Modeling, design and optimization of computer-generated holograms with binary phases

Thèse

Jing Wang

Doctorat en physique
Philosophiæ doctor (Ph. D.)

Québec, Canada

Modeling, design and optimization of computer-generated holograms with binary phases

Thèse

JING WANG

Sous la direction de :

Yunlong Sheng

Résumé

L’hologramme généré par ordinateur (HGO) a été démontré à jouer un rôle important depuis son invention par Lohmann dans les années 1960 dans de nombreuses applications telles que l’ingénierie du front d'onde, l’éclairage structuré et l’affichage optique, etc. Dans le travail de thèse ci-présent, la modélisation, la conception et l’optimisation d’HGO avec des phases binaires sont étudiées.

Nous avons examiné un système pratique de projection d’image avec certaines spécifications de travail, par exemple, une distance de travail de 40 cm, une profondeur de champ de 10 cm et un angle de diffraction de 53 degré pour une longueur d’onde de travail de 632 nm, et ensuite conçu et optimisé un hologramme de phase binaire en passant par une recherche directe binaire pour ce système d’image. L’hologramme a été fabriqué par la lithographie à faisceau d’électrons. Pour atteindre l’angle de diffraction requis, nous avons discuté de l’architecture optique dans le système de projection d’image holographique. L’HGO conçu et le système de projection d’image holographique ont été validés expérimentalement par reconstruction optique.

Étant donné que les pixels finiront par se regrouper pour former des ouvertures polygonales en hologramme, qui peut être vu clairement dans le processus de recherche directe binaire, nous avons proposé une nouvelle approche pour la conception directe des ouvertures polygonales basée sur la disposition triangulaire en HGO de grande taille en pixels. La diffraction de l’ouverture a été calculée par la transformation analytique d’Abbe. L’image reconstruite peut être exprimée comme une addition cohérente de motifs de diffraction à partir de tous les bords droits d’orientations et de longueurs différentes. Une optimisation en deux étapes comprenant l’algorithme génétique avec la recherche locale de codage des phases binaires des ouvertures, suivie par la recherche directe de co-sommets flottants des ouvertures triangulaires élémentaires a été développée.

Nous avons en outre proposé une disposition d’ouverture quadrilatérale, qui fournit plus de degrés de liberté et peut former des ouvertures polygonales plus diverses en hologrammes. L’algorithme génétique parallèle avec la recherche locale a été adopté dans une première étape pour assigner des phases binaires, et la recherche directe a ensuite été utilisée pour

optimiser des emplacements de co-sommets d'ouvertures quadrilatérales lors de la deuxième étape. Trois schémas différents pour l'algorithme en deux étapes ont été discutés pour fournir des moyens flexibles afin d’équilibrer la performance de l’optimisation et la durée nécessaire.

Abstract

The computer-generated hologram (CGH) has been demonstrated to play an important role, since its invention by Lohmann in 1960s, in many applications such as wavefront engineering, structured illumination and optical display, etc. In this thesis, the modeling, design and optimization of CGH with binary phases are studied.

We considered a practical projection image system with certain working specification, e.g. working distance of 40 cm, depth of field of 10 cm and a diffraction angle of 53 degree for 632 nm working wavelength, and then designed and optimized a binary-phase hologram by direct binary search for this image system. The hologram was fabricated by E-beam lithography. To achieve the required diffraction angle, we discussed the optical architecture in holographic projection image system. The designed CGH and holographic projection image system were validated experimentally by optical reconstruction.

Since the pixels will eventually cluster to form polygonal apertures in hologram, which can be seen clearly during the process of direct binary search, we proposed a new approach to directly design polygonal apertures based on triangular layout in CGH of a large number of pixels. The diffraction of aperture was calculated by analytical Abbe transform. The reconstructed image can be expressed as a coherent addition of diffraction patterns from all the straight edges of different orientations and lengths. A two-step optimization including genetic algorithm with local search for encoding binary phases of apertures, followed by direct search for floating covertices of the elementary triangular apertures was developed.

We further proposed a quadrilateral aperture layout, which provides more degrees of freedom and can form more diverse polygonal apertures in holograms. The parallel genetic algorithm with local search was adopted to assign binary phases in the first step, and direct search was then used to optimize of locations of covertices of quadrilateral apertures in the second step. Three different schemes for the two-step algorithm were discussed to provide flexible ways to balance the optimization performance and time cost.

Contents

List of figures

List of tables

Abbreviations

HGO	Hologramme généré par ordinateur
CGH	Computer-generated hologram
2D/3D	Two-dimensional/Three-dimensional
CCD	Charged coupled device
DH	Digital holography
DOE	Diffractive optical element
AR	Augmented reality
VR	Virtual reality
SLM	Spatial light modulator
FoV	Field of view
DoF	Depth of field/focus
SBWP	Space bandwidth product
DBS	Direct binary search
SA	Simulated annealing
GA	Genetic algorithm
IFTA	Iterative Fourier transform algorithm
CMOS	Complementary metal oxide semiconductor
HGA	Hybrid genetic algorithm
PGA	Parallel genetic algorithm
CNC	Computer numerical control
DLW	Direct laser writing

CAD	Computer aided design
EBL	Electron beam lithography
UV	Ultraviolet
LCOS	Liquid crystal on silicon
AM	Amplitude modulating
PM	Phase modulating
DFT	Discrete Fourier transform
FFT	Fast Fourier transform
MSE	Mean squared error
DMD	Digital micromirror device
RMS	Root-mean-square
GPU	Graphical processing unit
FDTD	Finite-Difference Time-Domain
RCWA	Rigorous Coupled Wave Analysis

Dedicated to my beloved family

for their love, support, encouragement and sacrifice

Acknowledge

I am deeply indebted to Professor Yunlong Sheng, my research supervisor during my PhD program at Université Laval, Quebec, for his great support and invaluable guidance of these work, and for his advice and ideas which gave me new insights in ways on how to do research.

