University of West Bohemia
Department of Computer Science and Engineering
Univerzitní 8
CZ 306 14 Plzeň
Czech Republic

Digital Hologram Synthesis

State of the art and a concept of doctoral thesis

Ing. Martin Janda

Technical Report No. DCSE/TR-2007-02
April, 2007
Distribution: public

To my Father

Abstract

The first part of this work summarises the basic knowledge about holography principles including wave optics and diffraction. The summarised knowledge is consequently exploited in the second part of this work that regards the methods of digital hologram synthesis. The synthesis methods were evaluated and compared and the perspective ones were announced. The perspective synthesis methods served as a starting point of the further research and the first output of the research is included in the third part of this work. It is a novel approach to synthesis based on ray casting. The last part of this work regards the roadmap for the future work.

První část práce obsahuje přehled principů holografie a to včetně základů vlnové optiky a difrakce. Principy jsou následně využity v druhé části práce, která obsahuje přehled známých postupů pro syntézu digitálních hologramů. Syntetizační metody byly vyhodnoceny a porovnány a byly vyhlášeny perspektivní metody. Ty potom sloužily jako počáteční bod dalšího výzkumu jehož první výsledky jsou uvedeny v třetí části práce. Jedná se o novou syntetizační metodu založenou na metodě vrhání paprsku. Poslední část práce obsahuje popis cílů a směrů dalšího výzkumu.

Contents

Acknowledgements

Work has been supported by these projects:

- 3DTV NoE, grant No. 511568
- LC CPG, MŠMT ČR project No.LC 06008

Chapter 1

Introduction

Display devices are widely used for presenting visual information. For example paper made very good service for a very long time until it was replaced with more universal electronic displays quite some time ago. The technology of the electronic displays evolved ever since until it reached contemporary quite advanced level. The problem of the paper and inherently the electronic, e.g. LCD liquid crystal display - LCD, cathode ray tube - CRT etc., display is their planarity, which restricts the kind of information possible to display without degradation.

The degradation of information is most significant if three dimensional information is displayed on two dimensional display, which is usually managed by some kind of projection operation, e.g. perspective projection. This projection is unfortunately done in many areas for instance computer aided design - CAD or multimedia, e.g. games. As a consequence, it is not surprising that many people are trying to develop a genuine three dimensional display, which would present three dimensional content without degradation.

The genuine three dimensional display would provide three dimensional reproduction of arbitrarily complex objects. Such reproduction is more natural to a human viewer who accordingly understands the displayed data faster and more reliably. In addition, a true three dimensional display device supports more than one viewer because each viewer perceives a correct image according to her location.

The research on three dimensional display is not a new thing. Some stereoscopy based techniques have been invented many years ago. The three dimensional perception originates form the fact that humans have two eyes slightly shifted in horizontal direction and therefore they deliver slightly different images to the brain which translates the difference into the depth perception. However, the images are again showing the object from one point of view and cannot be observed from other angles unless some head tracking is involved.

Another way is a direct extension of the two dimensional display. Instead of lighting up a point of a 2D grid a point of 3D grid is lit up instead. Such displays are usually referred as volumetric displays. There are many methods of lighting up a point in a space. For instance, there are methods relying on the persistence of view of a human eye and use fast moving projection screen. Other methods uses special materials which emits light at the intersection of two laser beams or a pulsed laser is used to create points of glowing plasma in air. To conclude, the volumetric displays have larger viewing angles but on the contrary the active volume itself is a constraint and the volumetric displays are incapable of handling occlusion.

Speaking generally, various three dimensional displays addresses some subset, usually mutually disjunctive, of the aspects of 3D vision but fails to provide others. Hence, a flawless genuine three dimensional display was not yet developed. There is however one promising technology, which could have remedy for all the mentioned problems. This technology is, as one might expect, holography.

Holography is the ultimate displaying technology since it is capable of reproducing the whole light field at the microscopic level. The visual perception from a hologram is therefore undistinguishable from a perception of a real object. The digital version of holography is especially useful because it allows creating holograms of artificial scenes without the quite strict restrictions imposed on optical holography.

This is maybe a good place to state that not everything called holography is actually holography! Especially the science fiction genre have introduced quite severe misconceptions about holography. The real holography is actually an analogy of photography. It even uses the same materials for recording as photography does. The detailed description of holography can be found in the Chapter 2.

Now one might rightfully ask what is so great about holography. The answer is in the word holography itself. It is derivation from two latin words holos and graf which mean approximately the whole recording. Hologram is recording of both amplitude and phase of a light field emerging from the recorded scene. For this reason holography is so convenient technology for displaying purposes. Holography is able to provide all depth cues, e.g. motion parallax, binocular parallax, accommodation, occlusion. In this aspect it surpasses all other displaying technologies.

Since hologram contains the full description of light field it can be therefore easily used as data source for other technologies. An image corresponding to some fixed viewing point can be extracted and displayed on a standard two dimensional display. Two images from two mutually shifted viewpoints can be extracted and used at the stereoscopic device. In a case of the volumetric displays the situation is more complicated. The intensity of a point in a space, which is the information the volumetric display requires, can be only estimated from a hologram because

light from other points interferes with the one in focus.

It is apparent that holography has a lot of advantages unfortunately there are also several major disadvantages that makes digital holography hard to apply directly. The first disadvantage and the second one is computational and storage requirements respectively. In other words, digital hologram is hard to compute and hard to store. The third problem is replaying the content of a hologram. From the three mentioned problems the first one is the most serious one and also it is the topic of this thesis.

The advertised advantages of holography are simultaneously the reason for the difficulties one encounters during hologram synthesis. Holography stands on principles of optical physics which operates are very small scales, usually in hundreds of nanometres. This fact implies utilisation of very high sampling frequencies and inherently high computation demands because each sample needs some non-trivial evaluation.

The main goal of this work is to design a compact and complete synthesis pipeline architecture similar to the one used for image synthesis where mathematical description of a scene serves as an input and 2D image is produced as an output. Analogically the holographic pipeline should have the same mathematical description as an input but hologram should be produced as an output instead.

The main area of research is the simulation of light propagation in a scene to obtain appropriate light field sufficiently accurate, from which an arbitrary hologram can be computed. The light propagation is the most problematic area because exact simulation is impossible to compute in a reasonable time. Drastic simplifications have to be applied to obtain the required speedup while ensuring that the impact of those simplification on the visual quality will be minimal.

Another area of research is implementing advanced rendering features known from classical computer graphics image synthesis like non-trivial illumination models, shading capabilities, reflections and refraction or global illumination. Unfortunately each one of the presented feature only adds complexity to the light field computation and therefore only increase stress on the simplification level.

The content of this document is organised in this manner:

- Chapter 2 Contains necessary introduction into holography principles
- Chapter 3 Contains overview of digital hologram synthesis methods
- Chapter 4 Contains a description of experimental synthesis method
- Chapter 5 Contains summary and conclusion

Chapter 2

Holography

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 provided images of poor quality and further 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 Sec. 2.2.

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 Sec. 2.3.

Holography is built on quite complex physical laws of optics. These laws are referenced many times in the text of this work and therefore basics of the wave optics are provided in the Sec. 2.1 of this chapter to avoid confusion when methods of digital hologram synthesis are described in the Chapter 3. The described topics are wave equation, interference, coherence and diffraction. Large portions of this chapter were taken from [JHS06b].

2.1 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, referred as 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. Moderate introduction into the interference phenomenon and related topics is included in this thesis because the methods of digital hologram synthesis refer to this phenomenon quite often. The interference in particular is described in the Sec. 2.1.2.

The principle of the interference phenomenon is based on the wave nature of light and therefore a short introduction into the mathematics of waves is provided in the Sec. 2.1.1. The relation of the wave calculus to the physics of light is also presented there.

From interference a more complex phenomenon of diffraction is derived. The diffraction is responsible for forming the object beam on a photographic plate, see Sec. 2.2 for reference. The exact and general 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 Sec. 2.1.5.

2.1.1 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 [Goo05] applies in 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 Eq. 2.5 and Eq. 2.6 are vector equations, but under certain circumstances, all components of each vector 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 Eq. 2.7 describes the behavior of a wave in a linear, uniform, isotropic, homogeneous, and non-dispersive material and it constitutes the core for the scalar wave theory that serves as base upon which all assumptions in this thesis are build on. Even though the scalar wave theory is an approximation rather than an exact description it is satisfactory as it describes the behavior of 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 $\mathbf{p}$. The Eq. 2.8 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. The term diffraction pattern refers to an array of complex notations for a wavefunction $u(\mathbf{p},t)$ defined by the Eq. 2.9. The wavelength of 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 vacuum, if not noted otherwise.

A wave defined by the Eq. 2.9 has to satisfy the scalar wave theory, i.e. Eq. 2.7. If the Eq. 2.9 is substituted into the scalar wave theory a relation known as the Helmholtz equation is obtained:

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

where $k = 2\pi/\lambda$ is known as a wavenumber. 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 the complete solution because $\tilde{u}(\mathbf{p}, t)$ 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 optics and the wave optics is straightforward. The relation is clearly visible from the specification of a wavefront. Wavefront is an iso-surface that consists of wave function points having 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 the given point $\mathbf{p}$. The normal also constitutes the direction of wave local propagation and thus it also constitutes the direction of a ray in a ray optics [Kra04].

Another important property of light is its optical power. This is important when hologram is captured because optical power determines the amount of energy delivered to a photographic material and/or sensor. Optical intensity is defined as the time average of an amount of energy that crosses an unit area 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$\tilde{u}^*$ [Har96]:

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

The optical intensity is the final product of digital hologram synthesis. One has to compute the diffraction pattern at the hologram's plane first and then the optical intensity is evaluated to create the actual interference pattern or fringe pattern.

2.1.2 Interference

In the Sec. 2.1.1 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 space are the same or near the same, the

constructive interference occurs at that point and intensity increases. If the phases are opposite or almost opposite the destructive interference occurs and intensity decreases, see Fig. 2.1.

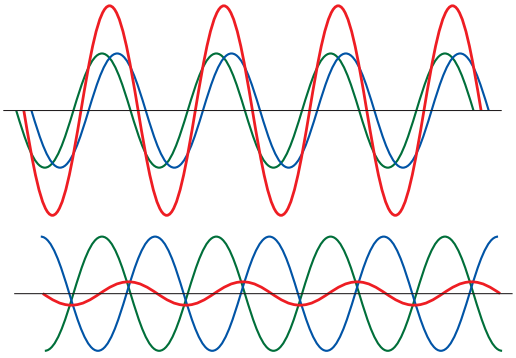


Figure 2.1: Two results of two waves interference (green and blue). Constructive interference in upper and destructive interference in the lower.

As a result of interference 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 light or none at all. However this is true only for the coherent light only.

The coherence is described in the Sec. 2.1.3 but basically it determines the stability of the interference effect in time. The coherent light produces stable interference pattern, i.e. 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 and vice versa. This flickering is so fast that the human eye or any other integrating sensor is incapable of registering it because the human eye integrates the incoming intensity and the average value is obtained.

The coherence is also essential for the holography because during the process of creating an optical hologram a photosensitive material is exposed to the interference pattern formed by interference. Such pattern has to be stable for the exposition time to be successfully recorded. 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 by exploiting the complex algebra. The advantage of the complex notation of the wave equation now pays off. The Eq. 2.13 demonstrates that interference can be written as a summation of the complex amplitudes of each interfering wave assuming that waves are monochromatic or in other words coherent.

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

The optical intensity due to the interference of two waves is therefore computed according to the Eq. 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 Eq. 2.14 is very important in holography. It rules the computation of intensity due to two interfering waves with complex amplitudes $\tilde{u}_1$ and $\tilde{u}_2$ constituting the scene wave and the reference wave respectively, see Sec. 2.2.1. The intensity is a result of adding the intensities of both waves and the variation term $\cos\left(\phi_1 - \phi_2\right)$, which constitutes 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 [Luc94].

2.1.3 Coherence

The coherence is quite important property light should exhibit to be useful for holography purposes. It is very important in 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 and 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. This visibility is quantified by 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 clearly visible interference pattern the incoherent ones won't create visible interference fringe at all.

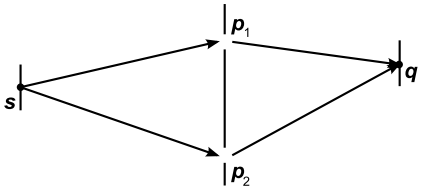


Figure 2.2: A configuration for exploring the coherence [Har96].

To formalise the coherence the optical setup in the Fig. 2.2 can be used. The ideal point source **s** emits a monochromatic wave for an infinite time period. The wave propagates through two slits $\mathbf{p}_1$ and $\mathbf{p}_2$ which became secondary sources. The coherence is examined according to the point **q**. In this theoretical example, the light from the secondary sources will be always perfectly coherent.

In the reality light source is not ideal because it will be never strictly monochromatic and it won't be point either. The real source will be of finite size and will be only quasi-monochromatic. The time varying field from such source 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 $\tilde{\gamma_{12}}$ of waves generated by two secondary light sources $\mathbf{p}_1$ and $\mathbf{p}_2$ positioned according to the Fig. 2.2 is defined as an normalised cross-correlation of the respective analytic signals $\tilde{v}_1$ and $\tilde{v}_2$. The cross-correlation R of two stationary time-dependent functions $g(t)$ and $h(t)$ is defined as Eq. 2.16 and normalised cross-correlation R_N is defined as Eq. 2.17, see [Har96].

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

$$R_N(\tau) = \frac{\langle g^*(t)h(t+\tau)\rangle}{[\langle g^*(t)g(t)\rangle\langle h^*(t)h(t)\rangle]} \tag{2.17}$$

The complex coherence is a function of time delay τ. In the Fig. 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 Eq. 2.15 and Eq. 2.17, the complex coherence, also known as the complex degree of coherence, of two light waves is computed as:

$$\tilde{\gamma}_{12}(\tau) = \frac{\langle \tilde{v}_1(t+\tau)\tilde{v}_2^*(t)\rangle}{[\langle \tilde{v}_1(t)\tilde{v}_1^*(t)\rangle\langle \tilde{v}_2(t)\tilde{v}_2^*(t)\rangle]^{1/2}} = \frac{\langle \tilde{v}_1(t+\tau)\tilde{v}_2^*(t)\rangle}{(I_1 I_2)^{1/2}} \tag{2.18}$$

The amplitude $|\tilde{\gamma}_{12}(\tau)|$ of complex coherence describes light in terms of coherency. If $|\tilde{\gamma}_{12}(\tau)| = 1$ then light is coherent, 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 the secondary point sources $\mathbf{p}_1$ and $\mathbf{p}_2$ and their distance the source **s** and the Eq. 2.14 it is possible to express the intensity at the point **q** as:

$$\begin{aligned} I &= I_1 + I_2 + \langle \tilde{v}_1(t+\tau)\tilde{v}_2^*(t) + \tilde{v}_1(t+\tau)\tilde{v}_2^*(t)\rangle, \\ &= I_1 + I_2 + 2\Re[\langle \tilde{v}_1(t+\tau)\tilde{v}_2^*(t)\rangle] \end{aligned} \tag{2.19}$$

where I_1 and I_2 are the intensities at $\mathbf{q}$ when $\mathbf{p}_1$ and $\mathbf{p}_2$ acts separately. From Eq. 2.18 and Eq. 2.19 one obtains:

$$\begin{aligned} I &= I_1 + I_2 + 2(I_1 I_2)^{1/2} \Re[\gamma_{12}(\tau)] \\ &= I_1 + I_2 + 2(I_1 I_2)^{1/2} |\gamma_{12}(\tau)| \cos \phi_{12}(\tau) \end{aligned} \tag{2.20}$$

where $\phi_{12}(\tau)$ is the phase of $\gamma_{12}(\tau)$.

Interference fringes are produced by the variation of the $\phi_{12}(\tau)$ across screen. 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]:

$$\mathcal{V} = \frac{2(I_1 I_2)^{1/2}}{I_1 + I_2} \cos(\psi), \tag{2.21}$$

where ψ is an angle between electrical vectors of both light waves and thus represents polarisation. Note that the visibility drops to zero if $\psi = \pi/2$, i.e. waves of polarised light do not create a visible interference pattern if polarisation directions are perpendicular to each other.

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

$$\mathcal{V} = |\tilde{\gamma}_{12}(\tau)|. \tag{2.22}$$

The coherence of the field produced by any light source can be studied from two aspects. First, the temporal coherence refers to the monochromacity of the light source. Providing the light source is very small but radiates over a range of wavelengths, the complex coherence depends only on τ and difference in transit times between each of secondary sources $\mathbf{p}_1$ and $\mathbf{p}_2$ and the primary source $\mathbf{s}$. Thus, the complex coherence is actually a normalised autocorrelation of the function $\tilde{v}(t)$.

If requirements for the Eq. 2.22 are fulfilled, the degree 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 fulfils the following condition:

$$\Delta_\tau \Delta_\nu \approx 1, \tag{2.23}$$

where $\omega = 2\pi\nu$. The time Δ_τ denotes the coherence time of the given radiation. From this quantity the coherence length is derived. 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:

$$\Delta_l \approx c\Delta_\tau \approx c/\Delta_\nu \approx \lambda_0^2/\Delta_\lambda. \tag{2.24}$$

Second, the spatial coherence refers to the spatial extend of the source. Providing the difference of optical paths $\mathbf{sp}_1$ and $\mathbf{sp}_2$ is small enough for the time

difference to be $\tau \approx 0$ the spatial coherence relates the range between two points and the 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 coherence area then waves generated by these two slit do not form visible interference pattern.

Any light source can provide coherent light. It is managed by the means of filtering. The spatial coherence is achieved by the spatial filter - pinhole and the temporal coherence is achieved by the wavelength filter - color filter. Both filters represents significant reduction of optical intensity therefore the light source has to be very powerful. This was a great problem before invention of laser, which provides highly monochromatic and collimated light of high intensity.

An important side effect of the coherence is that if light is coherent both spatially and temporally it is possible to neglect the temporal component $-\mathrm{i}\omega t$ in the wavefunction Eq. 2.9 and leave only the wave distribution to be examined or computed. This can be interpreted as an exploration of the wave distribution for an infinitely short time period. And since 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 neglecting the temporal term. Thus, in the following text the complex amplitude constitutes a full description of the monochromatic wave distribution, if not noted otherwise.

2.1.4 Elementary waves

Elementary waves represents the simplest solution of the Helmholtz equation. There are distinguished two simplest waves: a planar wave and a spherical wave.

The planar wave is a wave where wavefront is an infinite plane. The complex amplitude of such wave is [Kra04]:

$$\begin{aligned} \tilde{u}(\mathbf{r}) &= \tilde{a}\exp(\mathrm{i}\mathbf{k}\cdot\mathbf{r}) \\ &= \tilde{a}\exp[\mathrm{i}(k_x x + k_y y + k_z z)], \end{aligned} \tag{2.25}$$

where $\mathbf{k} = (k_x, k_y, k_z)$ denotes wavevector and $\tilde{a}$ is a complex amplitude that defines the phase and the amplitude at the origin of the wave. The vector $\mathbf{r} = (x, y, z)$ denotes radius vector. The length of the wavevector $|\mathbf{k}|$ is referred as wavenumber. Intensity I of the planar awave is constant and equals to $I = |\tilde{a}|^2$. The wavefunction governing the planar wave is then following:

$$u(\mathbf{r}, t) = |\tilde{a}|\cos(\varphi_{\tilde{a}} + \mathbf{k}\cdot\mathbf{r} - \omega t). \tag{2.26}$$

The spherical wave is a wave where wavefronts have a form of concentric spherical surfaces centred at the point source. The complex amplitude of such spherical wave is:

$$\tilde{u}(\mathbf{r}) = \frac{\tilde{a}}{r}\exp(\mathrm{i}kr), \tag{2.27}$$

where $r = |\mathbf{r}|$, i.e. distance from the source. The fraction $\tilde{a}/r$ reflects the fact that the surface of the propagating spherical surface grows quadratically while energy radiated by the source is constant. Therefore the intensity has to decrease quadratically with distance from source, i.e $|\tilde{a}|^2/|\mathbf{r}|^2 = \tilde{a}/r$

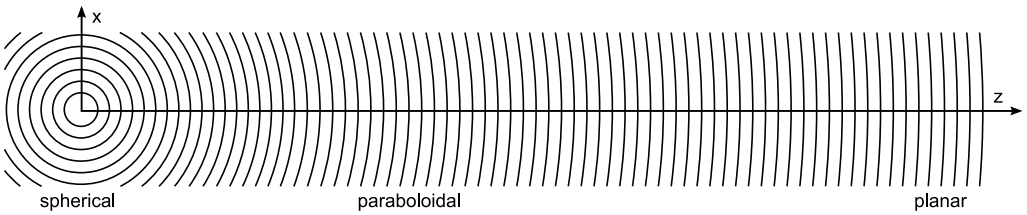


Figure 2.3: Relation between spherical and planar wave [Kra04].

Both spherical and planar waves are related to each other. This relation is illustrated in the Fig. 2.3. If an observation is done along the propagation direction the spherical wavefronts becomes gradually a planar wavefronts. 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 exploited by the Fresnel approximation, see Sec. 2.1.6. If the distance increases even further it is possible to approximate the spherical wave with the planar one. This is a mechanism exploited by the Fraunhofer approximation, see Sec. 2.1.6.

2.1.5 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 [Goo05] 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 Fig. 2.4 for reference. This principle was modified by A. Fresnel who stated [Goo05] 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 [Goo05].

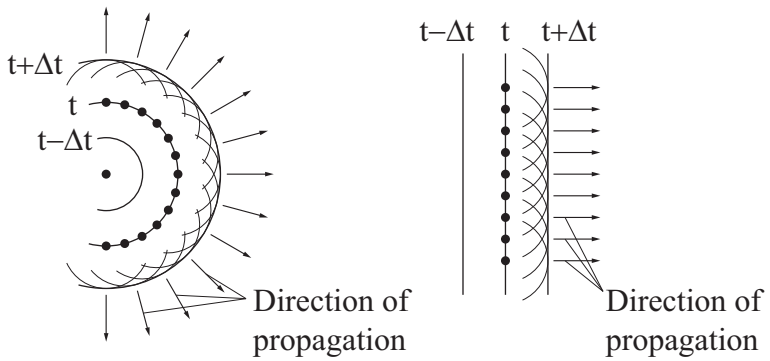


Figure 2.4: Huygens principle demonstrated on spherical (left) and planar (right) waves.

