx z$ provides reasonable amount of error. This leads to a resulting expression known as the **F**resnel diffraction integral:

$$\tilde{u}_z(x,y) = \frac{\exp(\mathrm{i}kz)}{\mathrm{i}\lambda z} \iint_{-\infty}^{\infty} \tilde{u}_0(\xi,\eta) \exp\left[\mathrm{i}k\frac{(x-\xi)^2 + (y-\eta)^2}{2z}\right] \mathrm{d}\xi\,\mathrm{d}\eta, \tag{2.41}$$

where $\tilde{u}_0(\xi,\eta)$ is a diffraction pattern at the aperture. This expression can be reorganized to a form that resembles the Fourier transform and thus it can be utilized for its

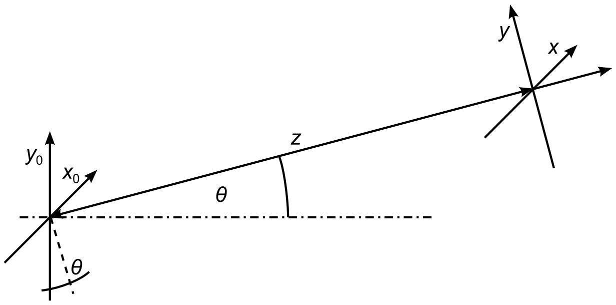

Figure 2.6: Tilted plane configuration for application of Fresnel approximation [YAC02].

computations reducing the computational complexity significantly at the same time:

$$\begin{aligned}\tilde{u}_z(x,y) &= \frac{\exp(\mathrm{i}kz)}{\mathrm{i}\lambda z}\exp\left(\mathrm{i}k\frac{x^2+y^2}{2z}\right)\\ &\times\iint_{-\infty}^{\infty}\left\{\tilde{u}_0(\xi,\eta)\exp\left(\mathrm{i}k\frac{\xi^2+\eta^2}{2z}\right)\right\}\exp\left(-\mathrm{i}k\frac{x\xi+y\eta}{z}\right)\,\mathrm{d}\xi\,\mathrm{d}\eta. \quad (2.42)\end{aligned}$$

As the distance in the Z-axis increases it is possible to widen the Fresnel approximation and omit another components of the integral. Such modification is known as the **F**raunhofer approximation and it is applicable only if:

$$z \gg \frac{k(\xi^2+\eta^2)_{\max}}{2}.$$

This condition has even higher demands on the Z-axis component value than the condition in the Equation (2.39). For a circular aperture with a diameter of 10mm and for a wave with wavelength 0.5μm the distance along the Z-axis has to be $z \gg 300$m. If z fulfills the condition the viewer is denoted to be in **f**ar field or **F**raunhofer region. By satisfaction of the condition the Equation (2.42) is reduced to the following form:

$$\tilde{u}_z(x,y) = \frac{\exp(\mathrm{i}kz)}{\mathrm{i}\lambda z}\exp\left(\mathrm{i}k\frac{x^2+y^2}{2z}\right)\times\iint_{-\infty}^{\infty}\tilde{u}_0(\xi,\eta)\exp\left(-j2\pi\frac{x\xi+y\eta}{\lambda z}\right)\,\mathrm{d}\xi\,\mathrm{d}\eta. \quad (2.43)$$

This form resembles a Fourier transformation of the aperture distribution $\tilde{u}_0$. Note, the if a normalized intensity is the required result then the Fraunhofer approximation leads to a Fourier transform of $\tilde{u}_0$. In that case, multiplicative phase factors are not applied as they have no influence on the intensity, see Equation (2.12). Given elements of the result distribution can be extracted on corresponding frequencies:

$$\begin{aligned}f_X &= x/\lambda z,\\ f_Y &= y/\lambda z.\end{aligned}$$

The Fresnel approximation explores a propagation of diffraction pattern between two planes along the Z-axis where both planes are parallel and their origins lie on the Z-axis. Yet, it is possible to enhance the Fresnel approximation formula so it can handle tilted planes such as that depicted in the Figure 2.6 as well [YAC02].

The approach is based on simplification of the expression for the distance. If the target diffraction pattern $\tilde{u}_{z,\theta}(x,y)$ is examined on a plane that is tilted as depicted in the Figure 2.6 then the expression for the distance required for the Rayleigh-Sommerfeld integral is:

$$r = \left[(y_0 \sin\theta - z)^2 + (x_0 - x)^2 + (y_0 \cos\theta - y)^2\right]^{1/2}. \tag{2.44}$$

By introduction of $r' = (x^2 + y^2 + z^2)^{1/2}$ to the Equation (2.44) a binomial expansion can be applied. After the expansion, it is possible to omit all terms but first two similar to the Fresnel approximation. The resulting expression is substituted to a modified Rayleigh-Sommerfeld integral:

$$\tilde{u}_{z,\theta}(x,y) = -\frac{\tilde{a}}{\mathrm{i}\lambda} \iint_{\Sigma} \tilde{u}_0(x_0,y_0) \frac{\exp(\mathrm{i}kr)}{r} \chi(x_0,y_0,x,y)\,\mathrm{d}x_0\,\mathrm{d}y_0,$$

where $\tilde{u}_0(x_0,y_0)$ is a source of the diffraction pattern and $\chi(x_0,y_0,x,y)$ is an inclination factor that is close to 1 if condition for the Fresnel approximation is valid. After a reorganization and further substitution a form is obtained that resembles the Fourier transform:

$$\tilde{u}_{z,\theta}(\xi,\eta) = \exp(\mathrm{i}kr') \iint_{\Sigma} \tilde{u}_0(x_0,y_0) \exp\left(\mathrm{i}k\frac{x_0^2+y_0^2}{2z}\right) \exp\left[-\mathrm{i}2\pi(\xi x_0 + \eta y_0)\right]\,\mathrm{d}x_0\,\mathrm{d}y_0,$$

where $\xi = x/(\lambda r')$ and $\eta = (y\cos\theta + z\sin\theta)/(\lambda r')$. Note, that the result obtained result is respective to a plane deformed by coordinates ξ and η and it assumes that the condition for the Fresnel approximation is valid.

2.6.3 Diffraction Condition and Diffraction Orders

A **d**iffraction condition specifies a result of a diffraction of a plane wave on a thin cosine grating [Goo05, Kra04]. The **c**osine grating is a thin cosine amplitude grating on a plane with a amplitude transmittance function:

$$t_A(\xi,\eta) = \exp\left[\frac{1}{2} + \frac{m}{2}\cos\left(2\pi\frac{\xi}{\Lambda_\xi}\right)\right] \mathrm{rect}\left(\frac{\xi}{2w}\right) \mathrm{rect}\left(\frac{\eta}{2w}\right), \tag{2.45}$$

where ξ and η are coordinates on a grating, $2w$ is the width/height of rectangular aperture, m represents a difference between maximum and minimum of the t_A, and Λ_ξ is the grating period.

If such grating is illuminated by unit amplitude plane wave then a diffraction occurs. By an application of the convolution theorem to t_A it is possible to obtain the Fourier transform of t_A that can be utilized to build the Fraunhofer diffraction pattern, i.e. a diffraction pattern in the far field:

$$\begin{aligned} \tilde{u}(x,y) =& \frac{a}{\mathrm{i}2\lambda z} \exp\left[\mathrm{i}kz + \mathrm{i}\frac{k}{2z}(x^2+y^2)\right] \mathrm{sinc}\left(\frac{2wy}{\lambda z}\right) \\ &\times \left\{\mathrm{sinc}\left(\frac{2wx}{\lambda z}\right) + \frac{m}{2}\mathrm{sinc}\left[\frac{2w}{\lambda z}\left(x + \frac{\lambda z}{\Lambda_\xi}\right)\right] + \frac{m}{2}\mathrm{sinc}\left[\frac{2w}{\lambda z}\left(x - \frac{\lambda z}{\Lambda_\xi}\right)\right]\right\} \end{aligned} \tag{2.46}$$

The intensity of the diffraction pattern is a squared magnitude of $\tilde{u}$ from the Equation (2.46). This means that the intensity is a sum of squared sinc functions such that in the Figure 2.7.