I greatly appreciate my beloved family for providing me with courage and strength over the years. They have been always so supportive of decisions I made, even sometimes they probably don't fully understand what I have done. Without their endless love and tireless support, it would not have been possible to finish my PhD.

I am also grateful to Y. Gravel, Dr. L. Yu and Y. Yang in Prof. Sheng's research group, for stimulating discussions and great assistance in both work time and spare time, and to Y. Kang, Dr. H. Liang, Q. Liang, Y. Dong at Ulaval for being (or, having been) there and making (or, having made) my days in Quebec unforgettable.

I would like to extend my gratitude to Dr. Z. Jan Jakubczyk at Optiwave Systems Inc., Ottawa. He offered me a work position as research scientist, and also helped me a lot after I joined the team on Dec. 1st, 2015. Special thanks go to Kevin Chu and Dr. Steve Dods in the company for their kind help with my work when I was with Optiwave in Ottawa. A wonderful experience there!

I acknowledge my thesis evaluation panel, Prof. Simon Thibault, Prof. Pierre Marquet, Prof. Yunlong Sheng at Ulaval, and Prof. Pascal Picart at Le Mans Université in France, for accepting to evaluate my work and giving constructive feedback, and the program director Prof. Laurent Drissen for organizing and presiding over my oral defense.

I want to thank all the staff in our department (Département de physique, de génie physique et d'optique) and COPL (Centre d'optique, photonique et laser) who have ever helped me when I was on campus between 2012 and 2015.

2019.04.24 in Quebec

Foreword

Two scientific papers, coauthored by Jing Wang & Yunlong Sheng and published on peer-reviewed international journals, are presented respectively in Chapter 3 and Chapter 4 of this thesis:

1) Jing Wang, Yunlong Sheng, "Computer-Generated Very Large Space-Bandwidth Product Binary Hologram for Laser Projector," IEEE Transactions on Industrial Informatics, Vol. 12, No. 1, pp. 179-186 (2016).
2) Jing Wang, Yunlong Sheng, "Design quadrilateral apertures in binary computer-generated holograms of large space bandwidth product," Applied Optics, Vol. 55, Issue 27, pp. 7636-7644 (2016).

For both papers, Prof. Sheng and I discussed and conceived ideas together; I implemented the work of modeling, design and optimization under the supervisor of Prof. Sheng; I wrote the first draft of each paper and Prof. Sheng finished the final version; I am the first author and Prof. Sheng is corresponding author of both papers.

The contents in Chapter 3 are basically from the paper published on IEEE TII with additional details. Specifically, Equations (3.7) – (3.9), Figure (3.3) – (3.5) and some related words are newly added in Chapter 3.

The contents in Chapter 4 are mainly from the paper published on AO with some supplementary. In Chapter 4, Figure (4.3), (4.10), (4.11) and subsection of Figure (4.6 a) are newly presented, as well as Equation (4.12) – (4.14) and Table (4.1).

General introduction

Historical events of holography

The record and storage of objects, especially three-dimensional (3D) objects, had been a recurring goal of humankind after the photography was invented in 1830s. It was not really achieved until holography was discovered by Denis Gabor in 1948 [1], when he worked to address the problem introduced by the spherical aberrations of the electron lenses, which limited the resolving ability of electron microscopes [1, 2]. The word "holography" was assembled from Greek words "holos" (whole) and "graphein" (to write). Thus, holography can be defined as a technique to write or record the whole optical information from an object source for the reconstruction of the original object later. "For his invention and development of the holographic method", Denis Gabor was awarded the Nobel Prize in Physics in 1971 [3].

The holographic process basically consists of two steps: record and reconstruction. In the first step, a holographically stored image, referred to as hologram, is produced by the interference between a wave field scattered from an object and a coherent reference wave. A hologram is usually recorded on a two-dimensional (2D) surface, for example, a photosensitive film. But it contains the whole information, i.e. both the amplitude and the phase, of the 3D object field. The information is encoded in the form of interference fringes with high spatial frequencies, which is usually invisible to human eyes. In the second step, the hologram is illuminated by original reference wave, and the diffracted light by hologram propagates to recover the object wave to form an image of object.

In Gabor's original set-up [1], the reference wave was incident normal to the recording medium and a predominantly transparent object was in its path. The axes of both the object wave and the reference wave were parallel. This is known as on-axis hologram. This set-up is simple, but it only works when the object is small enough and sufficiently transparent so that it will not disturb the reference wave significantly. Besides, the reconstruction of this hologram gives a real image superimposed on the undifffracted part of wave and a so-called "twin image" on the same optical axis.