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 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 exist. 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 governs the behavior of both magnetic and electric field.

There is one other condition that further simplifies the diffraction model: considered diffraction structures are assumed to be large compared to the wavelength of the diffracted wave. All those simplifications and constraints turn the diffraction model into an approximation but even though the simplifications are significant ones they cause only small loss of accuracy and thus they are more then appropriate in many situations.

There are two fundamental descriptions of the diffraction: the Kirchhoff formulation and the Rayleigh-Sommerfeld formulation [Goo05, LBL02]. Both formulations describe the light field in front of a screen or an aperture properly and accurately. Nevertheless, the Kirchhoff formulation has a certain limitation as it fails to provide correct result if the examined point is closer to the screen than

the distance of several wavelengths. Also, it assumes that the field just behind the aperture is zero and this is in contradiction with the physical experiments. Despite its limitations, the Kirchhoff formulation is widely used in practice.

The Kirchhoff formulation of diffraction is based on the integral theorem of Helmholtz and Kirchhoff. The Helmholtz and Kirchhoff theorem 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 Eq. 2.11.

While the Helmholtz equation describes behavior of waves, the Green's theorem defines a relation between two complex functions $\tilde{u}(\mathbf{p})$ and $\tilde{g}(\mathbf{p})$ of position $\mathbf{p}$. Let S be a closed surface surrounding a volume V. If $\tilde{u}$, $\tilde{v}$ and their first and second partial derivatives in the inward normal direction are single-valued and continuous within and on S then:

$$-\iint_S \frac{\partial \tilde{u}}{\partial \mathbf{n}}\tilde{g} - \tilde{u}\frac{\partial \tilde{g}}{\partial \mathbf{n}}\,\mathrm{d}s = \iiint_V \tilde{g}\nabla^2\tilde{u} - \tilde{u}\nabla^2\tilde{g}\,\mathrm{d}v. \tag{2.28}$$

The functions $\tilde{u}(\mathbf{p})$ and $\tilde{g}(\mathbf{p})$ provide twice continuously differentiable scalar fields mappings between V and S. If both functions satisfy the Helmholtz equation Eq. 2.11 then [LBL02]:

$$-\iint_S \frac{\partial \tilde{u}}{\partial \mathbf{n}}\tilde{g} - \tilde{u}\frac{\partial \tilde{g}}{\partial \mathbf{n}}\,\mathrm{d}s = 0 \tag{2.29}$$

The goal of the diffraction formulation is do find a field at a point $\mathbf{p}_0$, see Fig. 2.5. For such purpose Kirchhoff formulation uses a boundary surface on one side of the aperture. This boundary surface consists of two parts: a plane S_p close to the aperture including its transparent portion Σ and a spherical surface S_ϵ. Next, a couple of boundary condition known as Kirchhoff boundary conditions are introduced in order to simplify the result. The boundary conditions describe behavior of field $\tilde{u}$ in close neighbourhood to the screen.

The first condition 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 condition 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 simplify the equation by replacing one part of the enclosing surface just with the surface Σ. Refer to [Goo05] for more details.

The influence of the spherical surface vanishes as the radius of the spherical surface increases towards infinity [BW05]. By applying the boundary conditions the expression Eq. 2.29 is simplified to:

$$\tilde{u}(\mathbf{p}_0) = \frac{1}{4\pi}\iint_\Sigma \left(\frac{\partial \tilde{u}}{\partial \mathbf{n}}\tilde{g} - \tilde{u}\frac{\partial \tilde{g}}{\partial \mathbf{n}}\right)\,\mathrm{d}s, \tag{2.30}$$

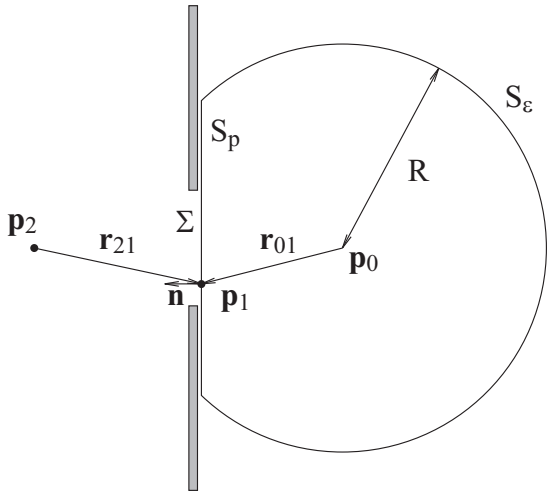


Figure 2.5: Kirchhoff formulation of diffraction by a plane screen. [Goo05]

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

A further simplification of Eq. 2.30 is based on exploiting a proper Green's function instead of the function $\tilde{g}$ in the expression Eq. 2.29. One 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 Fresnel-Kirchhoff diffraction formula:

$$\tilde{u}(\mathbf{p}_0) = \frac{\tilde{a}}{\mathrm{i}\lambda}\iint_\Sigma \frac{\exp[\mathrm{i}k(r_{21}+r_{01})]}{r_{21}r_{01}}\left[\frac{\cos(\mathbf{n},\mathbf{r}_{01}) - \cos(\mathbf{n},\mathbf{r}_{21})}{2}\right]\mathrm{d}s, \qquad (2.31)$$

where $\tilde{a}$ is a complex amplitude 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 $\mathbf{b}$. The vectors $\mathbf{r}_{01}$ and $\mathbf{r}_{21}$ are 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 $\mathbf{p}_2$.

More common formulation of the Fresnel-Kirchhoff formula can be obtained by reorganising and substituting to Eq. 2.31 accordingly:

$$\tilde{u}(\mathbf{p}_0) = \frac{\tilde{a}}{\mathrm{i}\lambda} \iint_{\Sigma} \tilde{u}'(\mathbf{p}_1) \frac{\exp(\mathrm{i}kr_{01})}{r_{01}} \,\mathrm{d}s. \tag{2.32}$$

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 light and this interpretation plays an important role in numerical reconstruction of the hologram, see below. Refer to [Goo05] for more details on the Kirchhoff formulation.

The Sommerfeld-Rayleigh formulation is a further enhancement of the Kirchhoff formulation that removes inconsistent boundary conditions mentioned earlier. It removes the boundary condition from the function $\tilde{u}$ by assuming that either $\tilde{g}$ or $\partial\tilde{g}/\partial\mathbf{n}$ in Eq. 2.30 vanishes on a portion of the boundary surface close to the aperture according to a proper definition of alternate Green's function, see below. Alike the Kirchhoff solution, it assumes that the screen is planar.

In order to fulfil this assumption the formulation uses a second point $\mathbf{p}_0{}'$ that is mirror image of the point $\mathbf{p}_0$. At this mirrored location a point source of the same wavelength as the original one is positioned. Both wave sources oscillate with π phase difference. For the π phase difference, a formula that describes the field commonly known as the first Rayleigh-Sommerfeld solution is:

$$\tilde{u}(\mathbf{p}_0) = \frac{-1}{4\pi} \iint_{\Sigma} \tilde{u} \frac{\partial\tilde{g}_-}{\partial\mathbf{n}} \,\mathrm{d}s, \tag{2.33}$$

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. If $|\mathbf{p}_0 - \mathbf{p}'_0| \gg \lambda$ is assumed then 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 applying this approximation to the alternate Green's function $\tilde{g}_-$ and by the fact that $\tilde{g}_-$ vanishes on the transparent portion Σ of planar screen it is possible to obtain two formulae known as the Rayleigh-Sommerfeld diffraction formula. For

more details on derivation of these formulae refer to [Goo05, LBL02, Mie02]. The first configuration that has a π phase difference in phases is:

$$\tilde{u}_I(\mathbf{p}_0) = \frac{\tilde{a}}{\mathrm{i}\lambda} \iint_{\Sigma} \frac{\exp[\mathrm{i}k(r_{21} + r_{01})]}{r_{21} r_{01}} \cos(\mathbf{n}, \mathbf{r}_{01})\, \mathrm{d}s. \tag{2.34}$$

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

$$\tilde{u}_{II}(\mathbf{p}_0) = -\frac{\tilde{a}}{\mathrm{i}\lambda} \iint_{\Sigma} \frac{\exp[\mathrm{i}k(r_{21} + r_{01})]}{r_{21} r_{01}} \cos(\mathbf{n}, \mathbf{r}_{21})\, \mathrm{d}s. \tag{2.35}$$

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

2.1.6 Wave propagation

The propagation of the wave in a free space plays an important role in the hologram reconstruction and synthesis. Propagation determines the final configuration of a light field at a hologram frame due to a given scene radiation. The consequent propagation in a reversed direction can reconstruct the recorded scene form a hologram. The propagation is driven by the diffraction formulae described in the Sec. 2.1.5.

The propagation of a wave is carried out by a change of phase. At this point is important to specify whetter the propagation is done by adding or subtracting a value from the phase. Since the time dependent component $\exp(-\mathrm{i}\omega t)$ of the wavefunction Eq. 2.9 rotates in a clockwise direction the 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 minimise the confusion from signs of phases, it is assumed that the propagation of the wave is examined in a direction of a positive Z-axis. This also simplifies equation for propagation in a direction parallel to the Z-axis. If other configurations are required then it is always possible to transform them so that eventually the propagation is done in a direction parallel to the Z-axis.

Huygens-Fresnel principle

The Huygens-Fresnel principle states that the wavefront of a disturbance in a time $t+\Delta t$ is an envelope of wavefronts of secondary sources emanating from each point

of the wavefront in a time t. 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 [Goo05]. This Huygens-Fresnel principle is supported by both Kirchhoff and Rayleigh-Sommerfeld diffraction formulae. Using the first Rayleigh-Sommerfeld solution from Eq. 2.34, the Huygens-Fresnel principle is [LBL02]:

$$\tilde{u}_I(\mathbf{p}_0) = \frac{1}{\mathrm{i}\lambda} \iint_\Sigma \tilde{u}(\mathbf{p}_1) \frac{\exp(\mathrm{i}kr_{01})}{r_{01}} \cos\theta \, \mathrm{d}s, \tag{2.36}$$

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 phases are $\pi/2$ before the phase of the original wave due to the 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 Eq. 2.36 is that it is basically a convolution integral and it can be expressed as:

$$\tilde{u}(\mathbf{p}_0) = \iint_\Sigma \tilde{h}(\mathbf{p}_0, \mathbf{p}_1) \tilde{u}(\mathbf{p}_1) \, \mathrm{d}s, \tag{2.37}$$

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

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

Fresnel and Fraunhofer approximation

As noted in the previous section, it is possible to express the Huygens-Fresnel principle in terms of the first Rayleigh-Sommerfeld solution. The angle θ is subtended by 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 assuming the situation depicted in the Fig. 2.6 it is possible to express the Huygens-Fresnel principle Eq. 2.36 as following:

$$\tilde{u}(x, y) = \frac{z}{\mathrm{i}\lambda} \iint_\Sigma \tilde{u}(\xi, \eta) \frac{\exp(\mathrm{i}kr_{01})}{r_{01}^2} \, \mathrm{d}\xi \, \mathrm{d}\eta, \tag{2.38}$$

where $r_{01} = [z^2 + (x - \xi)^2 + (y - \eta)^2]^{1/2}$. All following approximations simplify the expression for r_{01} because it contains square root function that does not allow exploiting the Fourier transform that greatly reduces the computation complexity of the expression.

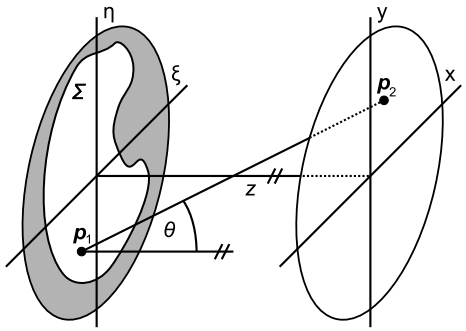


Figure 2.6: Configuration of the Fresnel/Fraunhofer approximation [Goo05].

The first approximation is known as the Fresnel approximation. It is based on the Z-axis component's value of the vector $\mathbf{r}_{01}$. If this value is large enough it is possible to apply the binomial expansion for the square root function. In order to apply the expansion, the expression for the distance r_{01} has to be rewritten first as:

$$r_{01} = z\left[1 + \left(\frac{x-\xi}{z}\right)^2 + \left(\frac{y-\eta}{z}\right)^2\right]^{1/2}. \tag{2.39}$$

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

$$\sqrt{1+b} = 1 + \frac{1}{2}b - \frac{1}{8}b^2 + \ldots. \tag{2.40}$$

If the extent in Z-axis is greater than the extent in both X-axis and Y-axis, i.e. $z \gg (x-\xi)^2 + (y-\eta)^2$, then it is possible to apply the approximation by keeping only the first two component of the binomial expansion Eq. 2.40. Such approximation does not cause a problem in a case of the denominator r_{01}^2 in Eq. 2.38 but severe error may occur for r_{01} appearing in the exponent because the wave propagation is sensitive to the phase [MNF+02]. The error stays under control if the distance z fulfils the following condition:

$$z^3 \gg \frac{k}{8}[(x-\xi)^2 + (y-\eta)^2]^2_{\max}, \tag{2.41}$$

where $k = 2\pi/\lambda$ is a wavenumber. Note that this condition requires a minimal distance of $z \gg 250$ mm for a circular aperture of diameter 10 mm while observing it from a region of size 1 cm. The wavelength in this case is $\lambda = 0.5$ μm. If z fulfils the condition Eq. 2.41 the viewer is denoted to be in a near field or Fresnel region.

The condition mentioned above 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 shorter distances than distances enforced by the condition Eq. 2.41.

Nevertheless, by fulfilling the condition it is possible to substitute the Eq. 2.39 to the first two components of the binomial expansion Eq. 2.40 thus obtaining:

$$r_{01} \approx z + \frac{(x-\xi)^2 + (y-\eta)^2}{2z}.. \tag{2.42}$$

Putting Eq. 2.42 and Eq. 2.38 together one obtains this resulting expression known as the Fresnel diffraction integral:

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

where $\tilde{u}_0(\xi,\eta)$ is a diffraction pattern at the aperture. This expression can be rewritten into a form that resembles the Fourier transform. The fast Fourier transform – FFT can be exploited for evaluating thus reducing the computational complexity significantly:

$$\begin{aligned}\tilde{u}_z(x,y) =& \frac{\exp(\mathrm{i}kz)}{\mathrm{i}\lambda z} \exp\left(\mathrm{i}k\frac{x^2+y^2}{2z}\right) \\ &\times \iint_{-\infty}^{\infty} \left\{\tilde{u}_0(\xi,\eta) \exp\left(\mathrm{i}k\frac{\xi^2+\eta^2}{2z}\right)\right\} \exp\left(-\mathrm{i}k\frac{x\xi+y\eta}{z}\right) \mathrm{d}\xi\,\mathrm{d}\eta.\end{aligned} \tag{2.44}$$

If the distance z is increased even more it is possible to apply another approximation that consist in neglecting another components of the integral. Such approximation is known as the Fraunhofer approximation and it is applicable only if:

$$z \gg \frac{k(\xi^2+\eta^2)_{\max}}{2}. \tag{2.45}$$

This condition has even higher demands on the distance z than the condition Eq. 2.41 has. For a circular aperture with a diameter of 10mm and for a wave with wavelength 0.5μm the distance z has to be $z \gg 300$m. If z fulfils the condition Eq. 2.45 the viewer at that distance is denoted to be in far field or Fraunhofer region. By satisfying the condition of the Fraunhofer region the Eq. 2.44 is reduced to the following:

$$\begin{aligned}\tilde{u}_z(x,y) =& \frac{\exp(\mathrm{i}kz)}{\mathrm{i}\lambda z} \exp\left(\mathrm{i}k\frac{x^2+y^2}{2z}\right) \\ &\times \iint_{-\infty}^{\infty} \tilde{u}_0(\xi,\eta) \exp\left(-j2\pi\frac{x\xi+y\eta}{\lambda z}\right) \mathrm{d}\xi\,\mathrm{d}\eta.\end{aligned} \tag{2.46}$$

The integral in the Eq. 2.46 resembles a Fourier transformation of the aperture distribution $\tilde{u}_0$. If a normalised intensity is the final product of a propagation then the Fraunhofer approximation constitutes only Fourier transform of $\tilde{u}_0$. The multiplicative phase factors are not applied as they have no influence on the intensity, see Eq. 2.12. The Fourier transform is evaluated at frequencies:

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

So far, the Fresnel approximation assumed 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 is capable of handling tilted planes such as that depicted in the Fig. 2.7 as well [YAC02].

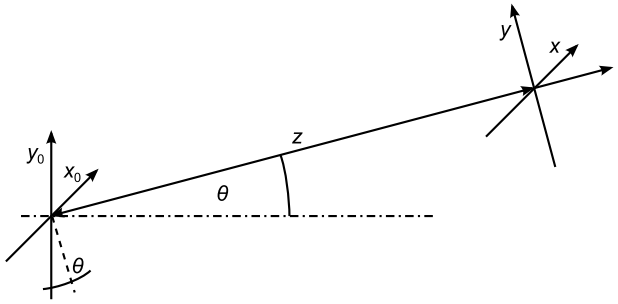


Figure 2.7: Tilted plane configuration for application of Fresnel approximation [YAC02].

The approach is based on simplifying 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 Fig. 2.7 then the expression for the distance required for the Rayleigh-Sommerfeld integral is:

$$r = \left[(y_0 \sin\theta - z)^2 + (x_0 - x)^2 + (y_0 \cos\theta - y)^2\right]^{1/2}. \tag{2.47}$$

By introducing the $r' = (x^2+y^2+z^2)^{1/2}$ to the Eq. 2.47 the binomial expansion can be applied. After the expansion, it is possible to neglect all terms but the first two as it is done in a case of the Fresnel approximation. The resulting expression is substituted to this 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 fulfilled. By reorganising and further substitution the Eq. 2.48 is obtained that resembles the Fourier transform:

$$\begin{aligned}\tilde{u}_{z,\theta}(\xi, \eta) = \exp(\mathrm{i}kr') \iint_{\Sigma} \tilde{u}_0(x_0, y_0) \exp\left(\mathrm{i}k\frac{x_0^2 + y_0^2}{2z}\right) \\ \exp\left[-\mathrm{i}2\pi(\xi x_0 + \eta y_0)\right] \mathrm{d}x_0\, \mathrm{d}y_0,\end{aligned} \tag{2.48}$$

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

Diffraction condition and diffraction orders

A term diffraction condition regards to a planar wave diffracted on a thin cosine grating [Goo05, Kra04]. The cosine grating is a thin planar cosine amplitude grating with a amplitude transmittance function:

$$t_A(\xi,\eta) = \exp\left[\frac{1}{2} + \frac{m}{2}\cos\left(2\pi\frac{\xi}{\Lambda_\xi}\right)\right]\operatorname{rect}\left(\frac{\xi}{2w}\right)\operatorname{rect}\left(\frac{\eta}{2w}\right), \tag{2.49}$$

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

If such grating is illuminated by a unit amplitude planar wave then diffraction occurs. By applying the convolution theorem to t_A it is possible to obtain the Fourier transform of t_A that can be utilised to get the grating's Fraunhofer diffraction pattern, i.e. a diffraction pattern in the far field:

$$\begin{aligned}\tilde{u}(x,y) =& \frac{a}{\mathrm{i}2\lambda z}\exp\left[\mathrm{i}kz + \mathrm{i}\frac{k}{2z}(x^2+y^2)\right]\operatorname{sinc}\left(\frac{2wy}{\lambda z}\right)\\ &\times\left\{\operatorname{sinc}\left(\frac{2wx}{\lambda z}\right) + \frac{m}{2}\operatorname{sinc}\left[\frac{2w}{\lambda z}\left(x+\frac{\lambda z}{\Lambda_\xi}\right)\right] + \frac{m}{2}\operatorname{sinc}\left[\frac{2w}{\lambda z}\left(x-\frac{\lambda z}{\Lambda_\xi}\right)\right]\right\}\end{aligned} \tag{2.50}$$

The intensity of the diffraction pattern $\tilde{u}$ is its magnitude squared. This means that the intensity of the Eq. 2.50 is a sum of squared sinc functions. The intensity is depicted in the Fig. 2.8.

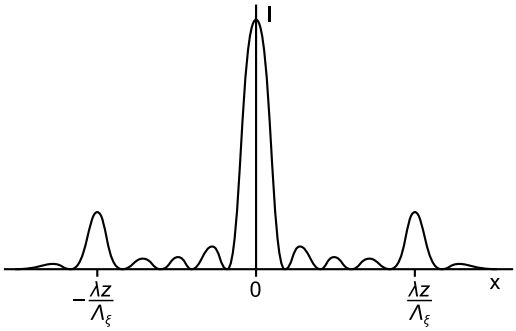


Figure 2.8: Intensity of a Fraunhofer diffraction pattern due to a thin amplitude cosine grating [Goo05].

Peaks in the Fig. 2.8 constitute so called diffraction orders and they represent energy deflected by the aperture. The largest central peak is called the zero order and represents the undiffracted wave, i.e. a wave having the same direction as the incident one. Zero order also contains the greatest amount of energy of the original incident planar wave. The two second largest peaks on sides are called

the first orders and represent planar waves diffracted according to the diffraction condition, see below.

The incident energy is divided into individual diffraction orders. The fraction for each order is determined as the squared coefficient of the delta function that appear in the Fourier transform of the expression Eq. 2.49. Note that while the sinc functions in the Fourier transform of Eq. 2.49 spreads the energy, the delta function determine the power in each order. The zero order obtains $1/4 = 25.0\,\%$ of the incident wave power, 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 diffracted as the first order is known as the diffraction 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 Eq. 2.49. This kind of grating also produces higher orders then the firsts. The approach for deriving the respective diffraction order's energy is similar to the amplitude grating so it will not be included here. Refer to [Goo05] for more details.