Peaks in the figure are related to the term of **d**iffraction orders and they represents energy deflected by the aperture. The central peak is known as the zero order and represents the undiffracted, original plane waves. It also contains the greatest amount of energy

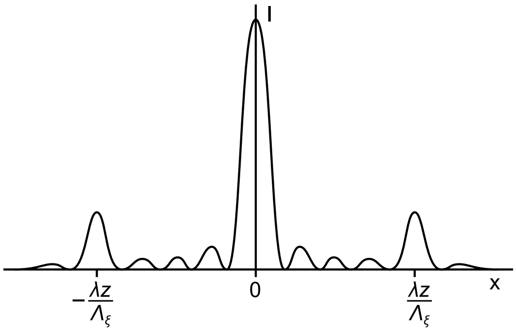

Figure 2.7: Intensity of Fraunhofer diffraction pattern for thin amplitude cosine grating [Goo05].

of the original undiffracted plane wave. The both side peaks are known as the first orders and represents original plane waves diffracted according to the diffraction condition, see below.

The portion of the energy that is divided into individual diffraction orders is $\frac{1}{2} + \frac{m}{2}\cos(2\pi\Lambda_\xi\xi)$ and can be found as the squared coefficient of the delta function from the Fourier transform of the amplitude modifier t_a in the Equation (2.45). Note, that while the sinc functions in Fourier transform of the t_A spreads the energy, the delta function determine the power in each order. The zero order obtains $1/4 = 25.0\,\%$ of the energy delivered by the incident wave, the maximum portion for the first order is $1/16 = 6.25\,\%$ of the incident power; the rest is absorbed by the grating or reflected. The percentage of the energy for the first order is known as the **d**iffraction efficiency of the grating.

A better diffraction efficiency of up to 33.8 % has the thin sinusoidal phase grating that employs complex amplitude transmittance instead of the real one in the Equation (2.45). This kind grating also leads to a formation of higher orders then just only the first and the zero ones. As the approach for derivation of diffraction order energy is similar to the amplitude grating it will not be exposed here. For more details, refer to [Goo05].

Basically, each diffraction order is the original illuminating plane wave with different portion of energy that is propagated to a different direction. The direction of the propagation for a given grating order is determined from the optical path difference of individual "rays". The difference can be obtained at integer multiplies of λ because only in such case a planar wavefront is formed. Thus, for a transmission grating, the incoming plane wave is diffracted according to:

$$\sin\theta_{\xi 2} = \sin\theta_{\xi 1} + q\frac{\lambda}{\Lambda_x}, \tag{2.47}$$

where $\theta_{\xi 1}$ is an angle of the incident wave, $\theta_{\xi 2}$ is an angle of the diffracted wave for given order **q**, and Λ_x is the period of the grating, see Figure 2.8.

2.6.4 Propagation in Angular Spectrum

The Fresnel approximation allows to compute a propagation of a wave but it has limitations on distance due to condition in the Equation (2.39). Even though this condition is unnecessarily strict and it is possible to apply the Fresnel approximation even for shorter distances, still it is not applicable for a region closer to a plane with known wave distri-

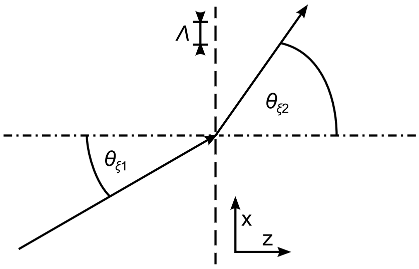

Figure 2.8: Diffraction grating and diffraction condition.

bution $\tilde{u}_0$ unless the third or higher components of the binomial expansion are taken into account.

For shorter distances the Rayleigh-Sommerfeld diffraction integral, see the Equation (2.32), offers a solution. Unfortunately, an application of this integral leads to a unpleasant high computation complexity and thus renders this solution almost unusable. Yet, a slightly different approach can be formulated if a Fourier transform of input wave distribution, which is also known as the **a**ngular spectrum, is considered [EO06, Goo05, TB93].

If a plane wave distribution $\tilde{F}(k_x, k_y)$ of plane waves is known then it is possible to determine a field $\tilde{u}$ at a given point $\mathbf{p}$ as:

$$\tilde{u}(\mathbf{p}) = \iint \tilde{F}(k_x, k_y) \exp(-\mathrm{i}\mathbf{k} \cdot \mathbf{p}) \,\mathrm{d}k_x \,\mathrm{d}k_y, \tag{2.48}$$

where $k_z = (k^2 - k_x^2 - k_y^2)^{1/2}$. It is assumed that $k_z^2 \geq 0$. If in any case k_z^2 becomes negative then k_z becomes a complex number. A wave with $k_z \in \mathbb{C}$ is known as the **e**vanescent wave [BW05]. This wave is propagated as well but its amplitude decays exponentially with increasing $|z|$, if it is propagated along the Z-axis[5].

For a $\varrho : z = 0$, the field is determined according to the following:

$$\tilde{u}_0(x, y) = \iint \tilde{F}(k_x, k_y) \exp[-\mathrm{i}(k_x x + k_y y)] \,\mathrm{d}k_x \,\mathrm{d}k_y. \tag{2.49}$$

It can be seen in the Equation (2.49) that the plane wave distribution $\tilde{F}$ is proportional to the Fourier transform of the distribution $\tilde{u}_0$ on the plane. More precisely, that $\tilde{F} = \mathcal{F}\left\{\tilde{u}_0\right\}/(2\pi)^2$.

By a knowledge of plane wave distribution it is possible to estimate a distribution on an arbitrary distance along the Z-axis. Waves are propagated according to the Equation (2.48). If the phase shift term is expanded properly then it is possible to obtain a form that resembles the Fourier transform as well[6]:

$$\tilde{u}(\mathbf{p}) = \iint \left\{\tilde{F}(k_x, k_y) \exp(-\mathrm{i}k_z z)\right\} \exp[-\mathrm{i}(k_x x + k_y y)] \,\mathrm{d}k_x \,\mathrm{d}k_y. \tag{2.50}$$

Then, computation of a field distribution $\tilde{u}$ on a plane that is parallel to the source plane at the distance z along positive Z-axis can be expressed as following:

$$\tilde{u} = \mathcal{F}^{-1}\left\{\mathcal{F}\left\{\tilde{u}_0\right\} \exp(-\mathrm{i}k_z z)\right\}, \tag{2.51}$$

[5] The exponential nature of attenuation is visible from substitution of complex-valued k_z to the Equation (2.48).

[6] Note, that individual wavevector components contains $2\pi/\lambda$ as the length of the wavevector is the wavenumber.