In 1956, Lohmann proposed the “single-sideband” holography [4] to handle the twin-image problem of Gabor holograms by the combination of communications theoretical and physical views of optics.

With the advent of laser by Theodore H. Maiman in 1960 [5] and the invention of off-axis hologram by Leith and Upantnieks in 1962 [6], high quality holograms could be produced and began to capture the interest of world. In their set-up, the reference waves with an oblique angle respect to normal of the recording medium did not pass through the object, which spatially separated the different diffraction orders and allows the capture of opaque objects.

The development of computer technology makes it possible to transfer the recording process and/or the reconstruction process onto the computer platform. In the middle of 1960’s Lohmann etc. realized the simulation of optical holograms by digitally generated binary transparencies, which is a big step forward in the path towards widespread applications [7-9]. This new approach was referred to computer-generated holography and the corresponding hologram was called computer-generated hologram (CGH). The CGH has the advantage that the object could be synthetic or fictitious. Its design can be optimized mathematically rather than experimentally.

Another important progress is the direct recording of holograms with Charged Coupled Devices (CCDs) by Schnars etc. in 1990s [10]. This method is known to us today as Digital holography (DH). Sometimes, computer-generated holography and digital holography are considered as the same concept, which is used to correspond to conventional optical holography.

Computer-generated holography and conventional optical holography

The first computer-generated holograms were binary holograms invented by A. Lohmann [8, 9]). They were printed on a computer line printer, then they were optically reduced and reconstructed in optical set ups using coherent laser illumination. The computer-generated holography shares the same basic principle and concept with the conventional optical holography. However, some differences [11-15] need to be noted.

From what was introduced above, the record process of conventional optical holography is by optical interference, and the reconstruction of object is an optical diffraction process. The object is usually a real one; the hologram, i.e. the interference pattern is recorded by light-sensitive medium.

For the computer-generated holography, the record process is usually numerically synthesized by inverse wave propagation theory and then encoded to hologram. The reconstruction process can be either optical or numerically diffractions. Sometimes, the numerically record and numerically reconstruction is specially referred to as digital holography; only the numerically record and optical reconstruction is named as computer-generated holography. Besides, the object can be anything you may imagine; it is not necessary to be limited to the physically existing objects. It also can radiate or be illuminated by an external source. More importantly, the hologram can be fabricated or loaded to a spatial light modulator (SLM) for dynamically display.

Categories of CGHs

Over the years, computer-generated holograms (CGHs) have demonstrated to be important components for its broad range of applications [11], such as diagnostic and testing (acoustic mapping of the earth, determining particle sizes and scattering properties, analyzing vibrations, visualizing aberrations etc.), imaging and display (map displays, artist display, 3D display, image processing and deblurring, stereoscopic display etc.), structured illumination, augmented reality (AR)/virtual reality (VR), imaging system, optical storage, sensors and security etc. Many kinds of CGHs have been proposed for these different applications.

In generally, depending on the different display types in time domain, the CGHs can be categorized into two groups [12-15]:

- static CGH. For static CGH, a permanent hologram is produced. After fabricated on the substrate, the static hologram is also called as diffractive optical elements (DOEs).
- dynamic CGH. For dynamic CGH, it is usually loaded onto spatial light modulators (SLMs), which are essentially small, television-like display devices.

Depending on the material with which the CGH is to be fabricated, there are typically three types [12, 13]:

- Phase-only CGH, where the material modulates only the phase of an incoming wavefront, and the transmittance amplitude is unity;
- Amplitude-only CGH, where the material modulates only the amplitude of an incoming wavefront, and the transmittance phase is constant;
- Complex-amplitude CGH, where such a material modulates both the amplitude and phase of the incoming wavefront.

The CGHs can be grouped as the way they utilize resources [14]:

- Point-oriented CGHs, in which each pixel is uniform; there is no sub-element. Each pixel will be assigned a value of the transmittance which corresponds to pixel value in the actual hologram.
- Cell-oriented CGHs, on the other hand, manipulate the sub-structure in each hologram cell so that each cell consisting of pixels controls amplitude and phase in some ways.

Depending on where the image reconstructed is observed, there are several types of CGH [14, 15], among which the most important two are:

- Fourier CGH: the image is reconstructed at back focal-plane of the second lens of the 4-F imaging system.
- Fresnel CGH: the image is reconstructed in at a distance far enough from the CGH so that the parabolic (or, Fresnel) approximation can be used.

Current research of CGHs

Since one of the main objectives of holography is to record and display of 2D/3D object, current research of CGH usually refers to the following 4 topics: object data acquisition, object wave representation, hologram encoding and image reconstruction.

The first research topic is about how to acquire the geometry of the object, especially 3D object. One of approaches is to acquire 2D projections of the object from different perspective viewpoints using 2D camera array [16] or by moving one single camera [17]. Another one is

to adopt 2D-plus-depth cameras [18] to get 2D views and depth information. The acquired object data can then be directly used for CGH.

The second one is to represent the object wave. The point-source approach [15] is one of the most commonly used to compute the object wave, which is the sum of spherical waves by each object point. Another technique for object wave is wave-field approach, which can be based on a layered model [19, 20] or a polygonal model [21] of 3D object. In the former model, the 3D object is sliced into a set of layers, which operates as a surface source of light. While in the latter one, 3D scenes are described as a set of oriented polygons acting as a surface source of light. The associated occlusion processing [22, 23], surface shading [24] and computation time reduction [25, 26] are widely studied as well.