Basically, each diffraction order is the original incident planar wave that is propagated to a different direction and is carrying different amount of energy. The direction of the propagation for a given diffraction order is determined from the optical path difference of individual "rays" [Goo05]. The difference for a given order has to be an integer multiple 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:

$$\sin\theta_{\xi 2} = \sin\theta_{\xi 1} + q\frac{\lambda}{\Lambda}, \tag{2.51}$$

where $\theta_{\xi 1}$ is an angle of the incident wave, $\theta_{\xi 2}$ is an angle of the diffracted wave for given order $\mathbf{q}$, and Λ is the period of the grating, see Fig. 2.9.

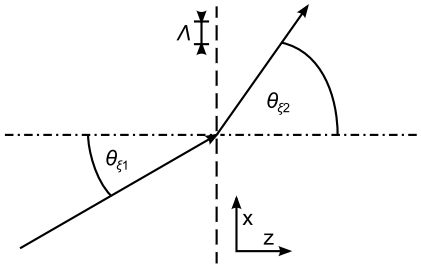


Figure 2.9: Diffraction grating and diffraction condition.

Propagation in Angular Spectrum

The Fresnel approximation allows computing a propagation of a diffraction pattern but it has restrictions on distance according to the condition Eq. 2.41. 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 diffracting plane with known diffraction pattern distribution $\tilde{u}_0$ unless the third or higher components of the binomial expansion are taken into account.

For shorter distances the Rayleigh-Sommerfeld diffraction integral, see Eq. 2.34, has to be employed. Unfortunately, an implementation of this integral has an unpleasantly high computation complexity and thus rendering this approach almost useless. However, a slightly different approach can be formulated using a Fourier transform of the diffraction pattern. The frequencies obtained from the Fourier transform are called the angular spectrum [EO06, Goo05, TB93].

If an angular spectrum $\tilde{F}(k_x, k_y)$ of a diffraction pattern is known then it is possible to determine a field $\tilde{u}$ at a given point $\mathbf{p}$ as:

$$\tilde{u}(\mathbf{p}) = \iint \tilde{F}(k_x, k_y) \exp(-\mathrm{i}\mathbf{k} \cdot \mathbf{p}) \,\mathrm{d}k_x \,\mathrm{d}k_y, \tag{2.52}$$

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 an evanescent wave [BW05]. Such wave is propagated as well but its amplitude decays exponentially with increasing $|z|$. The exponential nature of attenuation is apparent from substitution of complex-valued k_z to the Eq. 2.52.

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

$$\tilde{u}_0(x, y) = \iint \tilde{F}(k_x, k_y) \exp[-\mathrm{i}(k_x x + k_y y)] \,\mathrm{d}k_x \,\mathrm{d}k_y. \tag{2.53}$$

From Eq. 2.53 is apparent that the angular spectrum $\tilde{F}$ is proportional to the Fourier transform of the diffraction pattern $\tilde{u}_0$ on the plane ϱ. More precisely, that $\tilde{F} = \mathcal{F}\{\tilde{u}_0\}/(2\pi)^2$.

From a knowledge of some diffraction pattern's angular spectrum it is possible to predict the diffraction pattern at an arbitrary distance along the Z-axis. Waves are propagated according to the Eq. 2.52. If the phase shift term $\exp(-\mathrm{i}\mathbf{k} \cdot \mathbf{p})$ is expanded properly then it is possible to obtain an expression that resembles the Fourier transform as well. Note that individual wavevector components contains $2\pi/\lambda$ since the length of a wavevector is the wavenumber.:

$$\tilde{u}(\mathbf{p}) = \iint \left\{ \tilde{F}(k_x, k_y) \exp(-\mathrm{i}k_z z) \right\} \exp[-\mathrm{i}(k_x x + k_y y)] \,\mathrm{d}k_x \,\mathrm{d}k_y. \tag{2.54}$$

According to the Eq. 2.54 the diffraction pattern $\tilde{u}$ on a plane that is parallel to the source plane and is located at the distance z along positive Z-axis is computed

by employing this expression:

$$\tilde{u} = \mathcal{F}^{-1}\left\{\mathcal{F}\left\{\tilde{u}_0\right\}\exp(-\mathrm{i}k_z z)\right\}, \tag{2.55}$$

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

This approach can be further modified so it is capable of handling propagation between two planes that are spatially shifted in the XY-plane and tilted as well. This is achieved by applying simple geometrical operation of rotation written in a form of a 3×3 matrix $\mathbf{R}$ and a translation written in a form of a vector $\mathbf{b}$:

$$\mathbf{p}' = \mathbf{p}\mathbf{R} + \mathbf{b}. \tag{2.56}$$

By substituting the expression for a point $\mathbf{p}$ on a target plane into the Eq. 2.52 and reorganising the result, an expression for distribution $\tilde{u}$ on a target plane is obtained:

$$\tilde{u} = \frac{1}{4\pi^2}\mathcal{F}^{-1}\left\{4\pi^2\mathcal{F}\left\{\tilde{u}_0\right\}\exp[\mathrm{i}(\mathbf{k}\mathbf{R})\cdot\mathbf{b}]J(k_z, k_z')\right\}, \tag{2.57}$$

where k_z' is a Z-axis component of the wavevector $\mathbf{k}$ transformed by the matrix $\mathbf{R}$ and function $J(k_z, k_z') = k_z/k_z'$, which is a Jacobian correction factor. The transformation of the wavevector $\mathbf{k}$ by the matrix $\mathbf{R}$ equals to a shifting a portion of hemispherical surface over the hemispherical surface, see Fig. 2.10, because endpoints of all possible wavevectors excluding wavevectors for evanescent waves form a hemisphere of radius equal to a wavenumber.

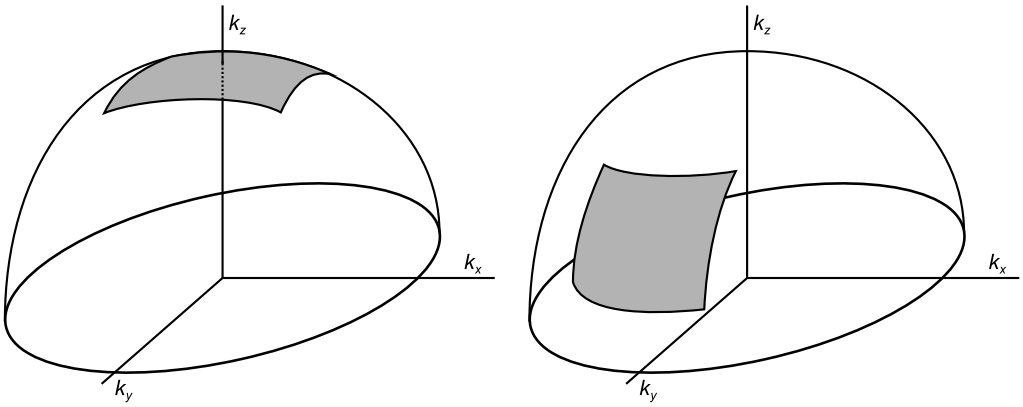


Figure 2.10: Original (left) and transformed (right) distribution of plane waves obtained by Fourier transform.

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 a proper propagation as mentioned in [TB93], so that miss in overlap does not occur at all.

2.2 Optical holography

The holography in general is quite wide scientific area. It can be divided into several groups. Each group have a different goal according to the primary application. For this work the relevant branch of holography is the one dealing with realistic image recording and reproduction. This branch can be further divided into two sub branches – optical holography and digital holography.

The optical holography is a real thing that needs actual lasers and photographic materials for recording and reproducing holographic images. Optical holography stands on the physical phenomenon of diffraction described in the Sec. 2.1.5. On the other hand, the digital holography is a numerical alternation of the optical holography. It replaces the real diffraction with a numerical simulation of the diffraction to obtain holograms. It also replaces static photographic material with computer driven electronic spatial light modulators for reproducing the holographic images.

This section describes the principle of the optical holography. Digital holography is described later on in the Sec. 2.3.

2.2.1 Holography principle

Holography is about recording and reproducing a light field. A light field is at each point determined by an amplitude and phase. In a case of classical photography the light field is focused onto a photographic material and incident energy is integrated over some time. The photographic material reacts accordingly and exposed places become dark. This process therefore records the intensity information but omits the direction information. 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, i.e. both intensity and direction. If a hologram is illuminated, the recorded amplitude and phase is reconstructed and the originally recorded light field is recreated. Since an observer has the whole light field available, the genuine three dimensional sensation is achieved, see Fig. 2.11.

Holography uses almost the same materials for recording as the photography and therefore the phase and amplitude cannot be recorded directly but rather in an encoded form - in a form of diffraction grating. The diffraction grating is formed by fringes produced from interference of the reference beam and the scattered beam reflected from a recorded scene. The interference fringes inflict variation of intensity across the recording medium. This variation of intensity is transformed into variation of transparency which effectively forms the wanted diffraction grating. Diffraction gratings diffract or in other words bend light, see Sec. 2.1.5, and most fortunately, or rather because of the physics of diffraction, one component of the diffracted light field matches the initially recorded one, see

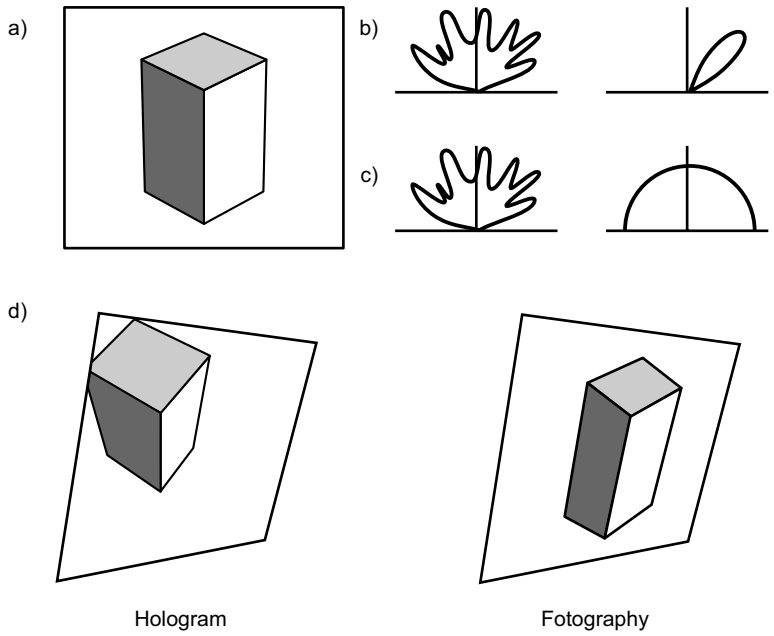


Figure 2.11: Difference between hologram and a common photo for a scene (a): a difference between incoming intensity from various direction for a given sample (b) and an outgoing intensity (c) for the same sample leads to a different resulting image for a tilted viewing screen (d).

Fig. 2.12.

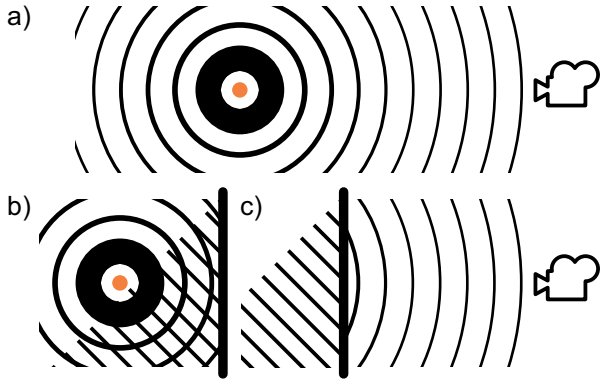


Figure 2.12: The essential principle of optical recording (b) and reconstructing (c). The reconstruction (c) generates light field that resembles the original scene (a).

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 field from the unwanted part is a research branch on its own in holography. One of the most common solution is recording hologram using the

off-axis configuration, see Sec. 2.2.3 for more details.

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

2.2.2 Inline hologram

The first hologram recording was done using the setup depicted in the Fig. 2.13. 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.

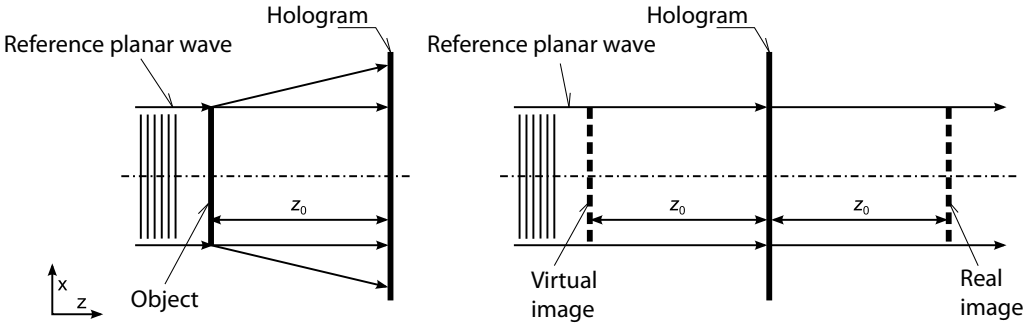


Figure 2.13: This is simplified depiction of capturing and reconstructing inline hologram. Adopted from [Har96].

This kind of setup works only for scenes that are 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 recorded scene and this scattered light then acts as the scene beam. Because of this restriction the inline configuration is used usually for recording scenes like aerosol, small particles floating in water etc. Scenes with the opposite characteristic, i.e. small holes in some opaque screen, are impossible to capture.

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 $\tilde{o}(x, y)$, where $|\tilde{o}(x, y)| \ll r$. 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 towards the scene.

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 Eq. 2.14 as

$$\begin{aligned} I(x,y) &= |r + \tilde{o}(x,y)|^2, \\ &= r^2 + |\tilde{o}(x,y)|^2 + r\tilde{o}(x,y) + r\tilde{o}^*(x,y), \end{aligned} \tag{2.58}$$

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

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

$$\mathbf{t} = \mathbf{t}_0 + \beta T I, \tag{2.59}$$

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 2.58 is substituted into 2.59 the amplitude of this transparency is

$$\mathbf{t}(x,y) = \mathbf{t}_0 + \beta T \left[r^2 + |\tilde{o}(x,y)|^2 + r\tilde{o}(x,y) + r\tilde{o}^*(x,y) \right]. \tag{2.60}$$

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

$$\begin{aligned} \tilde{u}(x,y) &= r\mathbf{t} \\ &= r\left(\mathbf{t}_0 + \beta T r^2\right) + \beta T r |\tilde{o}(x,y)|^2 \\ &\quad + \beta T r^2 \tilde{o}(x,y) + \beta T r^2 \tilde{o}^*(x,y). \end{aligned} \tag{2.61}$$

The expression Eq. 2.61 consists of four terms. The frist term $r\left(\mathbf{t}_0 + \beta T r^2\right)$ constitutes the directly transmitted beam. The second term $\beta T r |\tilde{o}(x,y)|^2$ is extremely small in comparison with the others since it has been assumed initially that $|\tilde{o}(x,y)| \ll r$ and can be therefore neglected. The third term $\beta T r^2 \tilde{o}(x,y)$ 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 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 superposes 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.

2.2.3 Off axis hologram

The problem of superposing 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 and served as the reference beam. This more complex configuration is depicted in the Fig. 2.14.

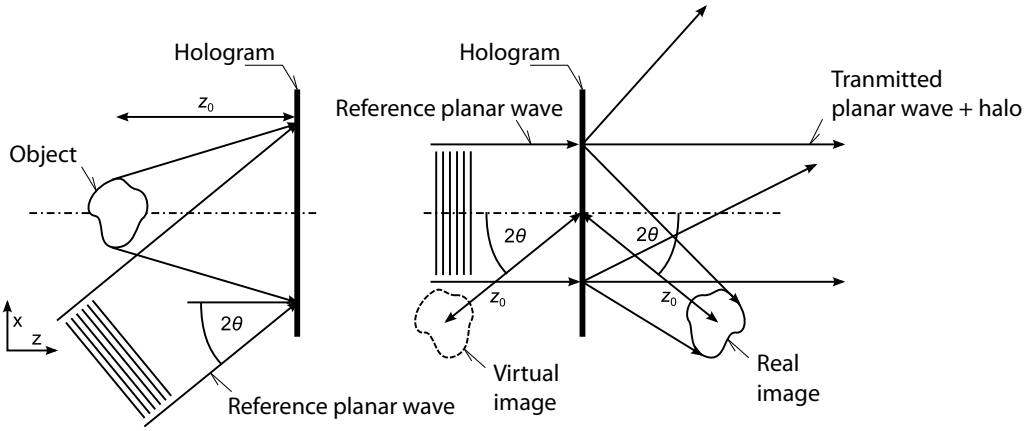


Figure 2.14: This is simplified depiction of recording (left) and reconstructing (right) offaxis hologram. Adopted from [Har96].

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

$$\tilde{o}(x,y) = |\tilde{o}(x,y)| \exp\left[-i\phi(x,y)\right], \tag{2.62}$$

while that due to the reference beam is

$$\tilde{r}(x,y) = r \exp\left(i2\pi\xi_r x\right), \tag{2.63}$$

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

$$\begin{aligned} I(x,y) &= |\tilde{r}(x,y) + \tilde{o}(x,y)|^2 \\ &= |\tilde{r}(x,y)|^2 + |\tilde{o}(x,y)|^2 \\ &\quad + r\,|\tilde{o}(x,y)| \exp\left[-i\phi(x,y)\right] \exp\left(-i2\pi\xi_r x\right) \\ &\quad + r\,|\tilde{o}(x,y)| \exp\left[i\phi(x,y)\right] \exp\left(i2\pi\xi_r x\right) \\ &= r^2 + |\tilde{o}(x,y)|^2 + 2r\,|\tilde{o}(x,y)| \cos\left[2\pi\xi_r x + \phi(x,y)\right]. \end{aligned} \tag{2.64}$$

The amplitude transmittance of the hologram can be written as

$$\begin{aligned}\mathbf{t}(x,y) = \mathbf{t}_0 + \beta T\{&|\tilde{o}(x,y)|^2 \\ &+ r\,|o(x,y)|\exp[-i\phi(x,y)]\exp(-i2\pi\xi_r x) \\ &+ r\,|\tilde{o}(x,y)|\exp[i\phi(x,y)]\exp(i2\pi\xi_r x)\}\end{aligned} \tag{2.65}$$

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

$$\begin{aligned}\tilde{u}(x,y) &= \tilde{r}(x,y)\,\mathbf{t}(x,y), \\ &= \tilde{u}_1(x,y) + \tilde{u}_2(x,y) + \tilde{u}_3(x,y) + \tilde{u}_4(x,y)\end{aligned} \tag{2.66}$$

where

$$\tilde{u}_1(x,y) = \mathbf{t}_0\exp(i2\pi\xi_r x) \tag{2.67}$$

$$\tilde{u}_2(x,y) = \beta T r\,|\tilde{o}(x,y)|^2\exp(i2\pi\xi_r x) \tag{2.68}$$

$$\tilde{u}_3(x,y) = \beta T r^2\tilde{o}(x,y) \tag{2.69}$$

$$\tilde{u}_4(x,y) = \beta T r^2\tilde{o}^*(x,y)\exp(i4\pi\xi_r x) \tag{2.70}$$

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 na 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

$$\xi_r \geq 3\xi_{\max}, \tag{2.71}$$

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 offaxis hologram. More complex and interesting scenes can be recorded 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.

2.2.4 Additional hologram types

Apart from the configurations described in Sec. 2.2.2 and Sec. 2.2.3, there are several other hologram recording configurations. In this section these configurations are described:

- Fourier hologram
- Image hologram
- Fraunhofer hologram

Fourier hologram

Setup for capturing Fourier hologram is similar to the setup used for recording inline hologram. Additionally, a lens is introduced between hologram plane and recorded object. Lens is placed such that the hologram plane is in the back focal plane of the lens and the recorded object and reference light source is in the front focal plane of the lens.

The lens works in such way that the back focal plane contains Fourier transform of the light field in the front focal plane. The Fourier hologram is therefore formed from interference of Fourier transforms of recorded object and reference beam.

Utilising the same formalism used in the cases of inline and off axis holograms, the complex amplitude leaving the object plane is $\tilde{o}(x, y)$, its complex amplitude at the hologram plate located in the back focal plane of the lens is

$$\tilde{O}(\xi, \eta) = F\{\tilde{o}(x, y)\}. \tag{2.72}$$

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

$$\tilde{R}(\xi, \eta) = e^{-i2\pi\xi b}. \tag{2.73}$$

The intensity of the interference pattern produced by those two waves is therefore

$$\begin{aligned} I(\xi,\eta) =& 1 + \left|\tilde{O}(\xi,\eta)\right|^2 + \tilde{O}(\xi,\eta)\exp(i2\pi\xi b) + \\ & \tilde{O}^*(\xi,\eta)\exp(-i2\pi\xi b). \end{aligned} \tag{2.74}$$

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 $I(\xi,\eta)$, the complex amplitude of the wave transmitted by the hologram is

$$U(\xi,\eta) = \mathbf{t}_0 + \beta T I(\xi,\eta). \tag{2.75}$$

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

$$\begin{aligned} \tilde{u}(x,y) &= F\{U(\xi,\eta)\}, \\ &= (t_0 + \beta T)\,\delta(x,y) + \beta T\tilde{o}(x,y) \star \tilde{o}(x,y) \\ &\quad + \beta T\tilde{o}(x-b,y) + \beta T\tilde{o}^*(-x+b,-y). \end{aligned} \tag{2.76}$$

The wave corresponding to the first term on the right-hand side of Eq. 2.76 comes to a focus on the axis, while that corresponding to the 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 be identified only by the fact that it is inverted. More details can be found in [Har96].

Image holograms

Image holograms are not recordings of object but recording of the real image of the object. The lens is used to project the real image of the object. With such arrangement it is possible to position the hologram plate in such manner that the image of the object straddles the plate. The reconstructed image is then formed in the same position with respect to the hologram, so that part of the image appears to be in front of the hologram and the remainder is behind it.

Image can be reproduced from an image hologram using 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].

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, see Sec. 2.1.6, to form 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:

$$z_0 \gg \left(x_0^2 + y_0^2\right)/\lambda. \tag{2.77}$$

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

2.2.5 Final notes on optical holography

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

2.3 Digital Holography

Holography is widely used in many areas for various purposes. It is not therefore surprising that efforts to digitalise optical holography 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 recording, the reproduction and digital hologram synthesis. In the following material, all three areas of the digital holography are described in more detail.