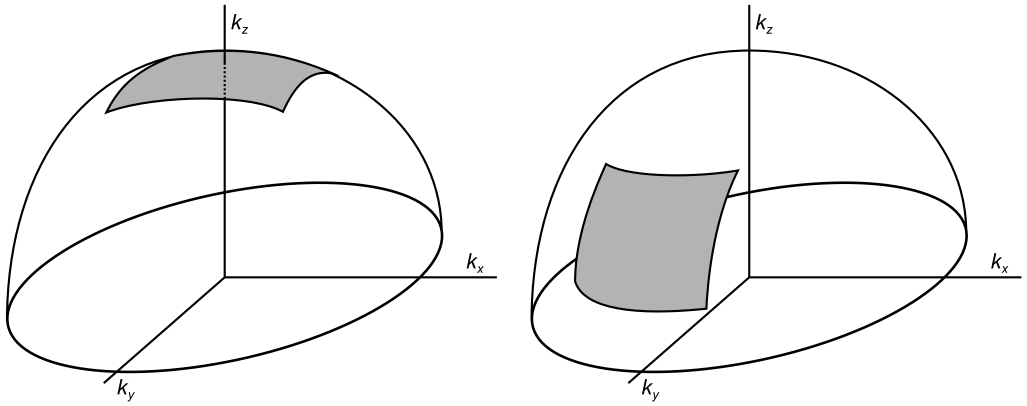

Figure 2.9: Original (left) and transformed (right) distribution of plane waves obtained by Fourier transform.

where $\tilde{u}$ is a wave distribution on a target plane while $\tilde{u}_0$ is wave distribution on a source plane.

This approach can be further modified to handle spatial shifting in the XY-plane and tilting as well. This is achieved by an application of simple geometrical operation of rotation stored in a form of a 3 × 3 matrix **R** and a translation in a form of a vector **b**:

$$\mathbf{p}' = \mathbf{p}\mathbf{R} + \mathbf{b}. \tag{2.52}$$

By substitution of the expression for a point **p** on a target plane to the Equation (2.48) and reorganization of the result, an expression for distribution $\tilde{u}$ on a target plane is:

$$\tilde{u} = \frac{1}{4\pi^2}\mathcal{F}^{-1}\left\{4\pi^2\mathcal{F}\left\{\tilde{u}_0\right\}\exp[\mathrm{i}(\mathbf{k}\mathbf{R})\cdot\mathbf{b}]J(k_z, k_z')\right\}, \tag{2.53}$$

where k_z' is a Z-axis component of the wavevector **k** transformed by the matrix **R** and function $J(k_z, k_z') = k_z/k_z'$ is a Jacobian correction factor. The transformation of the wavevector **k** by the matrix **R** is equal to a shifting of a portion of hemispherical surface over the hemispherical surface because all possible wavevectors excluding wavevector for evanescent waves forms a hemisphere of diameter equal to a wavenumber, see Figure 2.9.

The drawback of the approach is that it is sensitive to overlapping of both target and source plane. Due to the assumption on periodic nature of functions processed by the Fourier transform a disturbance appears if the target and source plane do not overlap each other after an orthogonal projection along the Z-axis. Yet, this can be avoided by combining of a proper propagation as mentioned in [TB93], so that miss in overlap does not occur at all.

2.6.5 Propagation in Lenses

Wave that propagates through an optically dense material or a material with different refractive index is slowed down, i.e. a wave propagating through such material is delayed. A lens is a optically dense material of a certain geometry. For purposes of simplicity, only a thin lens are considered. For the ray-based optics a **t**hin lens is a lens that causes a ray to exit the lens at a given point that is approximately the same as the point of entry. This means that in the case of a wave the thin lens causes only a slowdown of portions of incoming wavefronts. The slowdown is demonstrate itself as a change in phase.

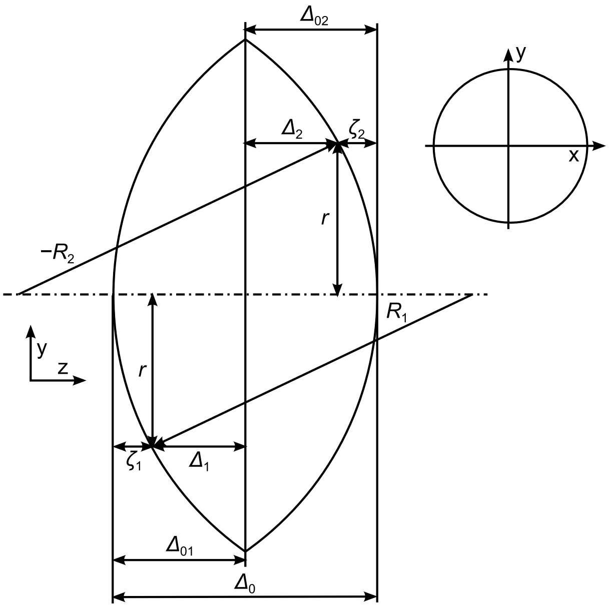

Figure 2.10: Lens and its coordinate system. Note, that the radius R_2 of the right-hand surface is negative as the ray is assumed to travel from left to right [Goo05, SJ05].

The amount of the slow down can be expressed in a form of a multiplicative phase factor [Goo05, SJ05]:

$$\begin{aligned} \tilde{t}_l(x,y) &= \exp[\mathrm{i}kn\Delta(x,y)]\exp\{\mathrm{i}k[\Delta_0 - \Delta(x,y)]\} \\ &= \exp[\mathrm{i}k\Delta_0]\exp[\mathrm{i}k(n-1)\Delta(x,y)], \end{aligned} \tag{2.54}$$

where $\Delta(x,y)$ is a lens thickness function, Δ_0 is maximum lens thickness, and n is a refraction index of the lens. Attenuation due to reflection and loss inside the lens is omitted. Note, that the factor causes a change in the phase that is equal to a sum of the phase changed due to the lens itself and phase change due to travel outside the lens, see Figure 2.10. The application of the factor on incident field $\tilde{u}_l$ leads to a field $\tilde{u}'_l$ that represents a field immediately after the lens, i.e. $\tilde{u}'_l(x,y) = \tilde{t}_l(x,y)\tilde{u}_l(x,y)$.

The most important component of the Equation (2.54) is the thickness function $\Delta(x,y)$. This function is a sum of thicknesses Δ_1, Δ_2 of both curved parts and thickness Δ_3 of the lens middle part:

$$\begin{aligned} \Delta(x,y) &= \Delta_1(x,y) + \Delta_2(x,y) + \Delta_3 \\ &= (\Delta_{01} - \zeta_1) + (\Delta_{02} - \zeta_2) + \Delta_3. \end{aligned} \tag{2.55}$$

As it is shown in the Figure 2.10, the thickness modification for the left-hand side of the lens is $\zeta_1 = R_1 - (R_1^2 - r^2)^{1/2}$; the right-hand side is similar. Thus, according to the Equation (2.55) the thickness function is:

$$\Delta(x,y) = \Delta_0 - R_1\left[1 - \left(1 - \frac{x^2+y^2}{R_1^2}\right)^{1/2}\right] + R_2\left[1 - \left(1 - \frac{x^2+y^2}{R_2^2}\right)^{1/2}\right], \tag{2.56}$$

where $\Delta_0 = \Delta_{01} + \Delta_{02} + \Delta_3$. The thickness function can be substituted to the Equation (2.54) in order to form an expression for a phase delay.