The third topic is how to encode hologram. The object wave computed is usually a complex-valued field. However, to be displayed on a screen or printed onto a transparency, the real positive values of hologram must be quantized to some levels. The previously mentioned phase-only hologram [27], amplitude-only hologram [28] and complex-amplitude hologram [29] are three encoding types. Another problem related is to use data compression techniques [30, 31] to reduce the amount of holographic information to be stored and transmitted since it could be very huge.

The last one is about image reconstruction. Current SLMs have limited resolutions and pixel pitches much larger than wavelength used, which prevent SLMs from providing a large diffractive angle [32, 33].

Design procedures of CGHs

As mentioned earlier, CGHs are numerically synthesized by inverse wave propagation theory. The design procedures of CGHs can basically be divided into three parts [12, 13]:

First, analysis the design problem and understand the physics behind it. In this part, the main task involves calculation of the fields that should be generated on the hologram plane produced by the object, i.e. the inverse wave propagation of the desired light pattern. One usually need to know if the scalar diffraction is accurate enough to solve the problem and may further decide the possible Fresnel or Fourier transform on the object fields. If scalar diffraction theory is not valid for some cases, full vector theory needs to be employed.

The second part of the design is the choice of encoding way into the hologram plane and then define an appropriate optimization algorithm. The fields calculated on the hologram plane in the last step are usually complex fields. One should choose the proper modulation method, which will be used to encode into hologram to form a desired transparency. The adoption of optimization algorithm needs to comprehensively consider the modulation way, the type of CGH, optimization speed and performance, computation resources etc. The main goal of the optimization is to obtain a hologram that will generate a reconstruction pattern as close as possible to the desired pattern.

The third part is to execute the design and fabricate the hologram. The fabrications of the hologram usually bring some extra errors due to the machine tolerance. In fact, the resolution of the machine plays on important role on determining the encoding and the optimization algorithm, which need to be considered in advance.

The motivation

Although great development and success has been made since CGH's invention, with the advancement of modern fabrication technology and emergence of application in new fields, some challenges still exist when design the CGHs. This thesis aims to model, design, and optimize CGHs regarding the two specific issues below.

One of the issues is how to design CGH for the practical projection system. Besides consideration of reconstruction error, diffraction efficiency of CGH and stray light from unwanted high diffraction orders, the specifications of practical projection system usually force designers to look upon more factors, typically including working distance, working wavelength (range), diffraction angles - field of view (FoV), and depth of field/focus (DoF) etc. Meanwhile, the abilities of facility equipment for fabricating CGH should also be considered before the design, such as the resolution can be achieved, the maximum capacity of data can be read, the speed of wring and sometimes even the financial cost etc. All these factors need to be proper treated and weighted, which greatly increase the complexity of designing CGH in a practical projection system.

Another issue concerned is that how to design and optimize CGH of large space bandwidth product (SBWP). With the advancement of industrial fabrication technology, the minimum

feature size can achieve is far less than the order of visible wavelength. For example, with the e-beam lithography, the pixel size can reach to 50 nm, or even 5 nm. With a pixel size of 50 nm, a CGH with one-inch square size will have 258 gigapixels, which provides a huge number of degree of freedom. The conventional optimization algorithm for CGH can not well handle. The main challenge is that how to manage such a huge number of degree of freedom to achieve the best performance in the scope of scale diffraction theory.

Organization of thesis

With respect to the content of the thesis, it will be organized as follows.

A general introduction on the holography and CGH is given at the beginning, including a short review of historical events, basic idea of holography, a list of categories of CGH and current research as well. This part also briefly introduces design procedures of CGH and the issues that this thesis attempts to answer, i.e. the design and optimization of CGH applied to a practical projection imaging system, and CGH with huge number of degree of freedom. The organization of thesis is presented in the end of the introduction.

In Chapter 1, the fundamental of holography and the underlying theory used - scalar diffraction theory are detailed. Besides, several classic algorithms of designing CGHs - DBS，SA, GA and IFTA, are introduced, together with two industrial technologies - dot matrix hologram and holographic prints. Also, four fabrication technologies for CGH are briefly presented.

In Chapter 2, we talked about how to design a CGH for a structured-light projector with certain specifications. This projector system has a working distance of 40 cm, and a 10 cm depth of field (DoF). The pattern angle is around 53 degree. The wavelength of used optical source is 632 nm. The direct binary search (DBS) algorithm is demonstrated first by designing a Fourier hologram with binary phases and then further applied to synthesizing Fresnel hologram with binary phases, which meet the requirements of the projector. The consideration of optical architecture design in the projection system was discussed. The fabrication of hologram and experimental results were shown as well.

In Chapter 3, we proposed a new approach for designing binary CGH with arbitrary-shaped polygonal apertures. With this method, the high number of degrees of freedom available in CGH of large space bandwidth product (SBWP) can be fully explored. The Abbe transform was introduced and used to compute the diffraction pattern. A two-step optimization algorithm including hybrid GA for assigning binary phase and direct search for floating co-vertices of the elementary triangular apertures was developed to reconstruct image with high performance.