2.3.1 Digital recording

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 optical sensor arrays have as much as tens millions of sensor elements. But although the resolution is important, the density is even more important to digital holography. The pitch between two sensors directly influences the frequencies that can be recorded and with larger frequencies recorded the reconstructed image has consequently a better quality.

2.3.2 Digital reproduction

The reproduction is more interesting but less completed area. The diffraction pattern of the incident reference beam has to be modulated according to the hologram recording. This is done by the spatial light modulator or SLM. SLM's are capable of changing phase of the light complex amplitude. If a digital hologram is used as the input for SLM, the recorded scene is reconstructed. There are numerous technologies that are exploited 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.

2.3.3 Hologram synthesis

Digital hologram synthesis is the main theme of this work. The synthetised hologram is not recording of an actual physical wave but rather it is a result of a numerical simulation of the diffraction and interference phenomenon. Digital hologram synthesis is an essential technology for the 3D holographic display development. The numerical nature of the synthesis process makes some parts of the hologram synthesis easier then recording the hologram by the optical means. However, there are disadvantages as well. The diffraction phenomenon in particular is very hard to simulate in a sense of computation time.

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 graphical processing units - 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 targeted 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 spec-

ifications. 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 Sec. 2.1.5. 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 Eq. 2.27. For each point at the hologram, distance r to the point source is computed and the Eq. 2.27 is evaluated. If scene contains more points, then the complex amplitude obtained from the Eq. 2.27 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 ray casting methods introduced in Chap. 3 and Chap. 4.

The basis for the ray casting method is casting of 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 combining the contributions from all primitives. This method has several unsolved problems that restrains the full exploitation of this methods.

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.

Detailed description of the approaches mentioned above plus several other can be found in the next Chap. 3.

2.3.4 Final notes on digital holography

For practical displaying 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 recorded scene inherent from 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 3

Hologram Synthesis Methods

In this chapter, the existing hologram synthesis methods are summarised and evaluated. The hologram synthesis is an important part of the digital holography. Ability of synthesising a hologram of an arbitrary objects and scenes is an important feature required for the holographic display to work. Hologram synthesis has the same importance for holographic displays as the computer graphics imaging - CGI has nowadays for the conventional two dimensional displays. The fundamental goal of a digital hologram synthesis is to compute such fringe pattern on a hologram surface that, once illuminated by an appropriate light source, provides a visually appealing virtual image of a virtual scene that originally served as a model and existed only as a mathematical description in a computer's memory.

In the whole chapter it is assumed that the input of all presented methods consists of a geometry, i.e triangular mesh, surface characteristics, i.e. color, texture, illumination model, and lighting specification, i.e. number, color and type of light sources. In addition, it is assumed that the output of all presented methods is two dimensional array of complex numbers that constitutes the state of a light field at the hologram frame due to the scene radiation only, i.e. no reference beam is added into the result. The array is referenced as diffraction pattern.

Although the final display device requires the fringe pattern as an input, the diffraction pattern is the output required from synthesis methods. The diffraction pattern is more general in the sense of the device independency. The various devices may work on a slightly different optical principles. Feeding them with the diffraction pattern instead of a hologram allow them to create the hologram according to their parameters. This is in most cases an easy operation consisting of simple addition of the reference beam. Some special hologram types need a bit more complicated treatment but it still represents a minor nuisance compared with the gain of device independency.

3.1 Synthesis methods overview

The computation of the diffraction pattern at the hologram plane is rather complicated for non-trivial scene content. The exact solution requires the evaluation of the diffraction formula Eq. 2.34 at every sampled point of the hologram plane. The principial complexity of such evaluation is, in general, $O(N^5)$ where N represents number of samples in one direction of the recorded cubical volume. For this reason, a lot of efforts have been put into finding out such approximation of the diffraction phenomenon that would reduce the complexity of the problem and at the same time wouldn't degrade the quality of the reconstructed image. Various methods achieved various degree of success in this problem.

What should be emphasised here is that the space for approximation is narrowed by the primary application of the synthetic holograms. In the context of this thesis, the primary application is three dimensional extension of the high quality imaging as it is known from the contemporary television or computer displays. Because of this application, the approximations that, for instance, neglect opacity or assumes only diffuse surfaces are not acceptable and methods that use such approximations are not considered as perspective since they are uncapable of delivering holograms that provides photorealistic images.

Another usual area for applying approximations is found in the issue of objects surface representation. Some methods approximate surfaces by such primitives that increase the efficiency of synthesis but restrict the possible shapes of the objects that can be represented accurately. Such methods are also unsuitable for the purposes of high quality imaging unless there is a possibility of their evolvement towards an exploitation of more flexible and universal primitive, e.g. triangle.

Quite large set of approaches for digital hologram synthesis were found in literature. The approaches exploit various techniques of synthesis but several generalised methodologies were extracted and they are presented in this section. The principle of each methodology is described along with the mathematical background if required. And finally, the advantages and disadvantages of the methodology are discussed and the suitability of each methodology are measured against these evaluation criteria sorted descending by their priority:

1. Photorealistic output,
2. Speed of evaluation,
3. Scalability of the algorithm,
4. Hardware implementation,
5. Scene description.

3.1.1 Diffraction integral evaluation

This methodology is inspired by the optical holography and it tries to simulate the process of diffraction of light from objects and interference of the diffracted beam and the reference beam as closely as possible. Such efforts are quite reasonable because if this principle works for optical holography then the numerical simulation should work as well.

However, an accurate simulation is not an easy task in general because the light field propagated from spatially rich objects is quite complex and analytical solution of the diffraction integral is not possible. Therefore, all approaches of this methodology, in some way, exploits the spherical wave equation Eq. 2.27. This equation describes behavior of a spherical monochromatic electromagnetic wave emitted by an infinitesimally small point source. Identically behaves light when diffracted by an infinitesimally small circular aperture. In the following text, the term *point source* will cover also the circular aperture. This approximation constitutes numerical evaluation of the diffraction integral.

Light field due to a point source can be computed directly in an analytic way. Point source can be therefore used as basic primitive because light field of more than one point sources is obtained according to the interference principle Eq. 2.13 which states that the final complex amplitude is obtained as a sum of complex amplitudes of the individual sources.

This is advantageous for scenes consisting of particles. Each particle is considered as a point source of the given intensity and its contribution to the final light field at the hologram plane is computed according to the spherical wave equation. Then appropriate reference wave is added and the intensity is computed to obtain the fringe pattern. The algorithm is summarised in Alg. 3.1.

foreach $\tilde{u}(x,y) \in U$ **do**
  $\tilde{u}(x,y) \leftarrow 0 + 0i$;
  foreach $p_S \in S$ **do**
    $\tilde{u}(x,y) \leftarrow \tilde{u}(x,y) +$ `SphericalWave(`p_S, x, y`)`;
  end
end

Algorithm 3.1: Spherical waves based hologram synthesis method.

At each sampled point $\tilde{u}(x,y)$ of the hologram frame U the contribution from all point sources p_S is evaluated and accumulated. It is a very simple algorithm. The obstacle in its wider exploitation is the number of point sources in S and samples in U. For very simple scenes consisting of several points the computation time is acceptable but such scenes have no use in the imaging applications.

Necessity of having the scene consisting of points only quite restricts the exploitation of this approach for scene consisting of a continuous geometry, i.e. triangular mesh. However, there is a whole area of classical CGI, called point

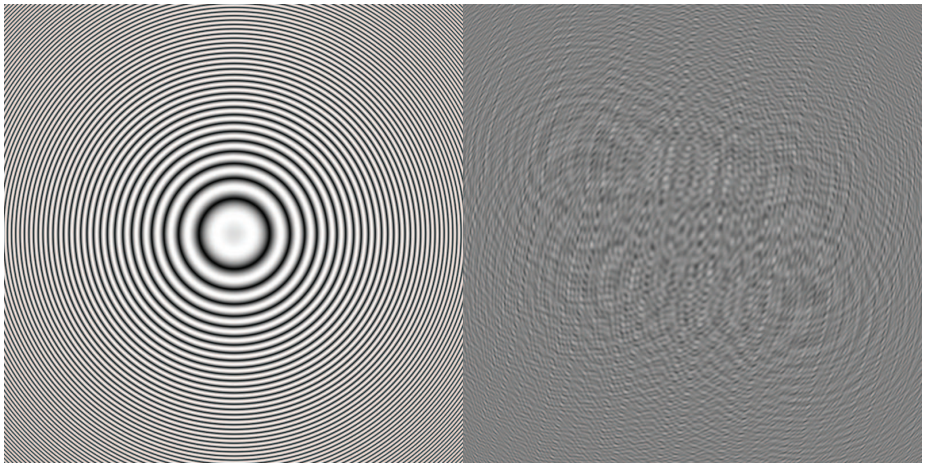


Figure 3.1: Inline holograms of a single point radiator on the left and several points (42) source radiators on the right.

rendering [GD98], which deals with this problem. The continuous surface can be densely populated with points and as the density of points increases and spaces between them decreases, the particle nature of the geometry disappears and the perception of a continuous surface is achieved.

Direct application of this approach is very inefficient in regard to the computation time. For example, the case where scene consists of 10000 points and hologram frame has resolution 1000×1000 requires 10^{10} evaluations of the spherical wave equation which involves square root and complex exponential function. This could take several hours to evaluate and 10000 points can accurately represent only the simplest geometry. Furthermore, the computation requirements could be much more demanding if illumination model evaluation and occlusion test are incorporated. Those two features are absolutely necessary for achieving realistic images.

To conclude, this approach is very simple in principle but at the cost of huge computational requirements. This is the reason, why it is addressed in literature so rarely and usually as an educative explanation of the synthesis process. However, the characteristics of the approach - a lot of simple operations done in parallel - fits to the scheme required for a successful implementation on specialised hardware so it may not be completely wrong approach especially when parallelised evaluation is easy to achieve in this case.

With regard to the evaluation criteria, this methodology has limited options in achieving photorealism. The occlusion solving and illumination model support is not straightforward to incorporate in the process. The speed of evaluation is not great compared to other methodologies and finally the scene has to be approximated by the point cloud which impose restriction on the scene description criterion. The evaluation criteria discussed until now were not fulfilled satis-

factorily. On the other hand, the methodology performs better in the last two evaluation criteria of scalability and hardware implementation. Methodology can be implemented as a parallel process which fits the hardware design.

3.1.2 Ray casting

In the previous section a quite simple principle of a hologram synthesis was introduced, however it is directly applicable only on scenes consisting of clouds of single points. This is due to the nature of the infinitesimal point source. It is omnidirectional with uniform intensity so the intensity is independent on a direction and it is infinitesimally small so it cannot be blocked by another source nor can block other sources so the occlusion is not considered. However, if more visually appealing images should be created, the lack of occlusion effect and uniform intensity distribution are no longer feasible.

However, the pitfalls related with the implementation of the occlusion effect and nonuniform intensity point sources are significant. It is because both occlusion and illumination models are view dependant and, as well known, each point on a hologram constitutes one viewpoint. This means that both occlusion and illumination must be reevaluated for each sampled point on a hologram frame. And that constitutes additional non-trivial processing for each point source radiator.

The more important feature from the advertised ones is occlusion. It is because the image containing only diffuse materials is more believable than the one without occlusion, which constitutes one of the major depth cues. Without occlusion a viewer can be completely confused. Solution for the problem of occlusion is well known from the field of CGI and there are several solutions available. However, not all of them suits the needs of the hologram synthesis.

One way of dealing with occlusion is the ray casting. The information about occlusion between two points is found out by casting a ray between those two points into the scene and if intersection is found between the ray and existing geometry the points do not see each other. This occlusion test has to be performed for all $[\mathbf{s}, \mathbf{u}]$ pair where $\mathbf{s} \in S$ is a point of a scene and $\mathbf{u} \in U$ is the point at the hologram plane. Occlusion test, which is quite complex operation, is therefore performed $S \times U$ times.

There is one problem related with the point nature of the scene. It is improbable that the infinitesimally small point will be intersected by any of the casted ray. The original geometry which served as a template for generating the point sources has to be therefore preserved and instead of computing the intersections of the ray with all the other point sources the intersection of the ray with the geometry is computed. The whole process is described in Alg. 3.2.

The ray casting method can be implemented in two different modes, see Fig. 3.2. The first mode iterates through all points constituting the scene and

```
foreach ũ ∈ U do
    foreach s ∈ S do
        l_{s,u} = Cast(Ray(s, u))
        if l_{s,ũ} == null then
            ũ ← ũ+ SphericalWave(s)
        end
    end
end
```

Algorithm 3.2: Ray casting method for synthesis with occlusion test.

casts rays from each of them towards each sample point of the hologram frame. This is the equivalent of the point based method and Alg. 3.2, however there are several pitfalls that comes with this methods. The first one is that the number of point samples influences the number of rays needed to cast. It is not so bad thing if there are several of them present but it becomes a bad thing if a lot of points are present. The second problem is the spatial uniformity of the point sources. If point sources are not distributed evenly across the surface, they need a weight coefficient which indicates how large surface they represent. Moreover the weight coefficient has to be corrected according to the projected area of the patch. This is not necessary if the second mode of ray casting approach is used.

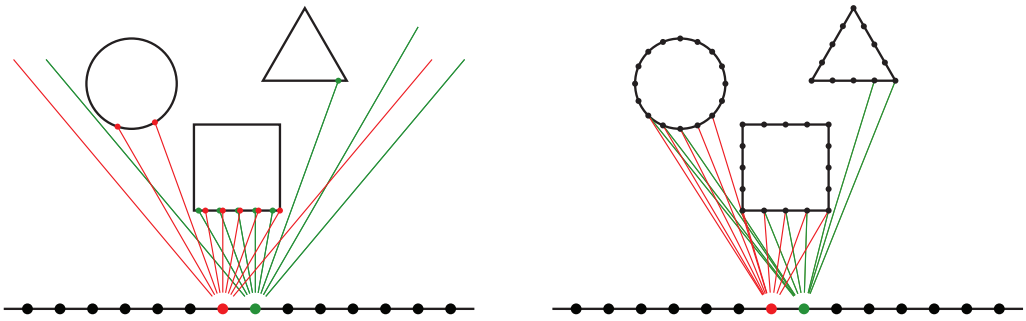


Figure 3.2: Two modes of ray the casting based approach operation.

The second mode of the ray casting method, see Fig. 3.2 does not have a static set of the point sources. Instead it generates point sources at the places of the intersections between the cone of rays originating at the currently computed sample point of the hologram frame and the geometry of the scene. In this mode, the number of rays is constant per sample point. The number of rays is independent on the geometry complexity, which is advantageous. The number of rays is rather dependant on the angle subtended by the scene's bounding box. The rays are distributed uniformly in the cone and therefore no weighting is necessary. If only the closest intersections are processed it also solves the occlusion problem. The algorithm for this mode is summarised in Alg. 3.3.

The other advertised feature which is hard to implement into the point based

```
foreach ũ ∈ U do
    for d ∈ D do
        l_d = Cast(Ray(d, ũ))
        if l_d != null then
            ũ ← ũ+ SphericalWave(l_d)
        end
    end
end
```

Algorithm 3.3: Reversed mode of ray casting method for synthesis.

synthesis method was the nonuniform intensity distribution support or in other words the illumination model support. Actually, it is not so difficult to implement in the algorithmic way. The surface normal is all what is required to evaluate the local illumination model. If each point source had the appropriate normal assigned the rest is simply evaluation of the illumination equation. The problem is more like the number of this evaluations performed for each sample point. The number of evaluations equals to the number of casted rays. This further stress the computation requirements. However, the proper illumination model is something which cannot be neglected without consequences to the realism of the final image.

The ray casting approach brings another level of possibilities to the illumination model extension. It is because the ray tracing methodology for the global illumination [Whi80, CTL84] can be easily incorporated. This means that more complex light interactions like reflection, transmission or refraction can be incorporated into the final image by the means of the geometric optics instead of much more computationally demanding wave optics.

To conclude, the ray casting based methods are far more computationally demanding than other methods but on the other hand they provide greater flexibility for other rendering operations like non-diffuse material properties and they supports the occlusion test natively. Another advantage of using the ray casting for the hologram synthesis is the large body of research about ray casting acceleration techniques done in the classical CGI so all existing ray tracing methods and approaches can be exploited directly with just one major modification which is the complex amplitude based mode of operation.

Ray casting provides natural solution for occlusion and illumination. It also provides an elegant way for solving other effects like reflection and refraction thanks to the ray tracing approach well known from the CGI. The dynamic scene sampling approach removes the scene description restriction of the previous methodology. The speed of evaluation is estimated as more favourable in comparison to the previous methodology, especially in a case of more complex scene. The scalability remains good because this methodology is also easily implementable as parallel process. The hardware criterion is the only one considered as problematic because the scene sampling includes ray/geometry intersection test which is

non trivial operation and hardware implementation is therefore problematic.

3.1.3 Diffraction between two coplanar planes

Diffraction in general is very complex phenomenon and it is impossible to compute values of a light field analytically but several most trivial cases. There are however situations, where the whole problem of diffraction may be redefined and solution may be found more efficiently. The diffraction between two coplanar planes is one such situation.

The diffraction between two coplanar planes may be evaluated by the means of Fourier transform. There are two possible approaches of exploiting the Fourier transform. The first approach is based on the Fresnel approximation of the diffraction integral and the second is based on the principle of angular spectrum propagation. Both approaches exploits Fourier transform to obtain the light field on the target plane due to the source plane. The Fresnel approximation is an approximation and works only for certain distances between two planes. The angular spectrum propagation is more accurate and works for any distances. Angular spectrum propagation also allows evaluating the diffraction between tilted planes.

Fresnel approximation

So far, all presented methods had to iterate through all point sources located in a scene to obtain one final value of a light field. This is very time consuming process. There is however one way to simplify the hologram synthesis process using the Fresnel approximation, see Sec. 2.1.6. The Fresnel approximation, from [Goo05], is about approximating the distance r between two points $P_1(x, y, z)$ and $P_0(\xi, \eta, 0)$. Normally, the distance is computed as $r = \sqrt{z^2 + (x-\xi)^2 + (y-\eta)^2}$. If condition Eq. 2.41 is fulfilled then the square root can be approximated according to the Eq. 2.40 and the expression for the value of light field at the point (x, y) can be written in the form of Eq. 2.43. If reorganised, the expression becomes the Eq. 2.44 which is:

$$\tilde{u}(x,y) = \frac{e^{jkz}}{j\lambda z} e^{j\frac{k}{2z}(x^2+y^2)} \iint \left\{ \tilde{u}(\xi,\eta) e^{j\frac{k}{2z}(\xi^2+\eta^2)} \right\} e^{-j\frac{2\pi}{\lambda z}(x\xi+y\eta)} d\xi d\eta, \qquad (3.1)$$

and which can be recognised into the Fourier transform of the product of the complex field just to the right of the aperture $\tilde{u}(\xi, \eta)$ and the quadratic phase exponential $e^{j\frac{k}{2z}(\xi^2+\eta^2)}$.

The Fresnel approximation therefore allows convenient computation of the light field from the whole array of point apertures by exploiting the FFT technique. This, of course, assumes that the point apertures are distributed in a

regular orthogonal grid. A method that utilise this Fresnel approximation for computing a hologram was published in [LHJ68].

The scene has to consist of point sources placed in three dimensional regular grid. Each point has an intensity assigned. The scene can be therefore described using a set of grayscale images of the same size. The black color represents zero amplitude and white color represents the largest amplitude allowed, see Fig. 3.1.3. Each image is then transformed into an array of complex amplitudes.

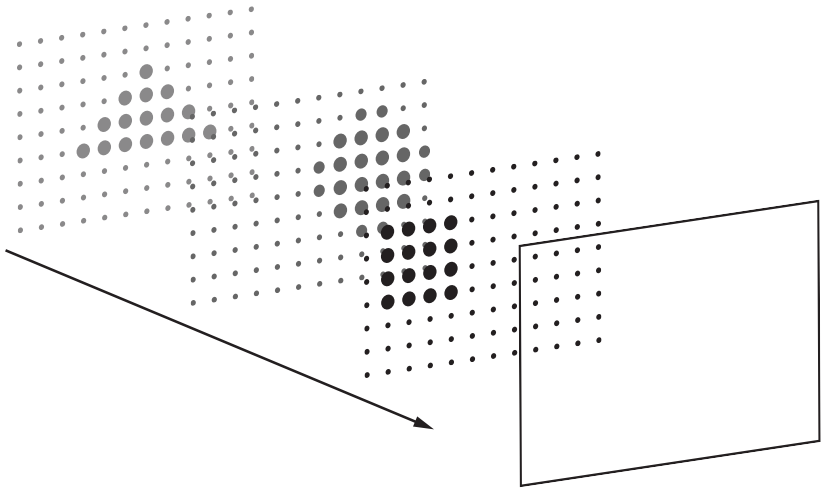


Figure 3.3: The scene consists of planar regular arrays of point apertures. The distance of the last array to the hologram is not in scale.

The following processing is based on the Eq. 3.1. Which means that each two dimensional stack of points having the same z coordinate is multiplied by the quadratic phase exponential $e^{j\frac{k}{2z}(\xi^2+\eta^2)}$ and FFT of the product is evaluated. This operation constitutes the inner part of the Eq. 3.1.

The result of FFT is then multiplied by the last exponential located in front of the integral in Eq. 3.1 and the product is accumulated into the resulting light field of the hologram frame. The process is repeated for all images of the scene grid. The particular hologram is then obtained as usual by adding the reference beam and computing the intensity. The whole process may be therefore expressed as the Alg. 3.4.

To conclude, this approach is very fast in comparison with the point wise or ray casting evaluation, the speed of evaluation criterion is therefore fulfilled satisfactorily. On the other hand, the photorealism criterion is not fulfilled. The most serious reason for the failure is the neglection of the occlusion. Each XY stack of the scene grid is processed absolutely independently on all others. The point sources at the stacks farther from the hologram plane are always visible, even though they shouldn't be. Another reason of the failure is the lack of illumination model support. The point sources emit their energy uniformly to all

```
// QuadExp represents e^{j k/2z (ξ^2+η^2)}
// Operator ⊗ represents element wise matrix multiplication
foreach I_z ∈ S do
    I'_z ← QuadExp(z) ⊗ I_z
    I''_z ← FFT(I'_z)
    ũ ← ũ+QuadExp(z) ⊗ I''_z
end
```

Algorithm 3.4: Fresnel approximation based synthesis method.

directions. This is enforced by the mathematics of the methodology and there is no resort to this issue.