Yet, the equation for the thickness is still complicated for practical use. Nevertheless, if the extent in both X- and Y-axis is sufficiently small then it is possible to consider only paraxial rays and apply an approximation similar to the Fresnel one:

$$\left(1 - \frac{x^2 + y^2}{R_1^2}\right)^{\frac{1}{2}} \approx 1 - \frac{x^2 + y^2}{2R_1^2}.$$

This approximation leads to a significant simplification of the Equation (2.56) to a form that is substituted to multiplicative phase factor $\tilde{t}_l$ defined in the Equation (2.54) and gives:

$$\tilde{t}_l(x, y) = \exp(\mathrm{i}kn\Delta_0) \exp\left[-\mathrm{i}k(n-1)\frac{x^2 + y^2}{2}\left(\frac{1}{R_1} - \frac{1}{R_2}\right)\right]. \tag{2.57}$$

In a praxis the first component $\exp(\mathrm{i}kn\Delta_0)$ is omitted as it is constant for the whole lens and thus it is equivalent to a constant phase shift. The second component can be further simplified by the **L**ens maker equation:

$$\frac{1}{f} \equiv (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right), \tag{2.58}$$

where f is the focal length. **F**ocal length is a point where all parallel rays intersects after passing a the lens. Analogically, in wave optics it is a place where incoming wavefront that were modified by the lens become a infinitely small point. This is obvious from the Equation (2.59) when applied to a normally incident unit-amplitude plane wave. The resulting expression is a form similar to the quadratic approximation of the spherical wave. The application of the Lens maker equation leads to a widely used expression for the phase transformation of the lens:

$$\tilde{t}_l(\xi, \eta) = \exp\left(-\mathrm{i}k\frac{\xi^2 + \eta^2}{2f}\right). \tag{2.59}$$

The lens described by the Equation (2.59) exhibits a certain aberration in a phase. This aberration case be neglected if the intensity is the desired output. Otherwise, a correction has to be applied. This phase correction can be determined from application of the Fresnel approximation and the Equation (2.59) to a field $\tilde{u}(\xi, \eta)$ at position $z = 2f$ in front of the lens. The field $\tilde{u}_v(x, y)$ at $z = 2f$ behind the lens has to be the original field with inverted coordinates [Wei], i.e. $x = -\xi$ and $y = -\eta$. A difference between fields $\tilde{u}$ and $\tilde{u}_v$ is eliminated by a correction factor [SJ05]:

$$\tilde{p}(x, y) = \exp\left(-\mathrm{i}k\frac{x^2 + y^2}{2f}\right). \tag{2.60}$$

The overall equation for a setup with a source plane $\tilde{a}(\xi, \eta)$ placed behind the thin lens of focal length f and propagated by use of a propagation kernel $\tilde{k}(\xi, \eta; x, y)$ is:

$$\tilde{u}(x, y) = \iint \tilde{a}(\xi, \eta)\tilde{l}(\xi, \eta)\tilde{k}(\xi, \eta; x, y)\,\mathrm{d}\xi\,\mathrm{d}\eta.$$

The side effect of the thin lens is its capability perform a Fourier transform in its back focal plane. Focal plane is a plane parallel to the lens at focal distance. The configuration that is capable of the Fourier transformation is depicted in the Figure 2.11.

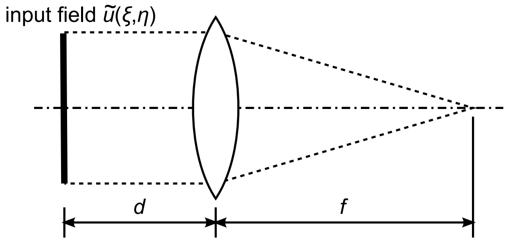

Figure 2.11: Geometry configuration for the Fourier transform of the input with positive lens.

As described by the Equation (2.50) a source angular spectrum is related to the target spectrum by a following phase factor:

$$\tilde{H}(f_X, f_Y) = \exp\left\{\mathrm{i}2\pi\frac{z}{\lambda}\left[1-(\lambda f_X)^2-(\lambda f_Y)^2\right]^{1/2}\right\}, \left(f_X^2+f_Y^2\right)^{1/2} < 1/\lambda.$$

If Fresnel or paraxial approximation is valid then above mentioned function can be simplified in similar manner as well[7]. The relation between angular spectrum of source $\tilde{F}_i = \mathcal{F}\{\tilde{u}_i\}$ and the angular spectrum of field in front of the lens $\tilde{F}_l = \mathcal{F}\{\tilde{u}_l\}$, i.e. source propagated to the lens, is following:

$$\tilde{F}_l(f_X, f_Y) = \tilde{F}_i(f_X, f_Y)\exp\left[-\mathrm{i}\pi\lambda d(f_X^2+f_Y^2)\right]. \tag{2.61}$$

The relation can be substituted to a Fresnel approximation applied to a propagation of the field $\tilde{u}_f(u,v)$ in the front focal plane to the lens, i.e. a propagation for a distance $z=f$ from the lens. Omitting the constant phase factor $\exp(\mathrm{i}kz)$ from the Equation (2.42) a following form is obtained:

$$\tilde{u}_f(u,v) = \frac{1}{\mathrm{i}\lambda f}\exp\left(\mathrm{i}k\frac{u^2+v^2}{2f}\right)\tilde{F}_l\left(\frac{u}{\lambda f},\frac{v}{\lambda f}\right). \tag{2.62}$$

Further substitution of the Equation (2.61) to the Equation (2.62) with $f_X = u/(\lambda f)$ and $f_Y = v/(\lambda f)$ leads to:

$$\tilde{u}_f(u,v) = \frac{1}{\mathrm{i}\lambda f}\exp\left[\mathrm{i}k\left(1-\frac{d}{f}\right)\frac{u^2+v^2}{2f}\right]\tilde{F}_i\left(\frac{u}{\lambda f},\frac{v}{\lambda f}\right). \tag{2.63}$$

It is clearly visible from the Equation (2.63) that the complex value of the field $\tilde{u}_f$ at coordinates (u,v) is related to a component at frequency $(u/\lambda f, v/\lambda f)$ of the input field angular spectrum. The Fourier transformation is disturbed by quadratic factor but this factor disappears if $d=f$. Note, that above mentioned equation is valid for the lens aperture with finite extent. For an aperture with limited extent, refer to [Goo05].

[7] The $\exp(\mathrm{i}kz)$ component of the Fresnel diffraction impulse response is omitted as it is constant for all plane waves.

Chapter 3

Optical Holography

Optical holography stands on the physical phenomenon of diffraction described in the section 2.5. The optical holography has one big advantage with respect to the digital holography which is that the diffraction is performed literally with the speed of light. It is so much not true for the numerical simulation of this phenomenon.

In this section, the principle of the optical holography is described. There are many ways of acquiring holograms so some most usual are presented. The mathematical principle of the reconstruction is provided.

3.1 Holography principle

Holography is about capturing and reproducing light filed. The light field is at each point determined by an amplitude and phase. In a case of classical photography the light field is integrated over some time and the adequate optical intensity is captured on a photographic material. When photograph is illuminated by a light source, the captured intensity is replayed into all directions. That is the reason, why reproduced image looks flat, without depth.

Holography, on the contrary, is a technique which records the phase and amplitude of the light field. When a hologram is illuminated by a proper light source, the exact amplitude and phase is reconstructed and the original light field recreated. Since the observer has the whole light field available, the genuine three dimensional sensation is achieved.

Holography uses almost the same materials for capturing as the photography and therefore the phase and amplitude cannot be recorded directly but rather in an encoded form, i.e. in a form of diffraction grating. The diffraction grating is formed by the fringes produced from interference of the reference beam and the scattered beam reflected from a captured scene. The interference fringes inflicts variation of intensity across the capturing medium. This variation results in variation of transparency which effectively forms the wanted diffraction grating. Diffraction gratings diffracts or in other words bends light and most fortunately, or rather because of the physics of diffraction, one component of the diffracted light field matches the initially captured one.