In Chapter 4, we further investigate to directly design quadrilateral aperture in binary CGHs of large SBWP, which exploit higher number of degree of freedom compared with that in CGHs based on triangular aperture layout. Coarse grained parallel GA with local search is used in the first step to assign phases and direct search is adopted in the second step to optimize the positions of vertices. Three schemes of this two-step algorithm were discussed to give us a flexible way to balance between the time cost and optimal results.

Conclusion. A summary of achieved work and future work was presented.

Chapter 1 Theory, algorithms and fabrications of holograms

1.1 Fundamental of holography

The fundamental problem addressed by holography is recording and later reconstructing, both the amplitude and the phase of an optical wave from a coherently illuminated object. A typical setup is shown in Fig. 1.1. Light with sufficient coherence is split into two waves of reduced amplitude by a beam splitter. The first wave illuminates the objects, is scattered at the object surface towards the recording medium. The second wave, i.e. the reference wave, directly illuminates the light sensitive medium. The waves interfere with each other to produce a characteristic interference pattern. In classical photographic holography the interference pattern is recorded on a photosensitive material such as silver halide films or plates and rendered permanent by wet chemical development of film. In digital holography the interference pattern is recorded directly onto an electronic sensor array such as a Charged Coupled Devide (CCD) or Complementary Metal Oxide Semiconductor (CMOS). The recorded interference pattern is the hologram [13-15].

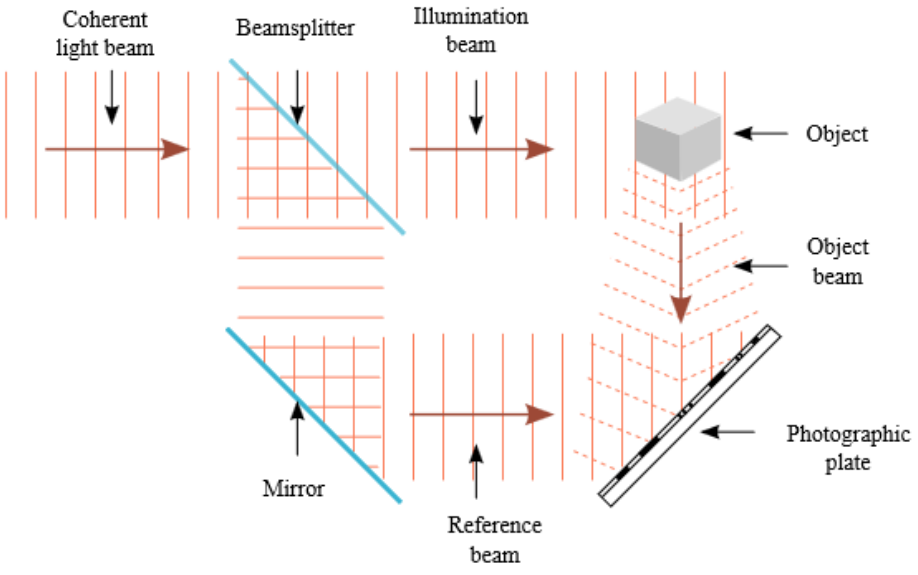


Fig. 1.1 A typical setup for holographic recording of an object by interference of two coherent beams

The holographic process is described mathematically by the formalism of interference of light. If the complex amplitude of the object wave is described by

$$\tilde{E}_O(x,y) = A_O(x,y)\exp\left[i\varphi_O(x,y)\right] \tag{1.1}$$

with real amplitude A_O and phase φ_O. And the complex amplitude of the reference wave with real amplitude A_R and phase φ_R is represented by

$$\tilde{E}_R(x,y) = A_R(x,y)\exp\left[i\varphi_R(x,y)\right] \tag{1.2}$$

Both waves interfere at the surface of the recording medium and the resultant intensity is calculated by

$$\begin{aligned} I(x,y) &= \left|\tilde{E}_O(x,y) + \tilde{E}_R(x,y)\right|^2 \\ &= \tilde{E}_O(x,y)\tilde{E}_O^*(x,y) + \tilde{E}_O(x,y)\tilde{E}_R^*(x,y) + \tilde{E}_R(x,y)\tilde{E}_O^*(x,y) + \tilde{E}_R(x,y)\tilde{E}_R^*(x,y) \\ &= A_O^2(x,y) + A_R^2(x,y) + A_O(x,y)A_R(x,y)\exp\left[i(\varphi_O - \varphi_R)\right] + A_O(x,y)A_R(x,y)\exp\left[-i(\varphi_O - \varphi_R)\right] \end{aligned} \tag{1.3}$$

while the first two terms of this expression depend only on the intensities of the individual waves, the third and fourth depends on their relative phases. Thus, information about both the amplitude and phase of $\tilde{E}_O$ has been recorded.