Problematic is also the scene description since it is limited onto planar rectangular elements coplanar with the hologram plane. The quality of hardware implementation is as good as is the quality of FFT implementation, which is not as efficient as it is needed for the purposes of digital hologram synthesis. The scalability criterion is fulfilled satisfactorily because this methodology can be implemented as parallel process.

Angular spectrum propagation

The synthesis methods based on the angular spectrum propagation, see Sec. 2.1.6 are very similar to the methods described in the previous Sec. 3.1.3. The difference between the Fresnel approximation and angular spectrum propagation is the approximation part. The angular spectrum propagation is not approximation and therefore it does not impose any limitations on the spatial extend of the source and the distance of propagation. Although the angular spectrum propagation provides exact solution it is only a little bit more complicated compared with the Fresnel approximation since the only additional principal step is the inverse Fourier transform at the end of the process, see Sec. 2.1.6 for details. The approaches based on the Fresnel approximation can be therefore enhanced by replacing the Fresnel approximation based propagation with the angular spectrum propagation. The algorithm for this method is listed as Alg. 3.5:

```
// PropagExp represents the propagation exponential
// Operator ⊗ represents element wise matrix multiplication
foreach I_z ∈ S do
    I'_z ← FFT(I_z)
    I''_z ← PropagExp(z) ⊗ I_z
    ũ ← ũ+ IFFT(I''_z)
end
```

Algorithm 3.5: Angular spectrum propagation based synthesis method.

An interesting application of the angular spectrum propagation was published

in [TB93]. Authors proposed a method for synthesising a hologram of a scene composed of not regular 3D grid but of an arbitrarily tilted planes. The core of the approach is the propagation between two tilted planes. The first one is some of the plane of the scene, and the other one is the hologram frame.

The angular spectrum decomposes the light field into individual frequency components, each constituting a direction of propagation of a single planar wave. If those directions are rotated, the light field on a plane rotated to the original one. This transformation therefore allows obtaining a light field due to the plane A on a plane B which is rotated and transformed by the means of the angular spectrum propagation.

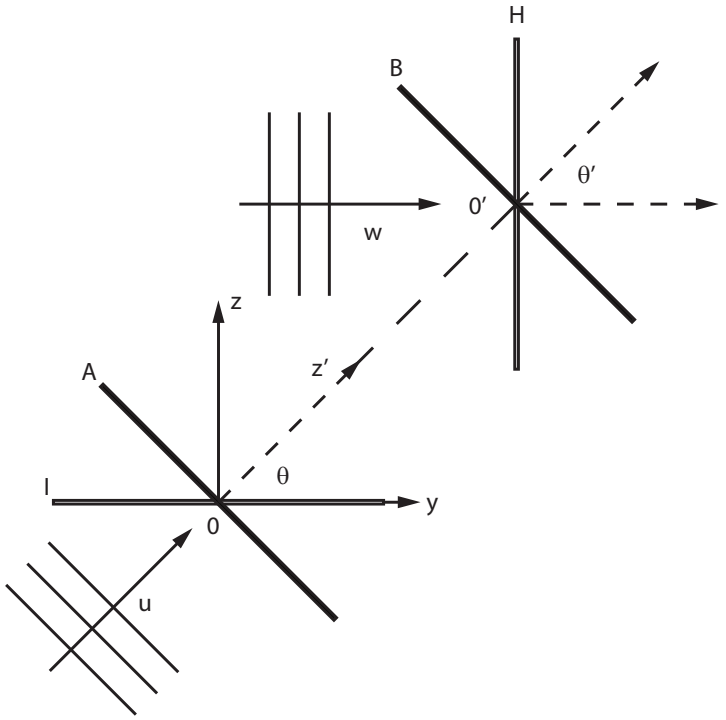


Figure 3.4: Geometry for the determination of the light diffracted by plane I onto hologram plane H. The angular spectra of virtual planes A and B are considered for this calculation [TB93].

According to the [TB93] two reference frames $(0; x, y, z)$ and $(0; \xi, \eta, \zeta)$ are assumed. The angular spectrum $A_2(f_\xi, f_\eta, 0)$ on the plane $2(\xi, \eta, 0)$ is calculated from the knowledge of $A_1(f_x, f_y, 0)$ which is the spectrum on the plane $1(x, y, 0)$. Let M be the rotation matrix relating (x, y, z) to (ξ, η, ζ) and the frequencies (f_x, f_y, f_z) to (f_ξ, f_η, f_ζ):

$$(x, y, z)^T = M\,(\xi, \eta, \zeta)^T \tag{3.2}$$

$$(f_\xi, f_\eta, f_\zeta)^T = M^T\,(f_x, f_y, f_z)^T \tag{3.3}$$

where $f_z = \left(1/\lambda^2 - f_x^2 - f_y^2\right)^{1/2}$.

The angular spectrum propagation states that:

$$A_2(f_\xi, f_\eta, f_\zeta) = \begin{cases} 0 \\ A_1(f_x, f_y, 0) f_z / f_\zeta \end{cases} \tag{3.4}$$

The propagation is valid only if $f_z > 0 \wedge f_\zeta > 0$.

The model of evaluating the diffraction pattern on any plane rotated and translated with respect to the source plane is illustrated in the Fig. 3.4 where only rotations about x axes are shown. The plane I with a transmittance $T(x, y)$ is illuminated by a monochromatic plane wave propagating in the direction u. The goal is to compute the diffraction pattern on the hologram plane H. It is done in the following approach.

The diffraction pattern $\tilde{u}(x, y)$ immediately behind the plane I is obtained from the expression:

$$\tilde{u}(x, y) = T(x, y) \exp(i 2\pi y \cos\theta / \lambda), \tag{3.5}$$

where θ is the angle between the y axis and the direction u. The Fourier transform of $\tilde{u}(x, y)$ follows and the angular spectrum of I is obtained. The plane A is considered perpendicular to the auxiliary optical axis z', which links the barycenters 0 and $0'$ of planes I and H respectively. A rotation transformation between the spectra of the planes I and A is done by exploiting the Eq. 3.3. The translation is accomplished by the angular spectrum propagation and the spectrum of the plane B is obtained. Another rotation transformation is executed between planes B and H thus the angular spectrum of the H is obtained. This is followed by the inverse Fourier transform of the angular spectrum which yield the diffraction pattern at the hologram plane H.

The process just described is repeated for all planes constituting the scene and their contribution is accumulated to the final diffraction pattern. The particular hologram is then obtained by adding the reference beam and evaluating the intensity. The algorithm for this method is listed as Alg. 3.6.

To conclude, although the angular spectrum based propagation is fast to evaluate and provides exact solution, it suffers by the same drawbacks described in the previous section. It is unable to handle occlusion and more advanced illumination models. This greatly reduces it's value as a hologram synthesis approach. It is however a great tool for numerical reconstructions from light field due to the accuracy of propagation between two coplanar planes.

```
// PropagExp represents the propagation exponential
// Operator ⊗ represents element wise matrix multiplication
foreach I_z ∈ S do
    I'_z ← FFT(I_z)
    R_z ← RotateSpectrum(I'_z)
    R'_z ← PropagExp(z) ⊗ R_z
    I''_z ← RotateSpectrum(R'_z)
    ũ ← ũ+ IFFT(I''_z)
end
```

Algorithm 3.6: Light field of arbitrarily tilted planes computation exploiting angular spectrum rotation and propagation.

Less problematic is the scene description since it is not limited to coplanar rectangular elements only. Elements can be arbitrarily tilted. This allows aligning each such plane with a triangle or planar polygon of the triangular mesh. The quality of hardware implementation is as good as is the quality of FFT implementation, which is not as efficient as it is needed for the purposes of digital hologram synthesis. The scalability criterion is fulfilled satisfactorily because this methodology can be implemented as parallel process.

3.1.4 Fourier hologram synthesis

Fourier holograms are a special kind of holograms, see Sec. 2.2.4 for details. They are special in the principle of recording. Both the reference and object beams are projected by a lens first before they hit the hologram frame. The lens is situated such that the reference point light source and object lies in one focal plane and the hologram plane incidents with the second focal plane of the lens. This is the optical equivalent of making the Fourier transform of the reference and object beam first and the hologram is then obtained from the interference of those transforms. The lens is also required for reconstructing the image.

The Fourier holograms have a very convenient mathematical description which allows using the signal processing approaches for their faster synthesis. One of the approaches was published in [LAR01, AR03]. Authors combines a set of scene projections by a special process, which is shown to be equivalent to the recording of the Fourier hologram.

The input of their approach is a set of scene's projections from various viewing angles. The viewing angles are not arbitrary but they are chosen in such way as it is illustrated in the Fig. 3.5. The viewpoint direction is parameterized by the angles ϕ_m and θ_m.

Each of the computed projections $p_{mn}(x_p, y_p)$ is multiplied by the exponential term $\exp[-i2\pi b(x_p \sin\phi_m + y_p \sin\theta_n)]$. This product is then summed to get a

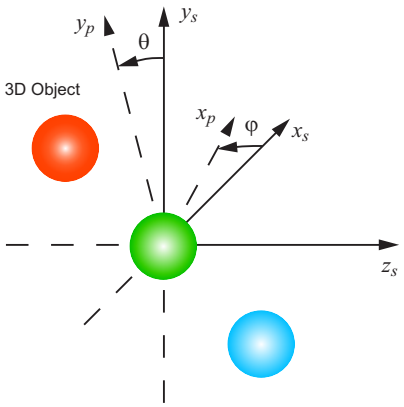


Figure 3.5: Scene projection from angles θ and φ [AR03].

single complex value in the following way:

$$\tilde{u}(m,n) = \iint p_{mn}(x_p, y_p) \exp[-i2\pi b(x_p \sin\phi_n + y_p \sin\theta_n)] dx_p dx_p, \tag{3.6}$$

where b is a real-valued constant. The values $\tilde{u}(m,n)$ are assembled into a complex matrix. Each element of the matrix corresponds to a different point of view and the matrix is arranged in the same order as the projected images are observed. The whole synthesis process is described in the Alg. 3.7.

```
// Operator ⊗ represents element wise matrix multiplication
foreach m do
    foreach n do
        I_{m,n} ← Project(θ_m, φ_n)
        I'_{m,n} ← I_{m,n} ⊗ Exponential(θ_m, φ_n)
        ũ(m,n) ← Σ_{i,j} I'_{m,n}(i,j)
    end
end
```

Algorithm 3.7: Multiple projections based method for Fourier hologram synthesis.

The complex matrix $\tilde{u}(m,n)$ constitutes the diffraction pattern on the plane of the Fourier hologram. The Eq. 3.6 is functionally equivalent with the optical setup depicted in the Fig. 3.6. According to [AR03, Goo05] the complex amplitude $\tilde{g}(u,v)$ on the back focal plane u,v due to the setup from Fig. 3.6 is given by:

$$\tilde{g}(u,v) = C \iiint t(x_s, y_s, z_s) \exp\left\{-j\frac{2\pi}{\lambda f}\left[ux_s + vy_s - z_s\frac{u^2+v^2}{2f}\right]\right\} \mathrm{d}x_s\,\mathrm{d}y_s\,\mathrm{d}z_s \tag{3.7}$$

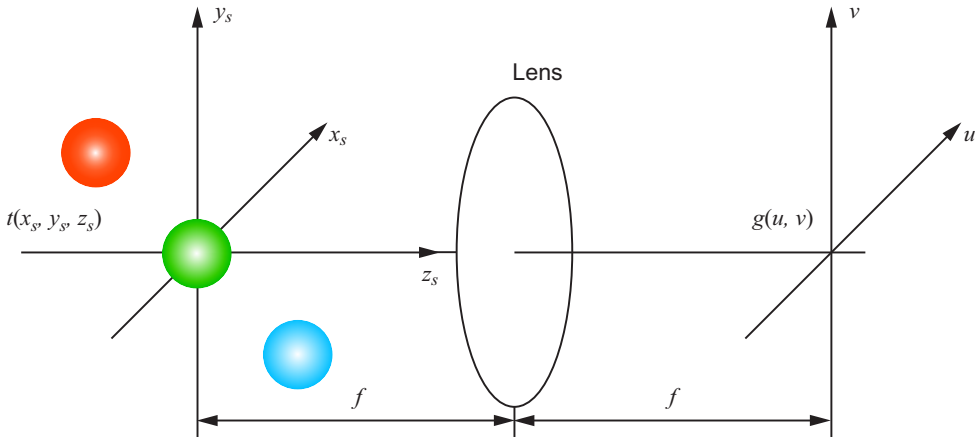


Figure 3.6: Equivalent optical system for the synthesis. [AR03]

Under some conditions the Eq. 3.6 can be rewritten into the form that is equivalent to Eq. 3.7 except some constants in the exponential power, see [AR03] for reference.

This approach uses multiple projections for computing the final diffraction pattern. It is advantageous because contemporary GPU are capable of rapidly generating 2D images of high quality. On the other hand, each final value of the diffraction pattern array needs one projection and therefore the number of projections required is significantly high.

Method that requires less projections for synthesising a Fourier hologram was described in [SIY04]. Authors claim that the full parallax Fourier hologram can be obtained by one dimensional scan of the scene. Authors relates the diffraction pattern on the hologram plane, see Eq. 3.8, to the 3D Fourier spectrum of the scene, see Eq. 3.9. From this relation follows that the values at the hologram plane lies on the paraboloid of revolution in the three dimensional Fourier spectrum of the object. From this relation is derived a method of indirect acquisition of the values using the 3-D central slice theorem that ensures that the partial components of the 3D Fourier spectrum are obtained from an orthogonal projection image of the object.

$$g\left(x_0, y_0\right) = \iiint O\left(x, y, z\right) \exp\left\{-\frac{i2\pi}{\lambda}\left[\frac{x_0 x + y_0 y}{f} - \frac{\left(x_0^2 + y_0^2\right) z}{2f^2}\right]\right\} dzdydz \tag{3.8}$$

$$
\begin{aligned}
g(u,v) &= \iiint O(x,y,z) \exp\left\{-i2\pi\left[ux+vy-\frac{\lambda}{2}\left(u^2+v^2\right)z\right]\right\} dxdydz \\
&= \left\{\iiint O(x,y,z) \exp\left[-i2\pi\left(ux+vy+wz\right)\right] dxdyd\right\}\Big|_{w=-\lambda(u^2+v^2)/2} \\
&= \mathscr{F}\left[O(x,y,z)\right]|_{w=-\lambda(u^2+v^2)/2}
\end{aligned}
\tag{3.9}
$$

At first, the object is projected onto a plane whose normal vector is inclined by θ to the Z axis. Subsequently, the projected image is 2D Fourier transformed. The 2-D Fourier transform corresponds with a sectional Fourier plane whose normal vector is also inclined by θ to the W axis in the 3-D Fourier space of the object. Therefore, it is possible to obtain partial components on the paraboloid of revolution in Eq. 3.9 from one projection image using this principle.

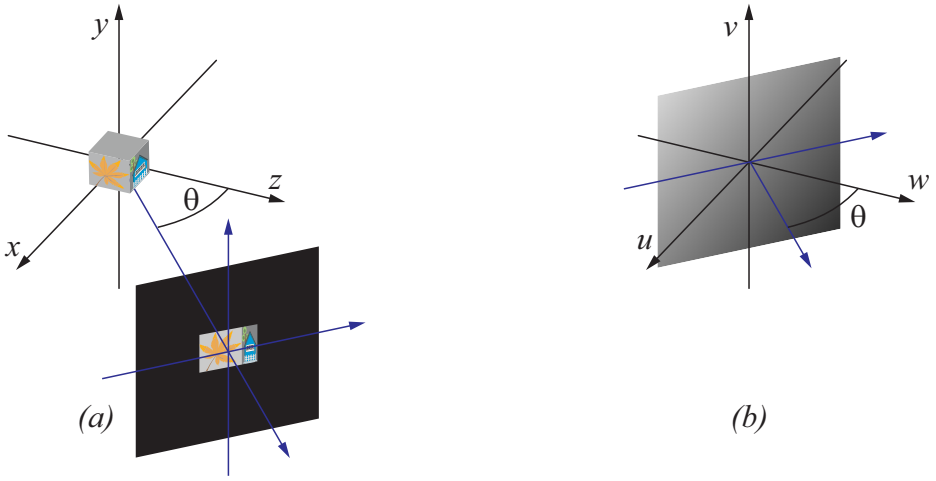


Figure 3.7: (a) Orthogonal projection in the real space and (b) a sectional plane in the 3-D Fourier space obtained from a projection image. [SIY04]

However, only the intersection between the sectional Fourier plane and the paraboloid of revolution can be extracted from the sectional Fourier plane, see Fig. 3.8. The intersections are calculated from the equation of the sectional plane and the equation of the paraboloid of revolution, see Fig. 3.8. The solution of these equation is an ellipse, see Eq. 3.10. When projected onto the u-v plane, the ellipse becomes a circle with radius $\tan\theta/\lambda$. The position of the centre of the circle depends on the direction of projection.

$$
\left(u-\frac{\tan\theta}{\lambda}\right)^2+v^2=\left(\frac{\tan\theta}{\lambda}\right)^2, \qquad w=-u\tan\theta \tag{3.10}
$$

To obtain all coefficients on the u-v plane many circles has to be extracted in such way that they fill the whole plane. The authors proposed a circular scanning

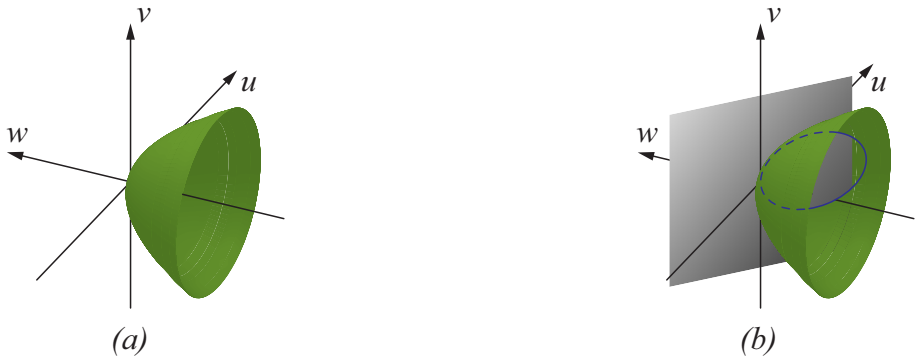


Figure 3.8: Paraboloid of revolution. (a) Components identical to objects waves and (b) intersection between the paraboloid of revolution and a sectional Fourier plane. [SIY04]

so that the extracted circles are obtained in the pattern illustrated in the Fig. 3.9. The whole process is described in the algorithm Alg. 3.8.

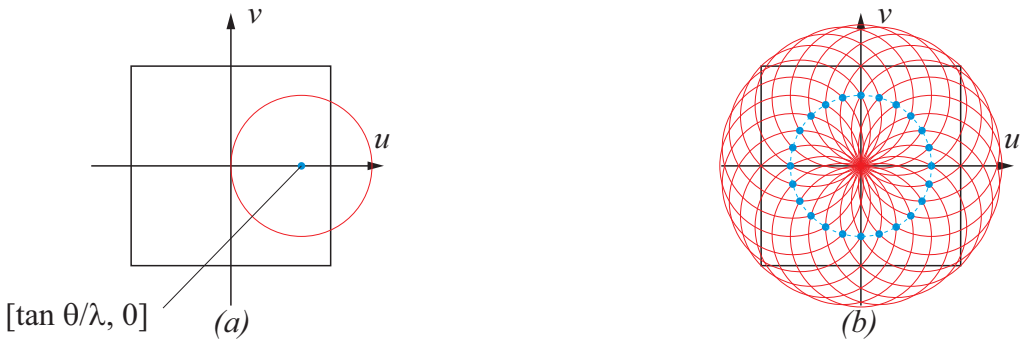


Figure 3.9: Extractive area on the u-v plane from (a) one projection image and (b) series of projection images. [SIY04]

begin
  foreach ξ **do**
    $I_\xi \leftarrow$ `Project(`θ_m, ξ`)`
    $I'_\xi \leftarrow$ `FFT(` I_ξ`)`
    $\tilde{u} \leftarrow$ `ExtractCircle(`I'_ξ`)`
  end
end

Algorithm 3.8: Values of the Fourier hologram obtained from sectional Fourier transform of scene projections.

This method uses different approach to the hologram synthesis. It allows to obtain holograms of real world scenes without need to use coherent light source.

The 1D azimuth scanning of the scene is easy an robust enough to be done easily. For synthesising a hologram of an artificial scene this method is not so efficient due to the computational requirements. Author claims, that they used 90 projection images to synthesise a hologram of resolution 256×256. However, the scalability of this method is discussable. The larger holograms needs more images to maintain the coverage of the plane. The utilisation of the computed Fourier transform is also very small since only the coefficients incident with the appropriate circle are actually used.

To conclude, the approaches from this group of synthesis methods uses fast signal processing techniques for obtaining holograms from a set of scene projections. It a very nice feature since contemporary GPU's are capable of computing such projections with great speed and visual quality including occlusion and very advanced illumination models. However, they have also disadvantages. One of them is the large number of Fourier transforms performed to compute the hologram. Each scene projection requires one. Another disadvantage is the more complicated optical setup required for reconstructing the Fourier holograms caused by the presence of the large lens.

The most common operation is Fourier transform of the some projection of the scene. The fact that two dimensional projections of the scene are utilised in the process of synthesis ensures the satisfactory fulfilment of the photorealism and scene description criteria. The speed of evaluation should be comparable to the ray casting methodology. The situation about scalability similar to all other methodologies, i.e. it is fulfilled satisfactorily because even this methodology can be implemented as parallel process.

The remaining criterion to evaluate is the hardware implementation criterion. There are two contrasting stages of the implementation. The positive stage represents the scene projection retrieval. It is satisfactorily solved by the nowadays GPU's. The negative stage is the FFT of the projection which has already been denoted as problematic. The definite evaluation of this criterion has to therefore wait till more rigorous testing will be completed.