It is important to emphasise the fact that the reconstructed light field is part of some more complex light field. Finding ways of separating the wanted part of diffracted light

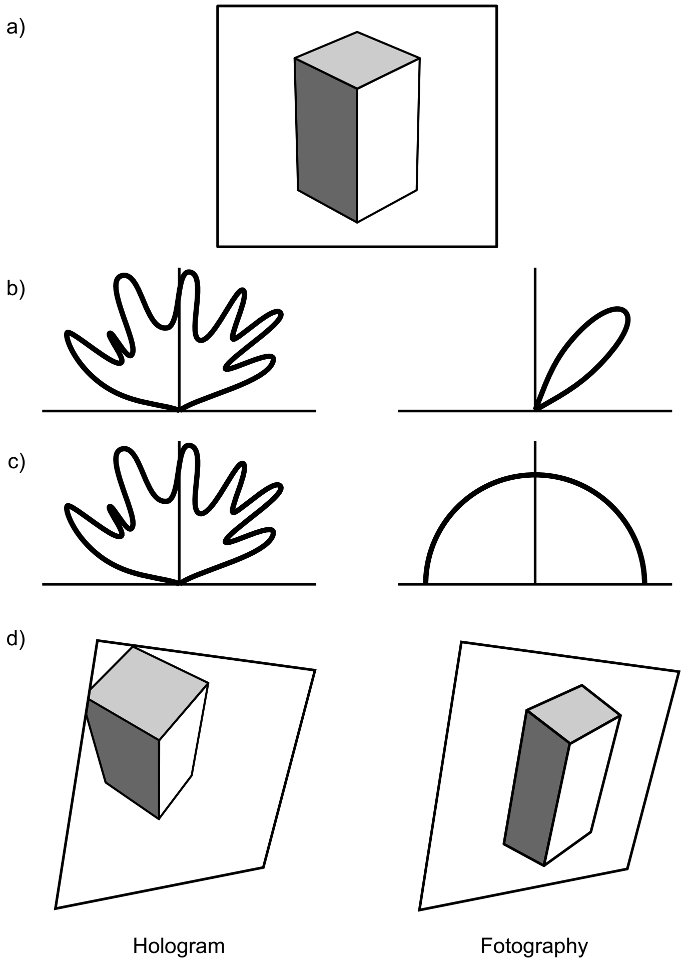

Figure 3.1: [v0.9]Difference between hologram and a common photo for a scene (a): a difference between incoming intensity from various direction for a given sample (b) and an outgoing intensity (c) for the same sample leads to a different resulting image for a tilted viewing screen (d).

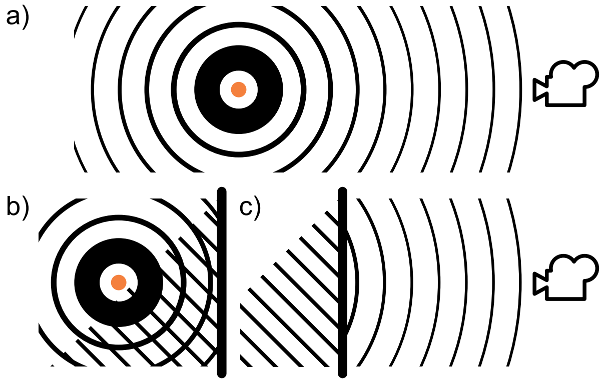

Figure 3.2: [v0.9]The most basic principle of optical recording (b) and reconstructing (c). Note, that the reconstruction (c) generates same impression as in the case of the original scene (a).

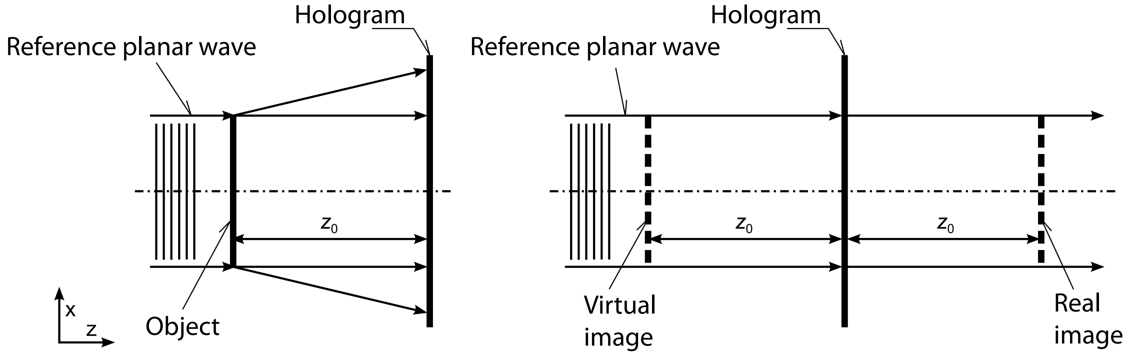

Figure 3.3: [v0.9]This is simplified depiction of capturing and reconstructing inline hologram. Adopted from [Har96].

field from the unwanted is one of the research branch on its own in holography. One of the most common solution is capturing hologram using the off-axis configuration, see Section 3.3 for more details.

The following material contains description of different hologram capturing principles. Advantages and disadvantages of each of them are discussed.

3.2 Inline hologram

The first hologram capturing was done using the setup depicted in the Figure 3.2. Because the light source, captured object and the hologram plate are aligned in one line this setup is called inline hologram. It is the simplest but also the least performing setup. It is because of the restriction placed on scene characteristics and the low quality of image reproduced from the inline hologram.

The most restricting constraint it that the captured scene has to be spatially sparse. It is because the majority of the incoming light has to get through so it can act as the reference beam. The minor part of the light is scattered on the obstacles formed by the components of the captured scene and this scattered light then acts as the scene beam. Because of this restriction the inline configuration is used usually for capturing scenes like aerosol, small particles floating in water, etc. Scenes with the opposite characteristic, i.e. small holes in some opaque screen, are not suitable for this method and cannot be captured properly.

The following text is adopted form [Har96]. The reference beam is collimated and therefore the complex amplitude does not vary across the hologram plane. It is written as a real constant r. The complex amplitude of the scattered wave varies across the hologram and is therefore written as $o\left(x,y\right)$, where $\left|o\left(x,y\right)\right| \ll r$. Note that the coordinate system is set in a such way that X axis corresponds to the horizontal direction of the hologram frame, Y axis corresponds to the vertical direction of the hologram frame and Z axis points from the scene to the hologram.

The complex amplitude at any point of the hologram frame is obtained as a sum of the reference and object beam complex amplitudes at that point. The resultant optical intensity is then obtained using Equation Equation (2.14) as

$$\begin{aligned} I(x,y) &= |r + o(x,y)|^2, \\ &= r^2 + |o(x,y)|^2 + ro(x,y) + ro^*(x,y), \end{aligned} \tag{3.1}$$

where $o^*(x,y)$ is the complex conjugate of $o^*(x,y)$.

The optical intensity is recorded on a transparency. If it is assumed that amplitude transmittance is a linear function of the intensity then it can be written as

$$\mathbf{t} = \mathbf{t}_0 + \beta T I, \tag{3.2}$$

where $\mathbf{t}_0$ is a constant background transmittance, T is the exposure time, and β is a parameter determined by the photographic material. When Equation (3.1) is substituted into Equation (3.2) the amplitude of this transparency is

$$\mathbf{t}(x,y) = \mathbf{t}_0 + \beta T \left[r^2 + |o(x,y)|^2 + ro(x,y) + ro^*(x,y) \right]. \tag{3.3}$$

To reconstruct the captured scene, the hologram is placed on the same position as during capturing and illuminated by the very same reference beam and the transmitted complex amplitude by the hologram can be then written as

$$\begin{aligned} u(x,y) &= r\mathbf{t} \\ &= r\left(\mathbf{t}_0 + \beta T r^2\right) + \beta T r\,|o(x,y)|^2 \\ &\quad + \beta T r^2 o(x,y) + \beta T r^2 o^*(x,y) \end{aligned} \tag{3.4}$$

The expression Equation (3.4) consists of four terms. The frist of the terms $r\left(\mathbf{t}_0 + \beta T r^2\right)$ constitutes the directly transmitted beam. The second term $\beta T r\,|o(x,y)|^2$ is extremely small in comparison with the others since it has been assumed initially that $|o(x,y)| \ll r$ and can be therefore neglected. The third term $\beta T r^2 o(x,y)$ is, except for a constant factor, identical with the object beam. This light field constitutes the reconstructed image. Since this image is located behind the transparency and the reconstructed light field appears to diverge from it, it is called the virtual image . The fourth term also represents the originally captured light field except it is complex conjugate of the field. This field converges to form the so called real image, which is inverted by the Z axis.