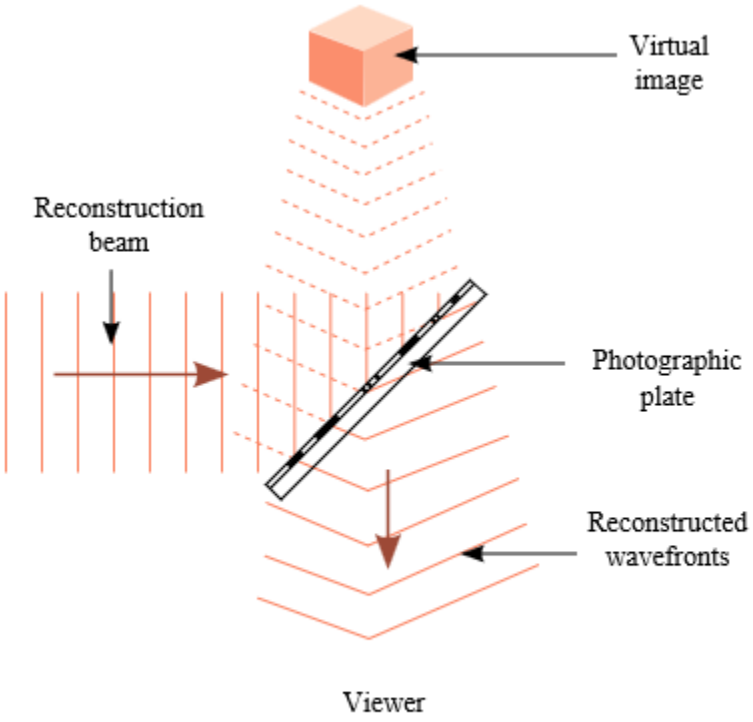
The material used to record the interference pattern will be assumed to provide a linear mapping of intensity. For simplicity, suppose $t(x,y) = 1 \cdot I(x,y)$.

Once the amplitude and phase information about the object wave have been recorded, it remains to reconstruct that wave. If the coherent reconstruction wave $\tilde{E}_C$ is represented by

$$\tilde{E}_C(x,y) = A_C(x,y)\exp\left[i\varphi_C(x,y)\right] \tag{1.4}$$

Then the light diffracted by the hologram is evidently

$$\begin{aligned} \tilde{E}_D(x,y) &= \tilde{E}_C(x,y)\cdot t(x,y) \\ &= \left[A_O^2(x,y) + A_R^2(x,y)\right]A_C(x,y)\exp\left[i\varphi_C(x,y)\right] + A_O(x,y)A_R(x,y)A_C(x,y)\exp\left[i(\varphi_O + \varphi_C - \varphi_R)\right] \\ &\quad + A_O(x,y)A_R(x,y)A_C(x,y)\exp\left[i(\varphi_R + \varphi_C - \varphi_O)\right] \end{aligned} \tag{1.5}$$



Released under public domain by http://en.wikipedia.org/wiki/User:Wykis

Fig. 1.2 An illustration for holographic reconstruction of an object

Note that if $\tilde{E}_C$ is simply an exact duplication of the original reference wave $\tilde{E}_R$,

$$\tilde{E}_C(x,y) = \tilde{E}_R(x,y) = A_R(x,y)\exp\left[i\varphi_R(x,y)\right] \tag{1.6}$$

Then, Eq. (1.5) can be rewriten as

$$\tilde{E}_D(x,y) = \left[A_O^2(x,y) + A_R^2(x,y)\right]\tilde{E}_R(x,y) + A_R^2(x,y)\tilde{E}_O(x,y) + A_R^2(x,y)\exp\left(i2\varphi_R\right)\tilde{E}_O^*(x,y) \tag{1.7}$$

The first term on the right side of Eq. (1.7) is the reference wave multiplied by a constant factor. It represents the non-diffracted wave passing through the hologram (zero diffraction order). The second term is the reconstructed object wave and forms the virtual image, as shown in Fig. 1.2. The real factor A_R^2 only influences the brightness of the image. The third term generated a distorted real image of the object (with 3D shapes). The reason for the distortion of the real image is the spatially varying complex factor $\exp\left(i2\varphi_R\right)$, which modulates the image forming conjugate object wave $\tilde{E}_O^*$.

In a similar fashion, if $\tilde{E}_C$ is chosen as the conjugate of the original reference wave $\tilde{E}_R$,

$$\tilde{E}_C(x,y) = \tilde{E}_R^*(x,y) = A_R(x,y)\exp\left[-i\varphi_R(x,y)\right] \quad (1.8)$$

Then we have

$$\tilde{E}_D(x,y) = \left[A_O^2(x,y) + A_R^2(x,y)\right]\tilde{E}_R^*(x,y) + A_R^2(x,y)\exp\left(-i2\varphi_R\right)\cdot\tilde{E}_O(x,y) + A_R^2(x,y)\tilde{E}_O^*(x,y) \quad (1.9)$$

The third term is proportional to the conjugate of the original object wave $\tilde{E}_O^*$, which represents an undistorted real image.