3.1.5 Basis diffraction footprint combination

Diffraction footprint methods focus on the reconstruction process instead of the construction process. Hologram fringes are decomposed into some elementary fringes with known properties. According to those properties, more complex fringe pattern is build from the elemental ones so that the intended content in the reconstructed image is achieved. Such fringes are highly artificial and couldn't be achieved by any optical experiment.

The most trivial example of the diffraction specific method is described in [PM03]. The diffraction pattern of a single point source respective aperture is dependant only on the perpendicular distance from the hologram plane. The

change of the diffraction pattern due to this distance manifests as a scale of the pattern. The diffraction pattern of a point can be therefore precomputed into a complex texture once, see Fig. 3.10 and the diffraction patterns of any other point within a scene can be obtained as a translation and a scale operation of the basis texture which is easily done on a modern GPU's.

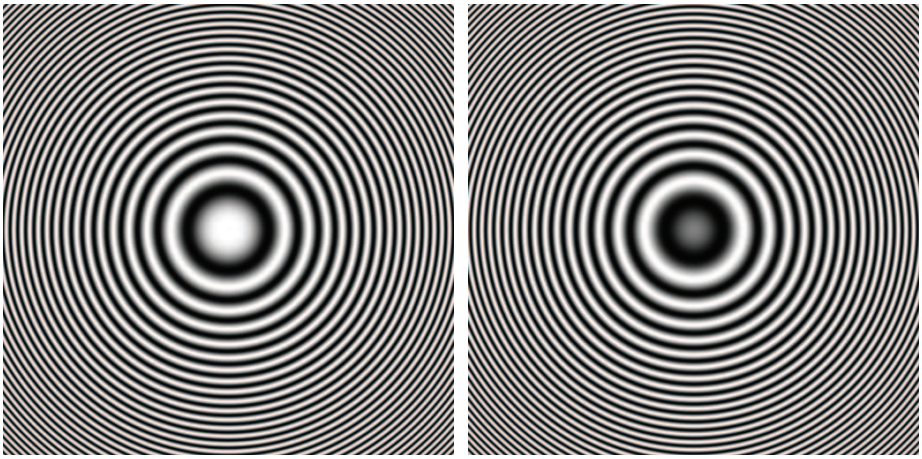


Figure 3.10: Two components of a complex texture representing diffraction footprint of a single point radiator. Imaginary component on the right and real component on the left.

The scale operation that simulates shift in the perpendicular distance is non-linear. If D_d represents the precomputed diffraction pattern at the distance d then the scale factor s for diffraction pattern at the distance d' will be $s = \sqrt{(d' + 0.5\lambda)/(d + 0.5\lambda)}$. The λ is very small compared to the distances and can be therefore neglected and the scale factor becomes $s = \sqrt{d'/d}$. The approximation is sufficiently accurate for some range of distance from the precomputed distance. The precomputed distance should be therefore chosen according to the spatial extend of a scene. The best distance to choose is the centre of the bounding box of the scene.

The translation operation changes the position of a point radiator according to the X and Y axes. The parameters of the translation operation are chosen such that the centre of the basis diffraction texture coincides with the position of the point radiator. It is obvious that the basis textures has to be sufficiently large to cover the whole target frame for all combinations of translation and scale that may occur for a given scene. The whole process of synthesis is simple and it is listed in the Alg 3.9.

This approach can be extended into more complex primitives. Diffraction patterns of a finite line has been used in [RBD+99] and [FLB86]. The same principle of transforming texture of an elemental diffraction pattern according to the spatial attributes of a primitive is used there. The diffraction pattern of

$FP \leftarrow$`PrecomputeFootprint(`D_d`)`;
foreach $p_S \in \boldsymbol{S}$ **do**
  $FP' \leftarrow$`Translate(Scale(`FP, $p_S.z/D_d$`)`, $-p_S.x$, $-p_S.y$`)`;
  $\tilde{u} \leftarrow \tilde{u} + FP'$;
end

Algorithm 3.9: Point diffraction footprint based method.

a line is still quite simple, see Fig. 3.11. While in [RBD+99] the precomputed pattern was stored in a complex texture, authors in [FLB86] have used analytic description of the conical waves, which is more accurate but the texture had to be recomputed for each line.

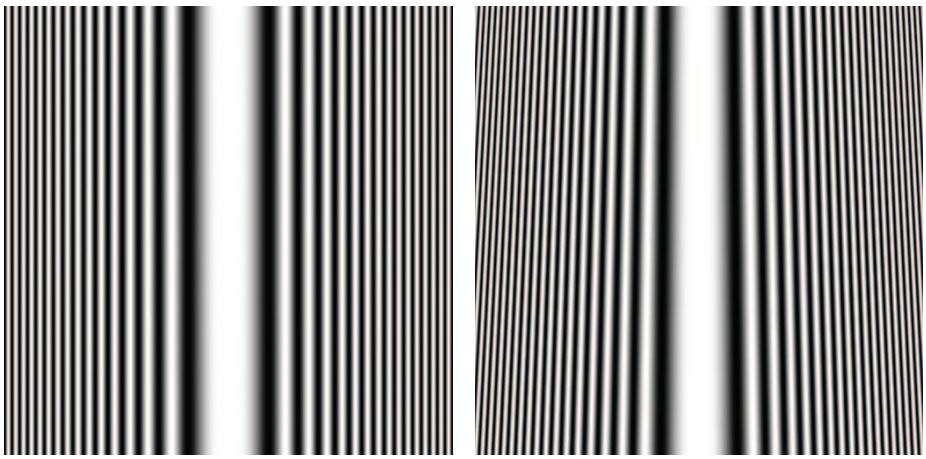


Figure 3.11: Real part of the complex texture of an infinite line. The pattern on the left reconstructs line parallel to the X and Y axis. The pattern on the right reconstructs line rotated with respect to the X axis.

The parametrisation of the line's diffraction pattern is similar to the parametrisation of the point's diffraction pattern. However, a line has more freedom in orientation than a point and therefore more transformations has to be done. Namely the rotation and taper operations has to be incorporated.

The rotation operation determines the orientation of the line's parallel projection into xy plane. The taper operation determines the angle subtended by the line and Z axis. If angle is other then $\pi/2$ both ends have different position in the Z direction. The parameters of the taper operation are scale factors for each end and they are computed in the same way as the scale factor for the point radiator. The taper transformation is depicted in the right side of the Fig. 3.11.

The problem constitutes the infinity of the lines. Such lines have a little use for imagining purpose. The finite lines are much more useful, however, the diffraction pattern from finite line is more complex and therefore an approximation is applied and diffraction pattern of a finite line is obtained by clipping the diffraction

pattern of an infinite line to the length of the line. This approximation causes blurring of the ends of a reconstructed line but tests proved this blurring feasible under certain conditions.

There is another problem with the clipping of the diffraction pattern. If line is short, the diffraction pattern will be inherently short as well. This means that the diffraction pattern will not cover the whole hologram frame. This has two unpleasant consequences. First, the intensity of the reconstructed line is proportional to the relative area covered by the corresponding diffraction pattern. The shorter the line is then less intense its reconstruction will be. And second, by not covering the whole hologram plane with the diffraction pattern, the corresponding reconstructed line will be visible only from certain viewpoints. Therefore, also the viewing area of a single line is directly influenced by the relative area a corresponding diffraction pattern covers.

All the problems noted above constitute significant drawback in exploiting line primitive as basic building block. More useful primitive would be a triangle. Most common description of a geometry is by a triangular mesh. Thus having a method for combining parametrised diffraction pattern texture of a triangle would be most convenient. An attempt for employing diffraction pattern of a triangle as the basic diffractive primitives is presented in [KDS01].

The diffraction pattern of a triangle is not universal, i.e. the diffraction pattern of an arbitrary triangle is not obtained as a simple transformation of the precomputed one. Therefore authors decided to precompute patterns for a triangle at different depths and different rotation angles. The synthesis algorithm therefore requires a huge lookup table parameterized by the depth and rotation of the triangle. Diffraction pattern texture for a triangle is depicted in the Fig. 3.12 and algorithm is summarised in Alg. 3.10.

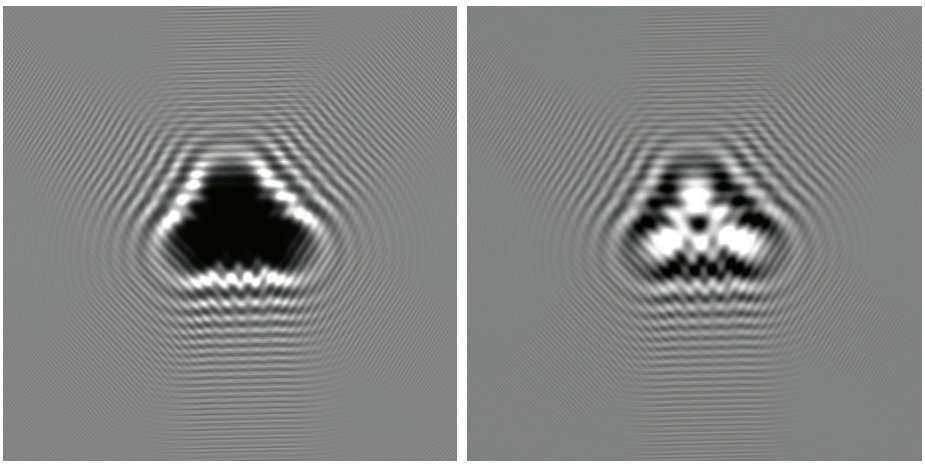


Figure 3.12: Complex texture for a triangle. The real part is on the left and imaginary part is on the right.

```
// d - depth
// ξ and η - rotation angles
T_{d,ξ,η} ←PrecomputeFootprints();
foreach Tri ∈ S do
    P ← T_{d,ξ,η}(Tri_d, Tri_ξ, Tri_η);
    ũ ← ũ + P;
end
```

Algorithm 3.10: Triangle diffraction footprint based method.

To conclude, the diffraction specific approaches try to reduce the complexity of the synthesis problem by focusing on the reconstruction output. If output conforms with the correct image, then the synthesised hologram is all right although the fringes are not physically correct according to the physics of the diffraction process. The approaches also exploits similar operations done by GPU's whose support is often incorporated into the implementation. But the diffraction footprint methods also have some serious drawbacks. One of them is the omnipresent problem of occlusion. The elemental patterns are precomputed as completely visible. Therefore, superposition of such element cannot yield an occluded surface. Another drawback is closely connected to the previous one. The basis elements are precomputed from a surface having a uniform intensity. As a result, no variation on the surface can be simulated. This makes the reconstructed image rather bland. Both problems haven't got straightforward solution which somewhat nullify the advertised speedup.

The basis diffraction footprint combination is very unbalanced methodology. It performs excellently in the speed of evaluation criterion, have a straightforward hardware implementation, is scalable and have no problem with the scene description criterion but on the other hand it performs terribly in the photorealism criterion. The methodology has difficulties in incorporating the occlusion solution and illumination into the process of hologram synthesis.

Although, the most important criterion of photorealism is completely failed it is still a subject of continuous research. If a solution for photorealism support is found without disrupting the outcome of the remaining criteria, this methodology becomes the one with the most potential.

3.1.6 HPO holograms synthesis methods

A hologram is very redundant. It is useful for the quality of the image but it burdens the synthesis with massive calculation requirements. One way of speeding up the synthesis process is therefore omitting some significant part of the redundancy and horizontal parallax only holograms do exactly that. The HPO hologram exhibit only the horizontal parallax and lacks the vertical one completely. This reduces the complexity of the synthesis by one order of magnitude.

The lack of vertical parallax is not so important as the horizontal one. It is because the human eyes are organised horizontally, thus the viewer does experience the binocular disparity in the HPO hologram, hence she does sense the depth. The drawback of HPO holograms is the necessity of a special reconstruction setup composed of cylindrical lenses [LAR01], otherwise the reconstructed image is heavily vertically blurred.

One of the greatest progress in the area of HPO hologram synthesis was done at Massachusetts institute of technology - MIT Media Lab, where the Spatial Imaging Group created a holographic system called Holovideo Display System. There are several approaches to digital HPO hologram synthesis that were tailored for this holographic device. They are summarised in the following sections.

MIT Holovideo Display System

The fundamental work about hologram synthesis for the MIT Holovideo device is described in [Luc94]. The fundamental contribution of this method is its capability of achieving interactive rate of synthesis. The interactivity was achieved by significant reduction of hologram redundancy, which is generally too high to be meaningfully utilised by human visual system. The complexity reduction is achieved by several factors.

First of all, the vertical parallax was sacrificed for reducing the dimension of the problem. Holograms without vertical parallax are commonly called horizontal parallax only or HPO holograms. HPO holograms are actually a vertical stack of one dimensional holograms which are much easier to synthesise. Also, vertical resolution of an HPO hologram can be much smaller than the horizontal one.

The disadvantage of HPO hologram is the need of special reconstruction setup [LAR01]. In a case of Holovideo system, this special reconstruction setup was replaced by the scanning process, where HPO hologram rows are processed sequentially. The final image is then perceived as a whole because of the persistence of vision.

Having the HPO hologram to synthesise instead of a full parallax one is not enough to achieve interactive frame rates. Further simplification is achieved by discretising the continuous signal of the hologram's fringe pattern. This dicretisation is motivated by the limited capabilities of the human visual system – HVS. Hologram in general provides more information then HVS is capable to process and therefore the hologram signal can be discretised without apparent degradation of the visual plausibility.

Let $S(x, f)$ represents the spatial spectrum of a single line of a hologram. This spectrum is continuous for a real hologram, but it is not necessary from the HVS point of view. According to the sampling theorem, the signal has to be sampled at least with two times higher frequency than the highest frequency contained in signal. In this case the sampling frequency should be two times

higher then the highest frequency usable by the HVS. This highest frequency should be found for both x and f.

First, for a given spatial frequency f_0, the spatial sampling spacing w_h in x must be small enough so that the most rapid amplitude variations in $S(x, f_0)$ can be reconstructed. Let the smallest variation in an image the HVS can see is set to 1 minute of an arc which is approximately 290 microradians. Considering a typical viewing distance 60 cm then the w_h is simply $0.000290 \times 0.6 = 0.000174$ m.

Second, for a given fringe region w_h centred at position x_0 the spatial frequency sample spacing Δ_f must be sufficiently small to reconstruct the most rapid variations in $S(x_0, f)$ as a function of f. The ability of the human eye to detect view-dependent variations is limited by the size of the pupil which is typically 3 mm. The angle subtended by the human pupil therefore determines the maximum allowable Δ_f. It can be computed from the diffraction grating equation Eq. 2.51 which can also be written as:

$$\sin\theta_q(\lambda) - \sin\theta_i = q\lambda f, \tag{3.11}$$

where f denotes the spatial frequency of the grating, λ denotes wavelength, q denotes diffracted order (can be negative), θ_q denotes the diffracted angle in the order q and θ_i denotes the incident angle.

Let $q = 1$, which is the order of interest in holography, and $\theta_i = 0$ then taking the derivative of both sides with respect to θ_1 one obtains:

$$\frac{df}{d\theta_1} = \frac{\cos\theta_1}{\lambda}. \tag{3.12}$$

For a viewing distance of 600 mm and a $\lambda = 0.633$ μm the calculation of Δ_f becomes

$$\Delta_f = \frac{(3mm)/(600mm)}{0.633\mu m} = 8mm^{-1}. \tag{3.13}$$

In a case of [Luc94], the fringe had width 1024 samples and each sample is spaced by 0.6 μm, the spatial frequencies therefore range from 0 to 850 mm^{-1} with spacing 0.8 mm^{-1}, which ten times exceeds the computed Δ_f. The fringe can be therefore encoded so that the bandwidth is compressed while useful image information is preserved.

Having the discretisation specified the term hogel can be introduced. Hogel is a small region of a fringe pattern. In a case of HPO hologram, it constitutes a small line segment of width w_h. Hogel is small enough to appear as a single point to the viewer and represents sampling of $S(x, f)$ at intervals of w_h and Δ_f which means that the hogel is completely described by its spectrum.

It was shown that diffracted light can be decomposed into a spectrum of planar waves, see Sec. 3.1.3. For each direction exists a spatial frequency that diffracts light into that direction. The range of direction is divided into small

intervals according to the computed Δ_f. For each of the interval is computed a basis fringe that diffracts light into a specific interval of directions. The spectrum of the hogel is then described by a set of weighting coefficients called hogel vector. Each component of the hogel vector corresponds with the appropriate basis fringe. The combination of the basis fringes then determines the diffraction duty of the hogel.

The computation of the hogel vector's components is done using a ray casting algorithm. From each imaged point a ray is cast to each hogel. The incident direction determines the hogel vector's appropriate component and the intensity of the imaged point determines the amount that should be added into that component. For example, point placed at the position of one of the hogels means that the hogel has to diffract into all directions so all components of the hogel vector are incremented. On the other hand, if point is placed at some distance from the hologram plane the light has to be diffracted in slightly different directions from each adjacent hogel. One or more components of each hogel vector must include a contribution for this point, see Fig. 3.13.

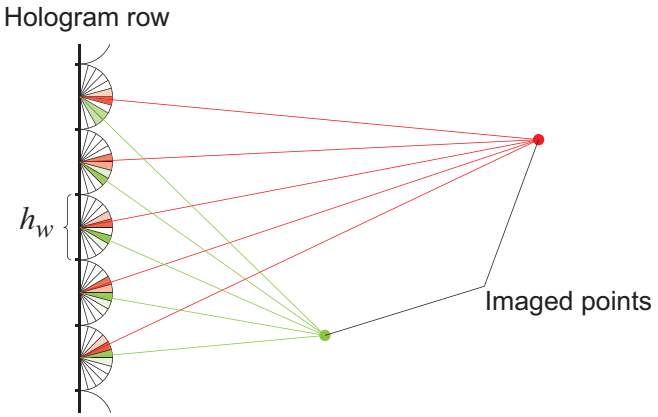


Figure 3.13: Hogel vector generation according to the imaged points positions.

This computation can be accelerated using a pre-computed diffraction table. This table maps (x, z) position of a point onto information about which components of which hogel vector should be incremented for how much. Another option is to use some existing rendering system and use its output as an input for this computation. This however needs a modified version of the diffraction table. The basic algorithm for synthesis is summarised in Alg. 3.11.

The interactivity of the MIT Holovideo system was also improved by the incremental update of the generated holograms published in [Ple03]. The incremental update is possible due to the nature of HPO hologram. The HPO hologram is actually a stack of independent one dimensional holograms. This independence is consequently used when change of the geometry occurs. Because it is expected that the change will be local it will therefore affect only some of the hologram

lines. Only those affected lines are consequently reevaluated.

```
T_{x,z} ←DiffractionTable();
foreach P ∈ S do
    W[] ← T(P_x, P_z);
    foreach w_i ∈ W[] do
        H_i[w_i.k] ← H_i[w_i.k] + w_i.value;
    end
end
```

Algorithm 3.11: Diffraction table based hogel vector generation.

The basis fringes are very important for high quality reconstruction. Their spectrum has to exhibit uniform amplitude across assigned frequency region and zero amplitude at all other. This is achieved by iterative approach in the first stage and by simulated annealing, see [Luc94] for more details. This process is quite lengthy but this is alright because once computed they can be used for the lifetime of the device for which they was computed. Example of eight basis fringes is depicted in the Fig. 3.14 along with their frequency spectrum.

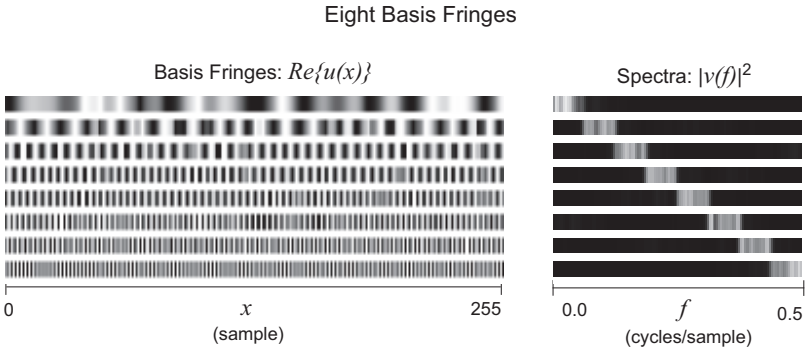


Figure 3.14: Set of 8 basis fringes and their spectrum. [Luc94]

The method just described was used for rendering HPO holograms. Author claims that it could be extended so it can be used for synthesising full parallax holograms. However, the number and size of basis fringes is increased quadratically as well as the number of hogels that are influenced by each imaged point. The extension therefore nullifies the acceleration features and makes the approach significantly less performing. Also the incremental update is not applicable in the extended mode of operation since the change in geometry will always affect the whole hologram.

Occlusion and shading support

The approach of M. Lucente was extended in [Und97]. The author augmented an occlusion solver into the Holovideo system and smooth shading was also incorporated. The situation is much more simpler in a case of HPO hologram when occlusion is discussed. The HPO hologram is a stack of horizontally oriented one dimensional holograms that do not interact with each other. The occlusion problem is therefore reduced from three dimensions into two dimensions only.

The occlusion solver has to decide, if the currently processed point is capable of contributing to any of the hogels. If point does not contribute to any of the hogels because it is fully occluded, then it is completely rejected. The opposite case where completely visible point contributes to all hogels is also easy since it is processed as described in the previous section. The more complicated case is when point contributes only to a subset of hogels. The information about hogels that are affected by the partially occluded point must be passed into next stage of the system, so that unaffected hogels are not processed.

To evaluate the occlusion, an auxiliary information is needed since Holovideo works pointwise and point is too small to occlude anything. The auxiliary information is extracted from the original geometry that was used for populating the input points. For each plane corresponding to the currently processed hologram line an intersection with the geometry is computed. This intersection produces a set of polygons respective line segments. Those segments constitutes the so called obstructor mask against which input points from the same plane are tested.

Each point has to be tested against each part of the obstructor mask. This is rather complicated so two simpler tests are performed first. The first test checks, if the processed point lies on the tested obstructor mask segment. If yes, then the segment is rejected from the following testing. The second test rejects all obstructor mask's line segments that have both endpoints farther from the hologram line then the precessed point and therefore cannot never occlude the point.