The low quality of images reproduced from inline holograms is caused by the fact that the reconstructed virtual image fully overlaps with the directly transmitted reference beam and with the blurred real image. This fact was the reason for low interest about holography at the beginning. The efficient way of separating the virtual image from its real counterpart and from the zero order beam was developed by Leith and Upatnieks in 60' and it is introduced in the following section.

3.3 Off-axis hologram

The problem of overlapping virtual image with the real image and transmitted beam was solved by creating more complicated setup. The source beam was divided and while one beam was used to illuminate the captured scene which scattered it onto the hologram plane as a scene beam the second one was directed onto the hologram without modification

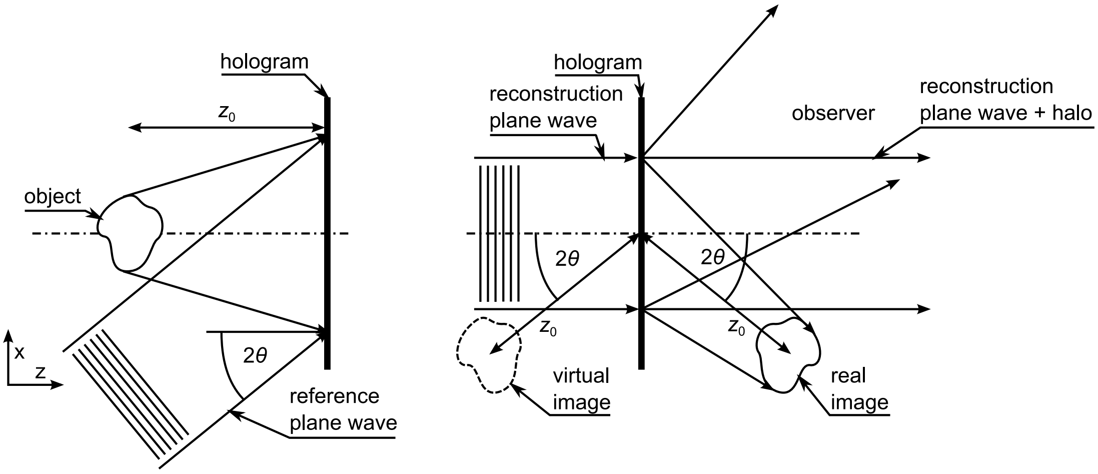

Figure 3.4: [v0.1]This is simplified depiction of capturing (left) and reconstructing (right) offaxis hologram. Adopted from [Har96].

and serves as the reference beam. This more complex configuration is depicted in the Figure 3.3.

The offaxis capturing principle can be described by the same formalism used in the previous section, text is adopted form [Har96]. The complex amplitude due to the object beam at any point on the hologram frame can be written as

$$o(x,y) = |o(x,y)| \exp[-i\phi(x,y)], \tag{3.5}$$

while that due to the reference beam is

$$r(x,y) = r\exp(i2\pi\xi_r x), \tag{3.6}$$

where $\xi_r = \sin\theta/\lambda$. The resultant intensity at the hologram plane is

$$\begin{aligned} I(x,y) &= |r(x,y) + o(x,y)|^2 \\ &= |r(x,y)|^2 + |o(x,y)|^2 \\ &\quad + r|o(x,y)|\exp[-i\phi(x,y)]\exp(-i2\pi\xi_r x) \\ &\quad + r|o(x,y)|\exp[i\phi(x,y)]\exp(i2\pi\xi_r x) \\ &= r^2 + |o(x,y)|^2 + 2r|o(x,y)|\cos[2\pi\xi_r x + \phi(x,y)]. \end{aligned} \tag{3.7}$$

The amplitude transmittance of the hologram can be written as

$$\begin{aligned} \mathbf{t}(x,y) = \mathbf{t}_0 + \beta T\{&|o(x,y)|^2 \\ &+ r|o(x,y)|\exp[-i\phi(x,y)]\exp(-i2\pi\xi_r x) \\ &+ r|o(x,y)|\exp[i\phi(x,y)]\exp(i2\pi\xi_r x)\}. \end{aligned} \tag{3.8}$$

To reconstruct the image, the hologram is illuminated by the same reference beam used for capturing. The complex amplitude $u(x,y)$ of the transmitted wave can be written as:

$$\begin{aligned} u(x,y) &= r(x,y)\,\mathbf{t}(x,y), \\ &= u_1(x,y) + u_2(x,y) + u_3(x,y) + u_4(x,y), \end{aligned} \tag{3.9}$$

where

$$\tilde{u}_1(x, y) = \mathbf{t}_0 \exp(i2\pi\xi_r x), \tag{3.10}$$

$$\tilde{u}_2(x, y) = \beta T r |o(x, y)|^2 \exp(i2\pi\xi_r x), \tag{3.11}$$

$$\tilde{u}_3(x, y) = \beta T r^2 o(x, y), \tag{3.12}$$

$$\tilde{u}_4(x, y) = \beta T r^2 o^*(x, y) \exp(i4\pi\xi_r x). \tag{3.13}$$

The first term $\tilde{u}_1(x, y)$ constitutes the directly transmitted reference beam attenuated by the constant factor. The second term $\tilde{u}_2(x, y)$ is responsible for some sort of halo surrounding the reference beam. The angular spread of the halo depends on the extend of the object. The third term $\tilde{u}_3(x, y)$ is identical with the original object wave so it is the virtual image. And finally the fourth term $\tilde{u}_4(x, y)$ is the conjugate of the original object wave so it is the real image. However, in a case of the off axis hologram, there is additional term $\exp(i4\pi\xi_r x)$ which indicates that the conjugate wave is deflected from the Z axis at an angle approximately twice that which the reference wave makes with it.

For this reason, the real and virtual image are reconstructed at different angles from the directly transmitted beam and from each other. If the offset angle θ is large enough the three will not overlap. This method therefore eliminates all the drawbacks of the Gabor's inline hologram.

The minimum value of the offset angle θ required to ensure that each of the images can be observed without any interference from its twin image, as well as from the directly transmitted beam and the halo of scattered light surrounding it, is determined by the minimum spatial carrier frequency ξ_r for which there is no overlap between the angular spectra of the third and fourth terms, and those of the first and second terms. According to the [Har96] they will not overlap if the offset angle θ is chosen so that the spatial carrier frequency ξ_r satisfies the condition

$$\xi_r \geq 3\xi_{\max}, \tag{3.14}$$

where $\xi_{\max}$ is the highest frequency in the spatial frequency spectrum of the object beam.

The restriction on scene characteristic found in the inline hologram does not apply in a case of the off-axis hologram. More complex and interesting scenes can be captured and consequently reconstructed. However, the off-axis configuration is more sensitive to the coherence length of the light source. The difference of path lengths of the reference and the scene beam must not exceed the coherence length of the light source used. Although the scene can be more or less arbitrary, the problem of laser speckle still persists.

3.4 Additional hologram types

Apart from the configurations described in the Sections 3.2 and 3.3, there are several other hologram capturing configurations. In this section these configurations are described:

- Fourier hologram
- Image hologram
- Fraunhofer hologram

3.4.1 Fourier hologram

Fourier hologram is the one in which the complex amplitudes of the waves that interfere at the hologram are the Fourier transforms of the complex amplitudes to the original object and reference waves. This implies an object that lies in a single plane or is of limited thickness.