1.2 Scalar diffraction theory

1.2.1 Wave equation

According to the differential form of Maxwell's equations for source-free medium [34, 35],

$$\begin{aligned}
&\vec{\nabla}\times\vec{E} = -\frac{\partial\vec{B}}{\partial t},\\
&\vec{\nabla}\times\vec{H} = \frac{\partial\vec{D}}{\partial t},\\
&\vec{\nabla}\cdot\vec{D} = 0,\\
&\vec{\nabla}\cdot\vec{B} = 0.
\end{aligned} \quad (1.10)$$

Taking the curl of Eq. (1.10.1) produces

$$\vec{\nabla}\times\vec{\nabla}\times\vec{E} = \vec{\nabla}\times\left(-\frac{\partial\vec{B}}{\partial t}\right) = -\mu\vec{\nabla}\times\frac{\partial\vec{H}}{\partial t} = -\mu\frac{\partial}{\partial t}\left(\vec{\nabla}\times\vec{H}\right) = -\mu\frac{\partial}{\partial t}\left(\frac{\partial\vec{D}}{\partial t}\right), \quad (1.11)$$

With D=εE, after Substituting Eq. (1.10.2) to Eq. (1.11), we get

$$\vec{\nabla}\times\vec{\nabla}\times\vec{E} = -\mu\varepsilon\frac{\partial^2\vec{E}}{\partial t^2} \quad (1.12)$$

with the property $\vec{\nabla}\times\vec{\nabla}\times\vec{E} = \vec{\nabla}\left(\vec{\nabla}\cdot\vec{E}\right) - \vec{\nabla}^2\vec{E}$, , we have

$$\vec{\nabla}^2\vec{E} = \mu\varepsilon\frac{\partial^2\vec{E}}{\partial t^2} \quad (1.13)$$

where $\upsilon = 1/\sqrt{\mu\varepsilon}$ is the velocity of the wave in the medium.

Suppose ψ represent a component E_x, E_y or E_z of the electric field E.

$$\frac{\partial^2\psi}{\partial x^2}+\frac{\partial^2\psi}{\partial y^2}+\frac{\partial^2\psi}{\partial z^2}=\frac{1}{\upsilon^2}\frac{\partial^2\psi}{\partial t^2} \tag{1.14}$$

The solution will be in form of

$$\psi(x,y,z,t)=\psi_p(x,y,z)\exp(i\omega_0 t) \tag{1.15}$$

Substituting Eq. (1.15) into Eq. (1.14) produces the well-known Helmholtz equation [35]:

$$\frac{\partial^2\psi_p}{\partial x^2}+\frac{\partial^2\psi_p}{\partial y^2}+\frac{\partial^2\psi_p}{\partial z^2}+k_0^2\psi_p=0 \tag{1.16}$$

where $\vec{k_0}$ is the propagation vector, and $\left|\vec{k_0}\right| = k_0 = \sqrt{k_{0x}^2+k_{0y}^2+k_{0z}^2} = \omega_0/\upsilon$, is called wave number.

1.2.2 Spatial frequency transfer function

Denote $\tilde{\psi}_p$ to be the Fourier transform of ψ with respect to the transverse spatial coordinates (x, y), as shown in Fig. 1.3. With the Fourier transform we have

$$\frac{d^2\tilde{\psi}_p}{dz^2}+k_0^2\left(1-\frac{k_x^2}{k_0^2}-\frac{k_y^2}{k_0^2}\right)\tilde{\psi}_p=0 \tag{1.17}$$

, which is a second-order, homogeneous, linear ordinary differential equation with constant coefficients. The solution can be written as

$$\tilde{\psi}_p(k_x,k_y,z)=\tilde{\psi}_p(k_x,k_y,0)\exp\left[-ik_0 z\sqrt{1-\frac{k_x^2}{k_0^2}-\frac{k_y^2}{k_0^2}}\right] \tag{1.18}$$

The spatial frequency transfer function [36] along z direction can be defined as

$$\tilde{G}(k_x,k_y,z)=\frac{\tilde{\psi}_p(k_x,k_y,z)}{\tilde{\psi}_p(k_x,k_y,0)}=\exp\left[-ik_0 z\sqrt{1-\frac{k_x^2}{k_0^2}-\frac{k_y^2}{k_0^2}}\right] \tag{1.19}$$

1.2.3 Fresnel diffraction and Fraunhofer diffraction

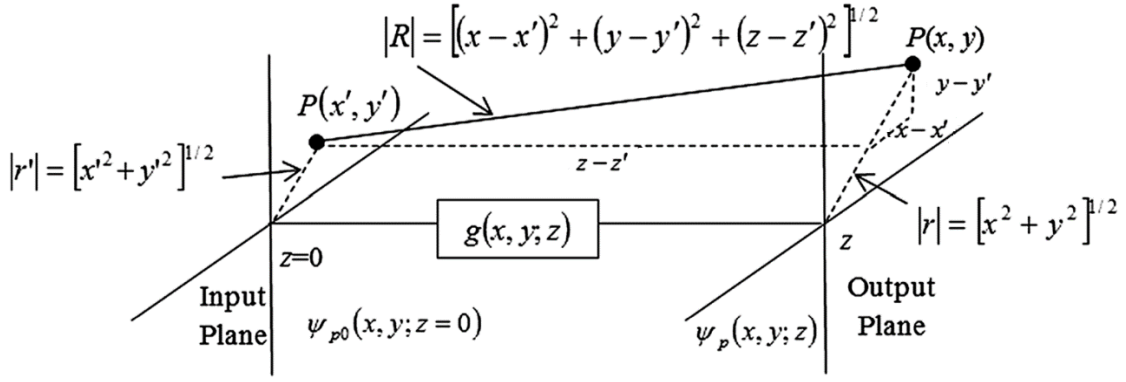


Fig. 1.3 Schematic showing the impulse response of propagation between an input and output plane.