The further processing consists of computing the geometric shadow of each segment against the $z = 0$ plane. If any part of the projection lies within the horizontal extend of the hologram then the projection added into the point sources occluded interval, see Fig. 3.15. The final interval is kept as simple as possible so each new interval is added logically.

In the end, each point has its occluded interval and can be passed into the next stage of processing. Those points having occlusion interval which completely covers the whole hologram line can be rejected right away. Other points contribute only to those hogels, that do not fall into the occluded part of the occlusion interval. The contributions are normally processed according to the diffraction table as described in the previous section.

The last extension of the Holovide system authors of [Und97] described was

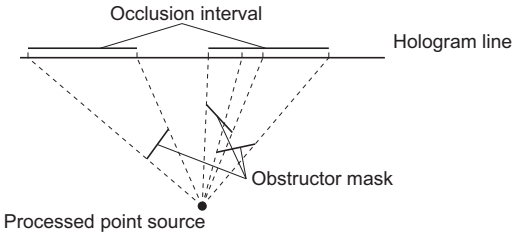


Figure 3.15: Occlusion interval determination [Und97].

shading and illumination. Both shading and illumination support is rather simple one. The intensity is computed according to the illumination model, Phong's model in this case, and assigned into each point source as its amplitude. This is done also for the specular component, which is view dependant and should be therefore recomputed for each hogel influenced by the processed point. However, authors evaluated this approach as too computationally demanding and decided to evaluate the specular component from one viewpoint and use that value for all hogels. The specular highlight is therefore static even though it should move along the surface according to the viewer's positon.

Conclusion

All synthesis methods dedicated to the MIT Holovideo system depend on the special features of the HPO hologram synthesis. Those special features, which are responsible for the phenomenal efficiency of the whole system, would not persist if the Holovideo system was transformed to support the full parallax hologram synthesis and reproduction. Therefore the significance of the methods described in the Section 3.1.6 for the purposes of this work is reduced. However, some of the concepts utilised in the HPO hologram synthesis are worth of incorporating into other methodologies, e.g. the basis diffraction footprint combination methodology may benefit from those concepts.

3.2 Summarisation and evaluation of the methods

The variation of the approaches for digital hologram synthesis is large. The common characteristic is the tradeoff between the quality of the output and the computational demands. The fastest approaches are the least flexible ones and vice versa. When evaluating the eligibility of the methodology, the application of the synthesised holograms plays a great role in the process. Since the main goal of this work is high quality imaging then the most important feature is the support for photorealistic rendering and the secondary feature is of course the speed of evaluation. However high quality and fast evaluation are in obvious contradiction. The advertised methods are therefore compared with respect to their ability to

produce photorealistic output or their potential extensibility in this direction and their speed of evaluation. The overview of the synthesis methods' parameters is presented in the table A.1.

According to the results of evaluation, as the most eligible methods for the purposes of this work were chosen these methods ordered by their eligibility:

1. Ray casting method of synthesis,
2. Fourier hologram synthesis,
3. Basis Diffraction footprint combination.

Ray casting was chosen as the best because it is capable of providing photorealistic output. Although the core speed of evaluation is not so good the scalability of this method compensates that issue. And finally, the scene content description does not require any additional information and standard description as used nowadays suffices. Successful hardware implementation also supports the choice of this methodology.

Fourier hologram synthesis was chosen for its ability to produce photorealistic output. The most interesting feature of the Fourier hologram synthesis is its input which consists of a set of two dimensional projections of the synthesised scene. This even allows synthesising holograms from the images provided by charge-coupled device - CCD camera. The reason for being chosen as second is necessity of more complex optical setup for reconstruction. Also the scalability and hardware implementation are more problematic than in the case of ray casting.

The basis diffraction footprint combination excels with the speed of evaluation and hardware support, however the realism of the output is in the current state unsufficient. If the issue of realism is sufficiently solved then this approach will become very well performing and might compete with the previous two methodologies. This approach was therefore also chosen to be included in the future research.

From the evaluation is apparent that the MIT Holovide also performs superbly. However, its specialisation to HPO holograms may become unfeasible in the future. The extension of the system into a full parallax supporting holographic system will probably spoil all advantages and interactivity implied by the HPO hologram synthesis specifics. This is a reason, why this methodology was not chosen as perspective one.

Chapter 4

Proposed ray casting synthesis method

Some work has already been done in the area of ray casting methodology for holograms synthesis. The ray casting was chosen as the core technology because it is closely related to the area of computer graphics and has a large body of research at disposal. The problem of synthesis was restricted to HPO hologram synthesis because of the computation complexity reduction. The reduced computation complexity implies shorter computation time and shorter computation time allows more rapid development. The ray casting method is also flexible enough to be extended to synthesise the full parallax holograms too.

Among the advantages of the proposed method belongs automatic processing, which doesn't require any special input. The input of the method is the common scene description including the triangular meshes together with material and light specifications. The light type is restricted to directional light but only due to the efficiency reasons. Any other light types, e.g. omnidirectional, spot, are supported but for the cost of increased computation complexity. This support for common scene description is a very useful feature because it allows seamless integration of the synthesis method into the existing 3D content authoring tools like 3D Studio Max.

The rest of the chapter deals with the description of the method, which has already been published in [JHS06c] and [JHS06a].

4.1 Basic principle

This section describes the proposed method for synthesising a hologram of a triangular mesh objects. This method is capable of evaluating an occlusion, i.e. a point is not considered if it is occluded by some other point and visibility,

i.e. less area is visible if a triangle is tilted or far away. These two features make reconstructed scene more reality like. The method is tailored for the HPO hologram synthesis because the problem of synthesis is much easier for HPO holograms. This is due to reduced dimension of the problem. Instead of three dimensional problem of general occlusion and visibility evaluation, one only deals with two dimensional problems in the case of HPO. The Fig. 4.1 depicts the situation for one single row of an HPO hologram.

If an HPO hologram synthesis is considered, then triangular mesh objects are decomposed into a set of planar polygons or polylines if mesh is not manifold. The complex amplitude at a hologram sample is then obtained as an integral of contributions from all polygon edges over half a circle, see Fig. 4.1, providing the edges face the sampled point. However, the problem of such integration is the complexity of the integrated function. More over, this function must exclude those edge parts, which are occluded by closer edges and thus invisible from the sample's location.

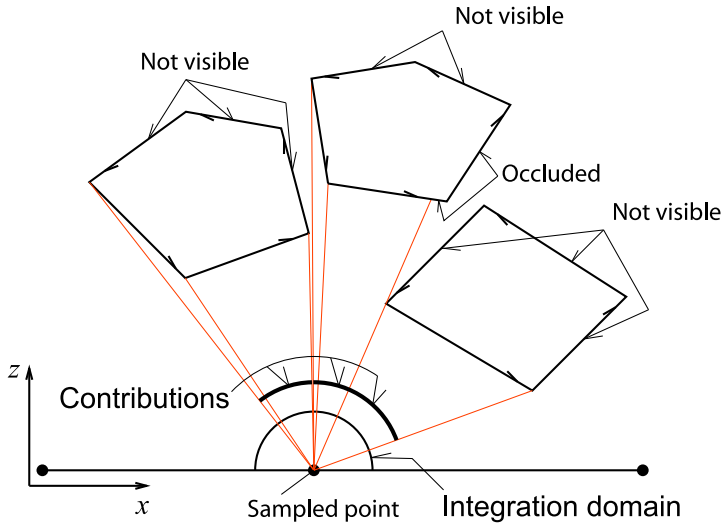


Figure 4.1: Integration domain of a single sample of an HPO hologram.

The integral can, however, be estimated by sampling the scene angularly. This sampling is depicted in the Fig. 4.7. The integration is then replaced with a summation. Since this summation has to be performed for each hologram sample, it has to be as efficient as possible. Several optimalisation approaches have been developed which perform this summation very efficiently and they are described in detail later on in the text.

The synthesis process is divided into several stages that forms some sort of hologram synthesis pipeline. The first stage of this pipeline is responsible for extracting the polygons and polylines from the intersection of the currently processed slicing plane and the geometry of a scene. The process is illustrated in

the Fig. 4.2.

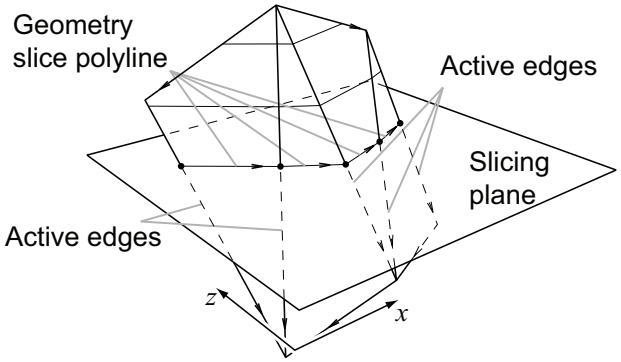


Figure 4.2: Geometry slicing.

Finding the intersection is a quite easy task and it can be performed in a similar way as the triangle scan line conversion is done [Wat00]. The scanning is performed in the Y axis direction. Vertices of each triangle have to be sorted by their Y coordinate. The first and the last vertex define the leading edge, the first and the second one define the top edge and the second and the third one define the bottom edge. A difference vector corresponding to the scan line spacing is computed for each of the triangle's edges. Then the initial intersection of the edge with the nearest scan line is computed, after that all consequent intersections are computed simply by adding the difference vector to the previous intersection. The process is illustrated in the Fig. 4.3.

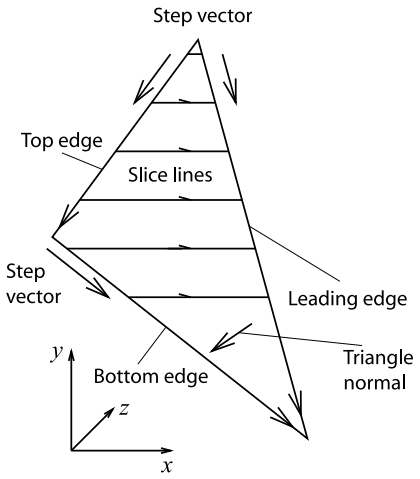


Figure 4.3: Modified triangle scanline slicing.

The scanning is edge based. A list of active edges is maintained for each slicing plane position. Each of the active edges has indices pointing at the opposing edge to the left and to the right. The left and right side is derived from the Y axis and

the triangle's normal. Singular triangles having the normal linearly dependent with the Y axis do not contribute to the hologram at all and can be therefore omitted.

When new edge is added into the list of active edges the integrity of the opposing edge indices must be checked. If the edge indexed by the new edge does not index the new edge back then the swap between the upper and lower edge has occurred and the index in the opposing leading edge has to be corrected.

The slice of a triangle, which is a line, is determined from the intersection points of the current scan line with the leading edge and the currently active opposing edge. The acquired line is required to be oriented in such way that its first vertex is on the left side and the second one on the right side. The left respective right side is derived from the triangle's normal.

An easy way to find out if the leading edge determines the first or the second vertex of a slice line is using some vector algebra. One has to compute a dot product of a triangle's normal and a vector product of the leading edge and the top edge considered as vectors. The sign of the dot product determines if the vertex obtained from the leading edge is the first one or the second one. This method of course requires uniform orientation of all triangles.

For each scanline and each active edge a corresponding slice vertex is created. Then slice edges are generated using the opposing edge information. Since triangles are aware of their orientation the slice edges are all oriented from left to right. This orientation is exploited when visibility is evaluated, see below. The vertices and edges from all objects are joined into a single list and passed into a next stage of the pipeline.

The complication at this stage constitutes the data structures required for storing the data required for shading purposes like normals or texture coordinates. The data need to be interpolated form values stored in vertices. The first phase of interpolation is performed during the geometry slicing stage. Each scanned edge has a data record associated and the interpolated values are obtained in a same fashion as the computation of intersections described above. The difference corresponding to the change in the Y component is computed in the initialization step and the actual value of datum is obtained from the previous one simply by adding the difference. This process constitutes a linear interpolation.

Once the slice edges are obtained, the interpolated data values d are obtained as $d(l) = A + l(B - A)$ where A are data at the first vertex of a slice edge, B are data at the second vertex of a slice edge, l is the interpolation parameter acquired as $l = \mathbf{L}_S/\mathbf{L}$, see Fig. 4.8. Difference $(A - B)$ is stored in the slice edge.

The texture coordinates were mentioned as an example of interpolated data, but any other attributes can be added such as vertex colours. Normals are also usually interpolated but the normals have to be renormalised, which is quite costly operation. As an alternative, the normals can be taken from a normal

map providing a texture fetch is faster than the normal normalisation.

The result of the slicing is a set of oriented lines. The situation is depicted in the Fig. 4.1. It is obvious that some of the lines are invisible from every sample point of the current row. The question is how to detect invisibility from a given point. This is the part where line orientation comes into play.

The angles subtended by the join of the examined x sample at the hologram line and the respective endpoints of the line and the X axis are computed. The visibility of the line is then determined by simple comparison of the computed angles. See Fig. 4.4 for reference. If α angle is smaller than the β angle then the line is not visible from the point x.

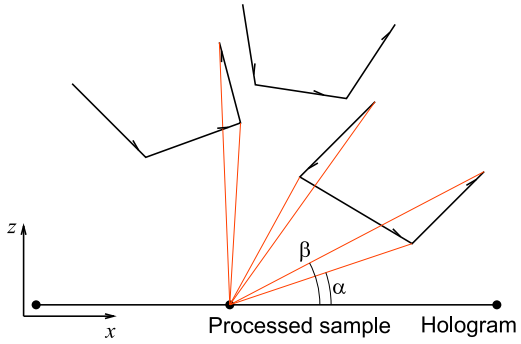


Figure 4.4: Principle of visibility check.

The line is a potential contributor if it is visible from at least one of the hologram extremes. See Fig. 4.5 for reference. The lines that are invisible from both hologram extremes are removed because they never contribute. Everything is set for the next, most computationally demanding phase, which is the angular sampling.

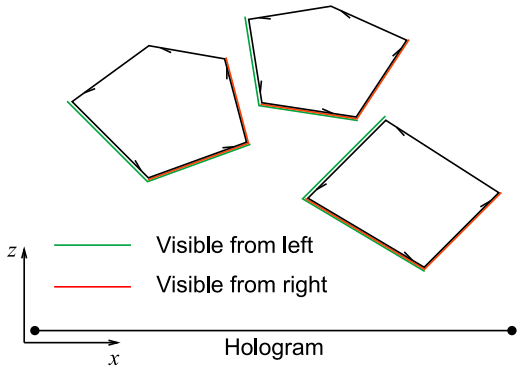


Figure 4.5: Candidate lines are those visible from at least one hologram row extreme.

One more preprocessing task, which is the lines length computation, has to be done. The direct distance computation, exploiting the square root operation,

is performed because the length has to be computed only once for each row. The length information is reused later on for indirect distance computation. Now everything is ready for computing the hologram samples at the current row.

The angles for each line have to be computed with respect to the current sample x position. Once the angles are acquired, the lines have to be sorted by the angle assigned to the second vertex. This sorting does not have to be so terrible operation to do because the angles are changing slowly since the hologram sample x position is changing slowly as well. The intensive sorting has to be therefore done for the first sample point only. The quick sort algorithm is used in this case. However, from this sample point further one can use a bubble sort, which is very efficient in sorting almost sorted set, which is this case.

The sorted list of lines is then used for maintaining the list of active lines, which is a list of lines that are potential contributors to the currently sampled angle. Whenever the angle is increased, all the lines, with the β angle smaller then the currently sampled angle, are moved from the sorted list into the list of active lines. Then the lines with the α angle smaller then the currently sampled angle are removed from the list of active lines. It should be noted, that the lines which are invisible from the current sample point are ignored, it is those where angle α is smaller than the angle β.

For each line in the active lines list the distance of the current angular sample to the hologram sample is computed. Then the closest sample is eventually accumulated because it is the non-occluded one. This is repeated for each sampled angle and in the end the phase and intensity at the sampled hologram position is evaluated and written into an output. The bottleneck of this process is the distance evaluation. But it can be significantly accelerated using the method described below.

There is one place, where pre-processing can be exploited. The sampled angles are the same for all hologram samples. When illumination model is evaluated, the view vector is required. This view vector correspond with the currently sampled angle and therefore the view vectors can be precomputed into a table. Furthermore the sine's of the sampled angles are also needed for aligning the Z-depth to the nearest wavelength multiple. This aligning is not performed in the Z direction but it is rather performed in a sampled direction. This distance correction d_s is computed from the triangle depicted in the Fig. 4.6 as $z_s / \sin \alpha$. The sine's can be also precomputed into a table.

The advantage of the angular sampling is that the lines, which are facing to the hologram and are closer, are sampled more densely then the tilted or distant ones. This complies with the fact that less area is visible from the tilted or distant triangles and therefore such triangles contribute less to the final sample value. One can object that this is only two-dimensional case but in three dimensions triangles may tilt in two directions. But this is actually handled during the scan

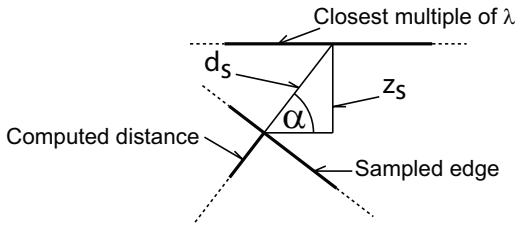


Figure 4.6: Aligning of the Z-depth.

line conversion. Tilted triangles end up with less scan lines then those facing the hologram plane more directly.

This angular sampling also provides uniform sampling for the whole scene. As a result, one can use radiance values obtained from illumination model as the amplitude value. And because the hologram is normalised at the end, the intensity distribution of a scene is maintained and it is independent on the angular sampling step. The angular sampling process is depicted in the Fig. 4.7

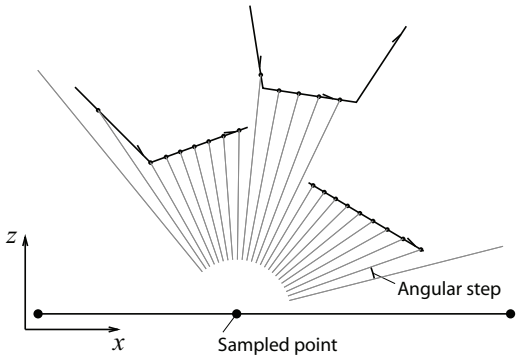


Figure 4.7: Angular sampling of the slice geometry.

Edges in a slice are sampled using the sine law in a triangle. The triangle is depicted in the Fig. 4.8. The β angle is gradually decreased and the α angle is gradually increased as the sampled angle advances counter clock wise. The α angle is used for recalculation of the triangle ratio. The γ angle is used for evaluation of the distance d and the β angle is used for evaluating the length $\mathbf{L_S}$ exploited for computing the parameter of a data record's linear interpolation.

The uniform change of the sampling angle is an advantage because the differential equation for cosine evaluation can be used. This differential equation is based on the well known reflection vector evaluation $\mathbf{R} = 2\mathbf{N}(\cos\alpha) - \mathbf{L}$. The cosine corresponds to the X coordinate of the reflection vector. The new cosine value is obtained from the two previous values as it is apparent form the Eq. 4.1 [Has06]. The sine function can be computed similarly only a value of $\pi/2$ is subtracted from the initial angle.

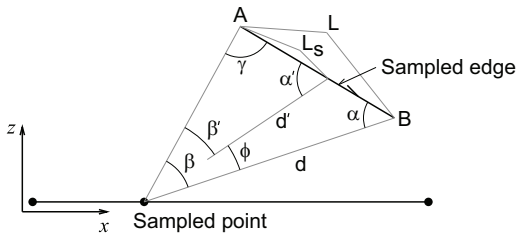


Figure 4.8: Distance evaluation.

$$\cos\beta_i = 2\cos\beta_{i-1}\underbrace{\cos\phi}_{const} - \cos\beta_{i-2} \tag{4.1}$$

There is another way of determining the distance to the intersection. It is based on the computation of the intersection between parametrised line equations. It leads onto couple of equations which can be easily presolved. The results showed small improvement against the sine law based evaluation.

Data needed for evaluating the illumination model are interpolated using the principle described above in the text. A view vector is precomputed in a table and since all light sources are assumed directional, light vectors are therefore also known. The one last missing component is a normal.

The normal can be interpolated in the same way as the texture coordinates but linear interpolation does not preserve the required unity of the normal vector. The interpolated vector has to be normalised first which requires square root operation. This can make the normal evaluation slow.

An alternative is to use a normal map. Texture coordinates can be interpolated without problem and the corresponding normal can be obtained form a texture. On the other hand, a texture fetch operation is not trivial either.

All necessary components are therefore known and the illumination model can be evaluated. A simple Phong illumination model and Phong shading [Pho75] was used for the purposes of testing. The results proved great improvement in a comparison to the originally used constant shading in [JHS06c].

4.1.1 Summary

A complete pipeline for rendering digital HPO holograms was proposed. It can be seamlessly integrated into the existing rendering systems because it consumes the same scene content description as the standard 3D graphics pipeline and no other information is required.

There are several issues which will be addressed in a future work. The major one is that the presented holograms are quite small. Bigger holograms have to be rendered to obtain more useful viewing aperture. An additional acceleration should be achievable with specialised HW or standard GPU as well.

Chapter 5

Summary and Conclusion

The purpose of this STAR document was the following. First, present a compilation of pieces of knowledge regarding the holography science in general. This compilation constitutes a supporting material for the synthesis method's overview and evaluation and furthermore, it may serve as a minimal knowledge pack useful for anyone, who wants to deal with holography.

Second, the overview and evaluation of the digital hologram synthesis methods. This overview constitutes the starting point for the next work and development. It reflects the actual state in the area of digital hologram synthesis. The evaluation of the method's features directly determines the direction of the following direction of research. The first results of this research is included in the Chapter 4.

And finally, the proposal of the forthcoming research plan. This proposal contains the intended research plan, that will eventually lead to the ultimate goal. The plane includes the near future and also includes the far future that is beyond the time scope of this work. This is caused by the dependence of this work on the technological advance of the holographic displays. The following section contains detailed description of those three purposes stated above.

5.1 Holography Overview

The Chap. 2 of this document contains a minimal bulk of knowledge about holography required for understanding the hologram synthesis principles. It is written in an educative spirit so that it may serve as an introductory reading for the potential holography adepts. The main purpose is however to support the overview of the digital hologram synthesis. This section of the work references the holography overview many times.