Using the same formalism used in the cases of the inline and the off-axis holograms, the light field's complex amplitude leaving the object plane is $\tilde{o}\left(x,y\right)$, its complex amplitude at the hologram plate located in the back focal plane of the lens and it is

$$\tilde{O}(\xi,\eta) = \mathcal{F}\left\{\tilde{o}(x,y)\right\}. \tag{3.15}$$

The reference beam is derived from the point source also located in the front focal plane of the lens. If $\tilde{\delta}(x+b,y)$ is the complex amplitude of the light field leaving this point source, the complex amplitude of the reference light field at the hologram plane can be written as

$$\tilde{R}\left(\xi,\eta\right) = \exp(-i2\pi\xi b). \tag{3.16}$$

The intensity in the interference pattern produced by these two waves is, therefore,

$$\begin{aligned} I\left(\xi,\eta\right) =& 1 + \left|\tilde{O}\left(\xi,\eta\right)\right|^2 + \tilde{O}\left(\xi,\eta\right)\exp\left(i2\pi\xi b\right) + \\ & \tilde{O}^*\left(\xi,\eta\right)\exp\left(-i2\pi\xi b\right). \end{aligned} \tag{3.17}$$

To reconstruct the image, the processed hologram is placed in the front focal plane of the lens and illuminated with a collimated beam of monochromatic light. If it is assumed that this wave has unit amplitude and that the amplitude transmittance of the processed hologram is a linear function of $I\left(\xi,\eta\right)$, the intensity is the interference pattern, the complex amplitude of the wave transmitted by the hologram is

$$\tilde{U}\left(\xi,\eta\right) = \mathbf{t}_0 + \beta T I\left(\xi,\eta\right). \tag{3.18}$$

The complex amplitude in the back focal plane of the lens is then the Fourier transform of $\tilde{U}\left(\xi,\eta\right)$,

$$\begin{aligned} \tilde{u}\left(x,y\right) &= \mathcal{F}\left\{\tilde{U}\left(\xi,\eta\right)\right\}, \\ &= \left(t_0 + \beta T\right)\tilde{\delta}\left(x,y\right) + \beta T\tilde{o}\left(x,y\right) \star \tilde{o}\left(x,y\right), \\ &\quad + \beta T\tilde{o}\left(x-b,y\right) + \beta T\tilde{o}^*\left(-x+b,-y\right). \end{aligned} \tag{3.19}$$

The wave corresponding to the first term on the right-hand side of equation comes to a focus on the axis, while the other that corresponds to a second term forms a halo around it. The third term produces an image of the original object, shifted downwards by a distance b, while the fourth term gives rise to a conjugate image, inverted and shifted upwards by the same distance b. Both images are real and can be recorded on a photographic film placed in the back focal plane of the lens. Since the film records the intensity distribution in the image, the conjugate image can identified only by the fact that it is inverted. More details can be found in [Har96].

3.4.2 Image holograms

Image holograms are special in the way they record the images and can be considered as a second step in creating a hologram. Instead of direct recording of the object, a previously hologram is used. The real image produced by that hologram is recorded through lens. The hologram plate then can be positioned in such manner that the image of the object straddles the plate. In such case the reconstructed image is formed in the same position with respect to the hologram, so a part of the image appears to be in front of the hologram and the remainder is behind it.

Image holograms are reconstructible by use of a source of appreciable size and spectral bandwidth and will still produce an acceptably sharp image. They have also increased luminosity. However, the viewing angles are limited by the aperture of the imaging lens. More details about image holograms can be found in [Har96].

3.4.3 Fraunhofer hologram

Fraunhofer holograms are special case of the inline holograms. This class of holograms impose further constraint onto the scene configuration. In this case, the captured scene has to be small enough for its Fraunhofer diffraction pattern to be formed on the photographic plate. The object is small enough if its distance z_0 and lateral dimensions x_0 and y_0 satisfies the far-field condition:

$$z_0 \gg \left(x_0^2 + y_0^2\right)/\lambda. \tag{3.20}$$

When such Fraunhofer hologram is illuminated the light contributing to the conjugate image is spread over a large area in the plane of the primary image. This causes a conjugate image to form only weak uniform background instead of disturbing noise. As a result, the primary image can be viewed without significant interference from its conjugate. More details can be found in [Har96].

3.5 Final notes

The optical holography is widely used in many areas. In the context of this work, the most interesting is object capturing and reconstructing for display purposes. In this area, optical holography has many advantages and also many limitations. Among the advantages belongs the speed and accuracy of capturing. Among disadvantages belongs unnatural illumination produced by the coherent monochromatic light produced by lasers.

For practical display purpose, the digital medium is more suitable. It has again many advantages and many disadvantages. Among the advantages belongs the universality of the digital medium. Also the synthesis process can neglect some restriction imposed on the captured scene by the physical matter of the problem. The disadvantages are immaturity of the hardware and high computation demands of the synthesis process. This work address the former disadvantage.

Chapter 4

Digital Holography

Holography is widely used in many areas for various purposes. It is not surprising that their digitalised variation appeared quite early. The digitalisation brought the possibility of computer processing but also brought the sampling based nuisances like aliasing. There can be identified three major areas of holography that are usually addressed in a context of digital holography. They are the capturing, the reproduction and hologram fringes synthesis. Furthermore, digital holography introduces one specific issue, that don't have parallel in optical holography. The area is numerical reconstruction.

4.1 Digital capturing

Electronic devices for capturing light intensity are known for some time. The advent of digital cameras that begun in 90' helped to evolve this area into cutting edge technology. The sensor arrays have as much as tens millions of sensor elements. However, the resolution is fine but the density more important for holography. The pitch between two sensors directly influences the frequencies that can be captured and large frequencies captured means better hologram.

4.2 Digital reproduction

The reproduction is more interesting but less completed area. The wavefront of the incident reference beam has to be modulated according to the hologram recording. This is done by the spatial light modulator or SLM. Such devices are capable of change phase of the light wavefront. If hologram is used as the input for SLM, the recorded scene is reconstructed. There are numerous technologies that are used to make SLM's work. The most important property of each SLM is again spatial density of the individual elements. For holography purposes the appropriate size of such element is measured in ones of micrometres 10^{-6}m which is quite technological challenge.

4.3 Hologram Synthesis

The synthesis of a hologram by a computer is vital for holographic display technology. It is also clear that such synthesis should be performed in a real time so that interactive work would be possible. Waiting even for several seconds for cursor position update is unbearable. This requirement is in contradiction with the computational requirements of the diffraction phenomenon simulation. This contradiction can be resolved either by employing massive computational power which is perfectly reasonable as proved by the contemporary GPU's, or by reducing the complexity significantly which is a preferred way in computer science.

The complexity of the synthesis could be illustrated by the following example. Let the target holographic display is a planar display similar to the 17" LCD. Usual resolution of 17" LCD is 1280×1024 so the pitch between pixels is 0.25 mm. A computer has to compute 1.3×10^6 samples if whole image is refreshed. The holographic display with 1.0 μm pitch between elements has resolution 320000×256000 so a computer has to compute 8.2×10^{10} samples if whole image is refreshed. That is 62500 times more samples in comparison to the LCD. It should be also noted that the minimal frame rate for interactive work is 15 fps so the hologram has to be computed under 60 ms and finally, the data stream for such frame rate is 1.2 TB per second.

The specifications in the previous paragraph are, of course, the final goal specifications. For the practical experiments more coarse parameters suffice. Moreover, there is no such holographic device in the world, that has such parameters. The holograms computed for this work have sizes measured in centimetres and usual pitch between samples is 10 μm.

The most straightforward method of the hologram synthesis is to compute light field due to the intended scene content and then the hologram is computed by adding the reference beam. The light field could be computed by numerical simulation of the diffraction phenomenon according to one of the diffraction models introduced in the section 2.5. It is fairly easy to do if the scene consists of planar object parallel to the hologram plane. The light field due to this object is computed using the Fourier transform. The complications arise if scene contains three dimensional object.

The simplest hologram is a hologram of a single point source of uniform spherical waves. Spherical waves are governed by the equation Equation (2.25). For each point at the hologram, distance r to the point source is computed and the equation Equation (2.25) is evaluated. If scene contains more points, then the complex amplitude obtained from the equation Equation (2.25) for each point is accumulated and the final value constitutes the final complex amplitude. This point source based hologram computation gave a birth to the raytracing methods.

The basis of the raytracing method is to cast rays from each sample point on a hologram frame in a uniform way into the scene. Rays that intersects the objects of a scene are evaluated as a contribution from a point source and the result is accumulated so after evaluating all rays, the final value is obtained. This is sort of reversal of the method described in the previous paragraph.

The next possible solution is to decompose the scene contents onto simple primitives. The diffraction pattern of the primitives can be computed analytically and the final light field is obtained by accumulating the contributions from all primitives. This method has several unsolved problems. They are surface intensity variation and occlusion.

The last known approach to the hologram synthesis is frequency based. Under some convenient conditions, the light field can be obtained from Fourier transform. These methods are fastest but they suffer from similar drawbacks as the pattern based methods. The most significant problem is occlusion.

4.4 Hologram Reconstruction

The hologram reconstruction is a process that leads to computation of wave distribution over an arbitrary surface according to a given diffraction pattern. The diffraction pattern is a result of an interaction between the reconstruction wave and a hologram, where the hologram modulates the reconstruction wave, see Section 3.

4.4.1 Diffraction pattern reconstruction

If both source and target surface are parallel or tilted planes then the wave propagation is a direct application of relations described in the Section 2.6. If an approach that is based on angular spectrum propagation is applied a loss of information occurs for tiled planes. The loss is caused by the transformation of the spectrum according to the transformation matrix so that the target frequency is falls outside the range of the target spectrum and thus it is clipped.

The occurrence of the loss depends on both angle of rotation and sampling step. For a discrete diffraction patterns with sampling step greater than $\lambda/2$ is the probability higher [TB93]. In such case a shifting of a central frequency is assumed solution. For example, if the transformation is a plain rotation around the X-axis then the center $(0, 0)$ of the source plane is actually mapped to a frequency $(\beta/\lambda, 0)$, where β is related to a cosine of the rotation angle.

In order to avoid the loss due to clip of frequencies a center is shifted to $(\beta/\lambda, 0)$ on the target plane. As all operations of propagation takes place in angular spectrum, the shifting of the center can be applied many times without a degradation cause by the clipping. On the other hand a degradation may occur as the target–source frequency back-mapping requires a sample that is positioned between two known samples and thus it has to be estimated. Even a bilinear interpolation can be utilized with reasonable results [TB93].

An arbitrary target or source geometry is an arbitrary surface described by a function then relations are complication introducing non-linearities as the source/target coordinates are expressed as functions [Ros99]. Such non-linearities causes inability to apply Fourier transform for improving of the computational complexity and thus a full approximation of the integral by the summation is required. If a given direction retains a linear nature then the Fourier transform may be utilized in its 1D form.

4.4.2 Retrieval of Phase and/or Amplitude

Hologram is an encoded form of wave distribution on a given surface such as the plane. If a diffraction pattern propagation is applied to the hologram after the reference wave then the wave proportional to the original scene wave is recreated. Besides that, also other components are recreated as well. Yet, in some cases only the original scene wave is the desirable result and such result can be extracted from the hologram numerically.

From the computational point of view, the information about the original scene encoded inside the intensity of the hologram can be extracted by two major approaches: one based on a phase-shift and second base on a solution of equation set for a small neighborhood. The phase-shift approach utilizes a set holograms of the same scene with a phase shifted by a fraction of the wavelength [YZ97]. It can be shown that if phase shifts are selected properly an accurate estimation of $\tilde{u}$, i.e. phase and amplitude, of the scene is equivalent to equation set solution.

Phase-shift methods offers a reasonable accuracy but requires a set of holograms to be created. The slightly different approach that is still capable to reasonable result has no such requirement but it assumes a certain conditions [LBU04]. It assumes an off-axis hologram and that in a small, approximately 3×3 neighborhood, there is no variation of phase and amplitude in the scene wave distribution at the hologram. By the knowledge of the recording wave it is able to express intensities at members of the neighborhood as a non-linear set of equations. Such set can be transformed to a set of linear equations by application of weights based on B-spline.

Bibliography

[BW05] M. Born and E. Wolf. *Principles of Optics.* Cambridge University Press, 7th edition, 2005.

[EO06] G.B. Esmer and L. Onural. Computation of holographic patterns between tilted planes. In *Holography 2005*, volume 6252, page 62521K. SPIE, 2006.

[Gab49] D. Gabor. Microscopy by reconstructed wavefronts. 1949.

[Goo05] J.W Goodman. *Introduction to Fourier Optics.* Roberts & Company Publishers, 3rd edition, 2005.

[Gra03] J.R. Graham. Wave optics. WWW http://grus.berkeley.edu/ jrg/ScalarWave/, 2003.

[Har96] P. Hariharan. *Optical Holography: Principles, techniques and applications.* Cambridge University Press, 2nd edition, 1996.

[Kra04] F. Krausz. Photonics: Lecture notes, 2004.

[LBL02] D.R. Luke, J.V. Burke, and R.G. Lyon. Optical Wavefront Reconstruction: Theory and Numerical Methods'. *SIAM Review*, 44:169–224, 2002.

[LBU04] M. Liebling, T. Blu, and M. Unser. Complex-wave retrieval from a single off-axis hologram. *J. Opt. Soc. Am. A*, 21(3):367–377, 2004.

[Luc94] M. Lucente. *Diffraction-Specific Fringe Computation for Electro-Holography.* PhD thesis, MIT, 1994.

[Mie02] K.D. Mielenz. Optical diffraction in close proximity to plane apertures. i. boundary-value solutions for circular apertures and slits. *J. res. Natl. Inst. Stand. Technol.*, 107(4):335–362, 2002.

[MNF$^+$02] O. Matoba, T. J. Naughton, Y. Frauel, N. Bertaux, and B. Javidi. Three-dimensional object reconstruction using phase-only information from a digital hologram. In *Three-Dimensional TV, Video, and Display.*, volume 4864, pages 122–128. SPIE, 2002.

[Ros99] J. Rosen. Computer-generated holograms of images reconstructed on curved surfaces. *Appl. Opt.*, 38(29):6136–6140, 1999.

[SJ05] U. Schnars and W. Juepner. *Optical Holography: Principles, techniques and applications.* Springer, 2005.

[TB93] T. Tommasi and B. Bianco. Computer-generated holograms of tilted planes by a spatial frequency approach. *Journal of the Optical Society of America A*, 10:299–305, February 1993.

[Wei] E. Weisstein. World of science. WWW http://scienceworld.wolfram.com/.

[YAC02] L. Yu, Y. An, and L. Cai. Numerical reconstruction of digital holograms with variable viewing angles. *Optics Express*, 10:1250–+, 2002.

[YZ97] I. Yamaguchi and T. Zhang. Numerical reconstruction of digital holograms with variable viewing angles. *Optical Letters*, 22(16):1268–1270, 1997.

List of Figures

List of Tables