The overview consists of several areas of knowledge. The first area concerns to the mathematical background of the wave nature of light. This is a brief introduction of the notation and terms. The next area concerns to the mathematics and principles of the diffraction phenomenon. The diffraction constitutes the backbone of the whole holography science so it is quite comprehensive. The last area covered in the Chap. 2 concerns to the principles of the actual holography. That includes the recording and reconstruction, optical configurations description and the brief introduction into the issues of the digital holography, where also falls the issue of digital hologram synthesis.

5.2 Synthesis methods

The Chap. 3 contains quite detailed description of the most common methods of digital hologram synthesis known up to date. The purpose of this overview is to provide a good starting point for the following development in the area of the digital hologram synthesis. The overview is aimed to identify the strong and weak points of the existing methods so that the most adequate one could be chosen as the central philosophy of the following research.

There is a large variety of approaches addressing the issue of digital hologram synthesis in the literature. From this variety, several basic methodologies has been extracted in the evaluation process. They are:

- Diffraction integral evaluation,
- Ray casting,
- Diffraction between two coplanar planes,
- Fourier hologram specific synthesis,
- Basis diffraction footprint combination,
- MIT holovideo specific synthesis.

Each section describing a method includes method's mathematical/physical background and a discussion about the method's suitability for the purposes of this work. The suitability of each method has been measured against these evaluation criteria sorted descending by their priority:

1. Photorealistic output,
2. Speed of evaluation,
3. Scalability of the algorithm,
4. Hardware implementation,

5. Scene description.

5.2.1 Diffraction integral evaluation

The diffraction integral evaluation methodology evaluates the diffraction integral and thus is close to the actual physical process. This fact more or less guarantees the validity of this methodology. The drawback of this methodology is the complexity of the diffraction integral. The solution is impossible to calculate analytically and therefore numerical approaches are used. The numerical approach introduces problems with the integration domain sampling, i.e. scene is decomposed into point cloud.

With regard to the evaluation criteria, this methodology has limited options in achieving photorealism. The occlusion solving and illumination model support is not straightforward to incorporate in the process. The speed of evaluation is not great compared to other methodologies and finally the scene has to be approximated by the point cloud which impose restriction on the scene description criterion. The evaluation criteria discussed until now were not fulfilled satisfactorily. On the other hand, the methodology performs better in the last two evaluation criteria of scalability and hardware implementation. Methodology can be implemented as a parallel process which fits the hardware design.

5.2.2 Ray casting

The ray casting methodology is akin to the diffraction integral evaluation methodology. Ray casting removes the weak point of the integral evaluation. Ray casting provides natural solution for occlusion and illumination. It also provides an elegant way for solving other effects like reflection and refraction thanks to the ray tracing approach well known from the CGI. The dynamic scene sampling approach removes the scene description restriction of the previous methodology. The speed of evaluation is estimated as more favourable in comparison to the previous methodology, especially in a case of more complex scene. The scalability remains good because this methodology is also easily implementable as parallel process. The hardware criterion is the only one considered as problematic because the scene sampling includes ray/geometry intersection test which is non trivial operation and hardware implementation is therefore problematic.

5.2.3 Diffraction between two coplanar planes

The diffraction between two planes is done by the means of FFT and therefore methodology exploiting this feature is considered as the fastest ones. This methodology therefore best fulfils the speed of evaluation criterion. However, this methodology suffers in the other criteria. Most importantly the photorealism criterion is greatly limited due to the diffuse surface restriction imposed by

the principle of evaluation. More seriously, the occlusion is completely ignored in this methodology. This constitutes serious degradation of the reconstructed image quality.

Problematic is also the scene description since it is limited to planar rectangular elements in better cases arbitrarily tilted. The quality of hardware implementation is as good as is the quality of FFT implementation, which is not as efficient as it is needed for the purposes of digital hologram synthesis. The scalability criterion is fulfilled satisfactorily because even this methodology can be implemented as parallel process.

5.2.4 Fourier hologram specific synthesis

The Fourier hologram synthesis is specific by the massive exploitation of the FFT in the process. The most common operation is Fourier transform of the some projection of the scene. The fact that two dimensional projections of the scene are utilised in the process of synthesis ensures the satisfactory fulfilment of the photorealism and scene description criteria. The speed of evaluation should be comparable to the ray casting methodology. The situation about scalability is similar to all other methodologies, i.e. it is fulfilled satisfactorily because even this methodology can be implemented as parallel process.

The remaining criterion to evaluate is the hardware implementation criterion. There are two contrasting stages of the implementation. The positive stage represents the scene projection retrieval. It is satisfactorily solved by the nowadays GPU's. The negative stage is the FFT of the projection which has already been denoted as problematic. The definite evaluation of this criterion has to therefore wait till more rigorous testing will be completed.

5.2.5 Basis diffraction footprint combination

The basis diffraction footprint combination is very unbalanced methodology. It performs excellently in the speed of evaluation criterion, have a straightforward hardware implementation, is scalable and have no problem with the scene description criterion but on the other hand it performs terribly in the photorealism criterion. The methodology has difficulties in incorporating the occlusion solution and illumination into the process of hologram synthesis.

Although, the most important criterion photorealism is completely failed it is still a subject of continuos research. If a solution for photorealism support is found without disrupting the outcome of the remaining criteria, this methodology becomes one with the most potential.

5.2.6 MIT holovideo specific synthesis

The MIT Holovideo system is probably the first working holographic display. The concept of synthesis developed for this device is efficient and capable of delivering reasonable images. The original concept was further developed and support for illumination and occlusion was gradually incorporated into the system. The solution at its current state fulfils all evaluation criteria satisfactorily.

This fact rises a question why invest more efforts into finding other methodologies for digital hologram synthesis when MIT Holovideo is a working solution. The answer to this question is simple. MIT Holovideo supports only synthesis of the HPO holograms only. The HPO holograms have the range of valid viewing angles quite restricted and therefore HPO hologram synthesis can be considered as only half solution to the whole holographic display problem.

All synthesis methods dedicated to the MIT Holovideo system depends on the special features of the HPO hologram synthesis. Those special features, which are responsible for the phenomenal efficiency of the whole system, would not persist if the Holovideo system was transformed to support the full parallax hologram synthesis and reproduction. Therefore the significance of the methods described in the Sec. 3.1.6 for the purposes of this work is reduced. However, some of the concepts utilised in the HPO hologram synthesis are worth of incorporating into other methodologies, e.g. the basis diffraction footprint combination methodology may benefit from those concepts.

5.2.7 Final statement

According to the results gathered in the evaluation, the ray casting methodology and the Fourier hologram specific synthesis constitutes the most promising directions of the following research. Additional research path should follow the basis diffraction pattern combination methodology providing the serious drawbacks of the methodology are removed satisfactorily.

5.3 Future direction of work

In the previous section the ray casting methodology, and Fourier hologram specific synthesis were evaluated as the most suitable for further development. Regarding the ray casting methods the first working algorithms were already developed and published in [JHS06c] and [JHS06a] and overview of this work can be found in the Chap. 4. Further exploration of the ray casting approach showed quite potential in application for synthesising full parallax holograms. The GPU based solutions were successfully tested which indicates that HW supported solution is possible providing that the increase of computational power will be further improved.

The Fourier hologram specific synthesis will be thoroughly tested if it is able to compete with the ray casting methods in efficiency and the quality of the reconstructed images. The different optical setup required for physical reconstruction has to be included into the survey.

The last direction of research considered as secondary deals with the basis diffraction pattern combination methodology. The speed and simplicity of this approach which is suitable for hardware implementation is very tempting. This is however conditioned by the finding of solution for incorporating the occlusion and illumination into the process.

The roadmap of the research is summarised in the next section where detailed plan of the research is presented.

5.3.1 Road Map

This section presents the roadmap for the future work in the area of digital hologram synthesis. The final goal of this work was advertised several times already. It is a proposal of a complete holographic pipeline, that would replace the contemporary CGI one. There are however some assumptions that need to be fulfilled first. It is the existence of a genuine holographic display with sufficient size and resolution. There are prognosis that such display could be available in 10 years. This is reasonably near future and the proposed synthesis technology will be sufficiently mature to be promptly employed at that time.

The road map is visualised in Appendix B and should be used for reference. There are several milestones in the roadmap. The milestone M-01 constitutes the identification of the final synthesis methodology. This is the most important milestone, since all the following actions will be derived from it, like data structures and data formats, transfer protocols, HW interfaces etc. At the moment, there are several methodologies that has been chosen from the research done for this thesis and they are summarised above.

The next important milestone is the milestone M-02 that constitutes the technology major advancement. It means that the chosen synthesis methodology has been extended and modified to suit the needs of the final goal. This milestone also represents validation and verification of the technology of digital hologram synthesis. Several experiments were already done using contemporary SLM with positive results. Additional thorough verification need to be done for larger holograms. If adequate electronic devices won't be available then classic photographic material could be used for verification. The problem of transferring the digital hologram onto photographic material is being currently addressed.

The technology will be then formalised and summarised into a technology specification represented by the milestone M-03. This specification should include the plan for the development of the underlying approaches, i.e. speed improvement, hardware implementation design etc. This specification will rule

the next development which should end up with final and complete solution for the digital hologram synthesis ready for application. The milestone M-03 also represents the final doctoral thesis.

The following milestones represents rough estimation of the following work. The Milestone M-04 constitutes the prototype of a hardware holographic processor HPU - the fruit of the efforts to create hardware implementation of the hologram synthesis. This processor represents an analogy to the contemporary GPU's. The first prototypes of holographic display should appear after that and this event is represented by the milestone M-05.

The following work should involve merging of both synthesis and display technologies into one working monolithic solution. This merge involves synchronisation of the data structures and interfaces. Other technologies like holographic stream compression and transfer are also applied at this point. The successful merging constitutes the last milestone M-07 of this work and it is the point where commercial application should be started searched and this is the spot where science stops and industry comes into play.

Bibliography

[AR03] D. Abookasis and J. Rosen. Computer-generated holograms of three-dimensional objects synthesized from their multiple angular viewpoints. *J. Opt. Soc. Am. A*, 20(8):1537–1545, 2003. [cited at p. 55, 56, 57, 100]

[BW05] M. Born and E. Wolf. *Principles of Optics*. Cambridge University Press, 7th edition, 2005. [cited at p. 18, 28]

[CTL84] R. L. Cook, Porter T., and Carpenter L. Distributer ray tracing. *Computer Graphics*, 18(3):137–145, July 1984. [cited at p. 49]

[EO06] G.B. Esmer and L. Onural. Computation of holographic patterns between tilted planes. In Y. Denisyuk, V. Sainov, and E. Stoykova, editors, *Holography 2005: International Conference on Holography, Optical Recording, and Processing of Information*, volume 6252, page 62521K. SPIE, 2006. [cited at p. 28]

[FLB86] Ch. Frere, D. Leseberg, and O. Bryngdahl. Computer-generated holograms of three-dimensional objects composed of line segments. *Optical Society of America*, 3(5):726–730, May 1986. [cited at p. 61, 62]

[Gab49] D. Gabor. Microscopy by reconstructed wavefronts. In *Proc. Roy. Soc.*, 1949. [cited at p. 7]

[GD98] J. P. Grossman and William J. Dally. Point sample rendering. In *9th Eurographics Workshop on Rendering*, pages 181–192, 1998. [cited at p. 46]

[Goo05] J.W Goodman. *Introduction to Fourier Optics*. Roberts & Company Publishers, 4950 S. Yosemite Street, F2 #197, Greenwood Village, CO 80111, United States, third edition, 2005. [cited at p. 8, 16, 17, 18, 19, 20, 21, 22, 23, 26, 27, 28, 50, 56, 100]

[Har96] P. Hariharan. *Optical Holography: Principles, techniques and applications*. Cambridge University Press, second edition, 1996. [cited at p. 10, 12, 13, 14, 32, 34, 35, 37, 38, 100]

[Has06] A. Hast. Fast cosine function evaluation. Personal communication, 2006. [cited at p. 79]

[JHS06a] M. Janda, I. Hanák, and V. Skala. Digital HPO hologram rendering pipeline. In *EG2006 short papers conf. proc.*, pages 81–84, 2006. [cited at p. 73, 85]

[JHS06b] M. Janda, I. Hanák, and V. Skala. Holography principles. Technical Report DCSE/TR-2006-08, University of West Bohemia, December 2006. [cited at p. 7]

[JHS06c] Martin Janda, Ivo Hanák, and Václav Skala. Scanline rendering of digital hpo holograms and numerical reconstruction. In *Proc. SCCG*, volume 1, pages 66–73, 2006. [cited at p. 73, 80, 85]

[KDS01] M. Koenig, O. Deussen, and T. Strothotte. Texture-based hologram generation using triangles. In S. A. Benton, S. H. Stevenson, and T. J. Trout, editors, *Proc. SPIE Vol. 4296, p. 1-8, Practical Holography XV and Holographic Materials VII, Stephen A. Benton; Sylvia H. Stevenson; T. John Trout; Eds.*, pages 1–8, 2001. [cited at p. 63]

[Kra04] F. Krausz. Photonics: Lecture notes, 2004. [cited at p. 10, 15, 16, 26, 100]

[LAR01] Y. Li, D. Abookasis, and J. Rosen. Computer-generated holograms of three-dimensional realistic objects recorded without wave interference. *Appl. Opt.*, 40(17):2864–2870, 2001. [cited at p. 55, 65]

[LBL02] D.R. Luke, J.V. Burke, and R.G. Lyon. Optical Wavefront Reconstruction: Theory and Numerical Methods'. *SIAM Review*, 44:169–224, 2002. [cited at p. 17, 18, 21, 22]

[LHJ68] L.B. Lesem, P.M. Hirsch, and J.A. Jordan. Computer synthesis of holograms for 3-d display. *Communications of the ACM*, 11(10):661–673, October 1968. [cited at p. 51]

[Luc94] M. Lucente. *Diffraction-Specific Fringe Computation for Electro-Holography.* PhD thesis, MIT, 1994. [cited at p. 7, 12, 65, 66, 68, 100]

[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. [cited at p. 21]

[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 B. Javidi and F. Okano, editors, *Three-Dimensional TV, Video, and Display. Edited by Javidi, Bahram; Okano, Fumio. Proceedings of the SPIE, Volume 4864, pp. 122-128 (2002).*, pages 122–128, 2002. [cited at p. 23]

[Pho75] Bui Tuong Phong. Illumination for computer generated pictures. In *Communications of the ACM*, volume 18, pages 311–317, June 1975. [cited at p. 80]

[Ple03] W.J. Plesniak. Incremental update of computer-generated holograms. *Optical Engineering*, 42:1560–1571, June 2003. [cited at p. 67]

[PM03] Christoph Petz and Marcus Magnor. Fast hologram synthesis for 3d geometry models using graphics hardware. In *SPIE*. SPIE, 2003. [cited at p. 60]

[RBD+99] A. Ritter, J. Böttger, O. Deussen, M. König, and T. Strothotte. Hardware-based Rendering of Full-parallax Synthetic Holograms. *Appl. Opt.*, 38:1364–1369, March 1999. [cited at p. 61, 62]

[SIY04] Y. Sando, M. Itoh, and T. Yatagai. Full-color computer-generated holograms using 3-D Fourier spectra. *Optics Express*, 12:6246–+, 2004. [cited at p. 57, 58, 59]

[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. [cited at p. 28, 29, 53]

[Und97] J. S. Underkoffler. Occlusion processing and smooth surface shading for fully computed synthetic holography. In S. A. Benton and T. J. Trout, editors, *Proc. SPIE Vol. 3011, p. 19-30, Practical Holography XI and Holographic Materials III, Stephen A. Benton; T. John Trout; Eds.*, pages 19–30, 1997. [cited at p. 69, 70, 101]

[Wat00] Alan H. Watt. *3D Computer Graphics*. Addisn-Wesley, third edition, 2000. [cited at p. 75]

[Whi80] T. Whitted. An improved illumination model for shaded display. *Communications of the ACM*, 23(6):–348, June 1980. [cited at p. 49]

[YAC02] L. Yu, Y. An, and L. Cai. Numerical reconstruction of digital holograms with variable viewing angles. *Optics Express*, 10:1250–+, 2002. [cited at p. 25]

Appendices

Appendix A

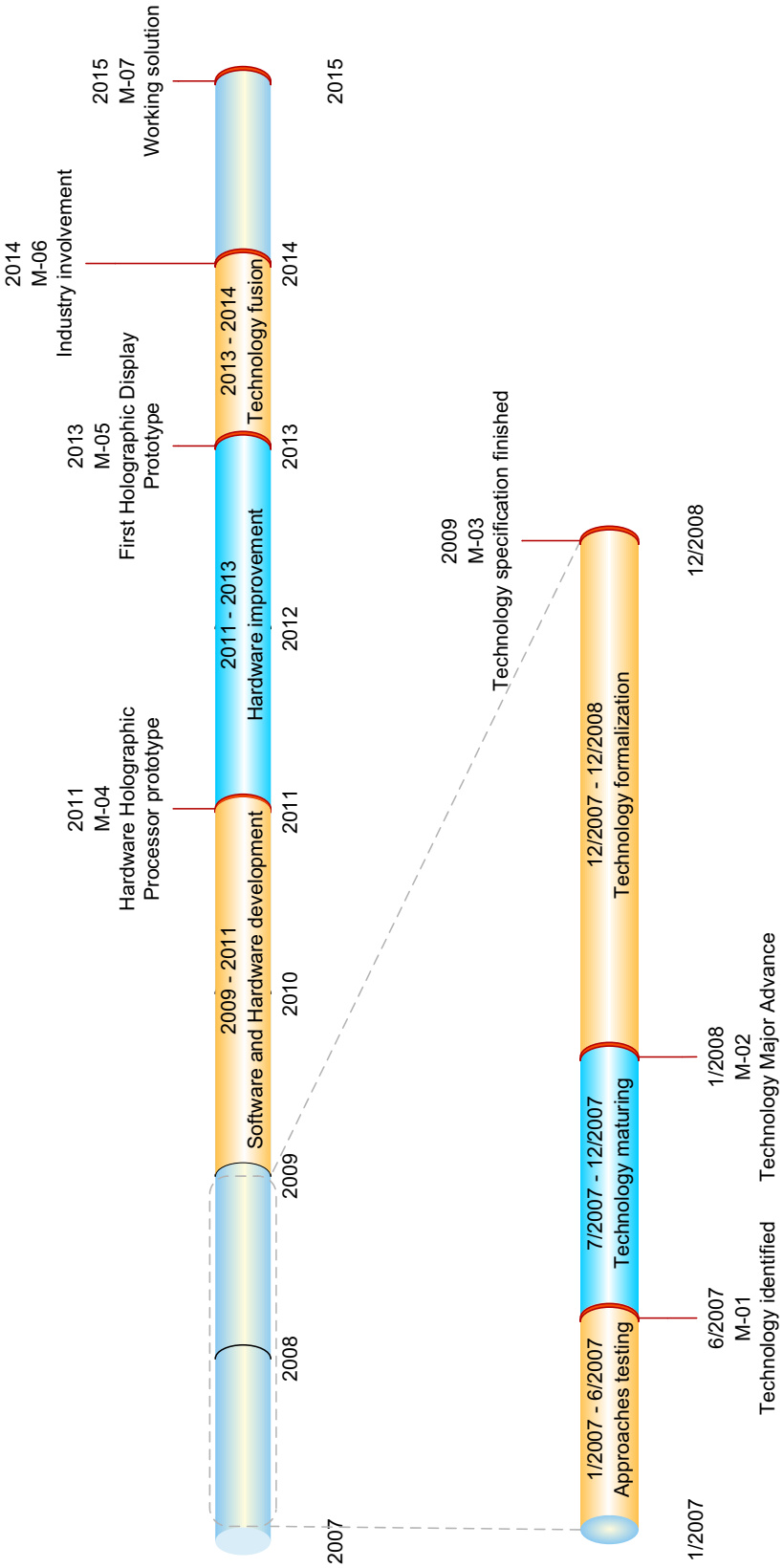
Synthesis methods comparison

	Photorealism	Speed of evaluation	Scalability	HW implementation	Scene description
Diffraction integral evaluation	★★☆☆	☆☆☆☆	★★★☆☆	★★★★☆	★☆☆☆☆
Ray casting	★★★★	★☆☆☆	★★★★★	★★★★☆	★★★★★
Diffraction between two coplanar planes	★☆☆☆	★★☆☆☆	★★★☆☆	★★★☆☆	★★☆☆☆
Fourier hologram synthesis	★★★★	★☆☆☆	★★☆☆☆	★★★☆☆	★★★★★
Basis diffraction footprint combiantion	★☆☆☆	★★★☆☆	★★★☆☆	★★★★☆	★★★☆☆
MIT Holovideo	★★★☆	★★★★★	★★★★★	★☆☆☆☆	★★★★☆

Table A.1: Synthesis methods evaluation

Appendix B

Research Road Map


2015
2015
M-07
Working solution
2014
M-06
Industry involvement
2013 - 2014
Technology fusion
2014
2013
M-05
First Holographic Display
Prototype
2011 - 2013
Hardware improvement
2013
2011
M-04
Hardware Holographic
Processor prototype
2009 - 2011
Software and Hardware development
2012
2010
2009
2008
2007
2009
M-03
Technology specification finished
12/2007 - 12/2008
Technology formalization
12/2008
7/2007 - 12/2007
Technology maturing
1/2008
M-02
Technology Major Advance
1/2007 - 6/2007
Approaches testing
6/2007
M-01
Technology identified
1/2007

List of Symbols and Abbreviations

Abbreviation	Description	Definition
HPO	horizontal parallax only	page 87
SLM	spatial light modulator	page 39
CCD	coupled charge detector	page 71
FFT	fast fourier transform	page 24
GPU	graphics processing unit	page 39
STAR	state of the art	page 81
MIT	massachusetts institute of technology	page 65
CGI	computer graphics imagery	page 43
LCD	liquid crystal display	page 3
CRT	cathode ray tube	page 3
CAD	computer aided design	page 3
LASER	light amplification by stimulated emission of radiation	page 7

List of Figures

List of Tables

List of Algorithms

Index