When propagating waves make small angles, i.e. under the so-called paraxial approximation, we have $k_x^2 + k_y^2 \ll k_0^2$ and then

$$\begin{aligned}\tilde{G}\left(k_x, k_y, z\right) &= \exp\left[-ik_0 z\sqrt{1-\frac{k_x^2}{k_0^2}-\frac{k_y^2}{k_0^2}}\right] \\ &\approx \exp\left[-ik_0 z\left(1-\frac{k_x^2+k_y^2}{2k_0^2}\right)\right] = \exp\left(-ik_0 z\right)\exp\left[iz\frac{\left(k_x^2+k_y^2\right)}{2k_0}\right]\end{aligned} \tag{1.20}$$

The impulse response of propagation $g(x, y, z)$ can be obtained by performing the inverse Fourier transform of $\tilde{G}$,

$$\begin{aligned}g\left(x, y, z\right) &= F^{-1}\left[\tilde{G}\left(k_x, k_y, z\right)\right] \\ &= \frac{ik_0}{2\pi z}\exp\left(-ik_0 z\right)\exp\left[-ik_0\left(\frac{x^2+y^2}{2z}\right)\right]\end{aligned} \tag{1.21}$$

Perform inverse Fourier transform to revert to space domain,

$$\begin{aligned}\psi_p\left(x, y, z\right) &= F^{-1}\left[\tilde{\psi}_p\left(k_x, k_y, z\right)\right] = F^{-1}\left[\tilde{\psi}_p\left(k_x, k_y, 0\right)\tilde{G}\left(k_x, k_y, z\right)\right] \\ &= \psi_p\left(x, y, 0\right) * g\left(x, y, z\right)\end{aligned} \tag{1.22}$$

then,

$$\psi_p\left(x, y, z\right) = \exp\left(-ik_0 z\right)\frac{ik_0}{2\pi z}\iint\psi_p\left(x', y', 0\right)\exp\left\{-\frac{ik_0}{2z}\left[\left(x-x'\right)^2+\left(y-y'\right)^2\right]\right\}dx'dy' \tag{1.23}$$

This is called the Fresnel diffraction formula [36]. The range of applicability of the Fresnel diffraction formula spans the near field of the object to the far field.

If only the far-field diffraction pattern is of interest, another approximation can be made to yield the Fraunhofer diffraction formula [36].

$$\begin{aligned}\psi_p(x,y,z) &= \exp(-ik_0 z)\frac{ik_0}{2\pi z}\exp\left[-\frac{ik_0}{2z}\left(x^2+y^2\right)\right] \\ &\times\iint \psi_p(x',y',0)\exp\left\{-\frac{ik_0}{2z}\left[\left(x'^2+y'^2\right)+\left(2xx'+2yy'\right)\right]\right\}dx'dy' \\ &= \exp(-ik_0 z)\frac{ik_0}{2\pi z}\exp\left[-\frac{ik_0}{2z}\left(x^2+y^2\right)\right] \\ &\times\iint \psi_p(x',y',0)\exp\left\{-\frac{ik_0}{2z}\left(x'^2+y'^2\right)\right\}\exp\left[\frac{ik_0}{z}\left(xx'+yy'\right)\right]dx'dy'\end{aligned} \tag{1.24}$$

For Fraunhofer approximation, the term $\pi\left(x'^2+y'^2\right)$ is like the maximum area of the source and if this area divided by the wavelength is much less than the distance z under consideration, the term $\exp\left\{-\frac{ik_0}{2z}\left(x'^2+y'^2\right)\right\}$ inside the integrand can be considered as unity, so

$$\begin{aligned}\psi_p(x,y,z) &= \exp(-ik_0 z)\frac{ik_0}{2\pi z}\exp\left[-\frac{ik_0}{2z}\left(x^2+y^2\right)\right]\iint \psi_p(x',y',0)\exp\left[\frac{ik_0}{z}\left(xx'+yy'\right)\right]dx'dy' \\ &= \exp(-jk_0 z)\frac{ik_0}{2\pi z}\exp\left[-\frac{ik_0}{2z}\left(x^2+y^2\right)\right]F\left\{\psi_p(x,y,0)\right\}dxdy\end{aligned} \tag{1.25}$$

Fresnel diffraction formula and Fraunhofer diffraction formula can also be obtained based on Rayleigh-Sommerfeld diffraction model of the well-known Huygens-Fresnel principle. From the standpoint of object wavefront reconstruction, Fresnel holograms differ from Fourier holograms in that they have focusing properties and are capable of reproducing the finite distance from the observation surface to the object.

1.3 Conventional algorithms for CGHs design

According to the optical propagation model from the CGH plane to reconstruction plane, the algorithms for CGHs design can usually be divided into two groups: unidirectional algorithm

and bidirectional algorithm. For unidirectional algorithm, the optical propagation will only be calculated in the forward direction. Typical examples of unidirectional algorithms are Direct Binary Search (DBS) [37, 38], Simulated Annealing (SA) [14, 39, 40] and Genetic algorithm (GA) [41, 42]. While for the bidirectional algorithm, the inverse propagation from the image plane to CGH plane will also be calculated. the classic Iterative Fourier Transform Algorithm (IFTA) [43] and IFTA-based phase retrieval algorithm [44, 45] belong to this group.

1.3.1 Direct Binary Search

The basic idea of DBS is to begin with a random CGH with binary phases and flip the value in a pixel-by-pixel order till all the pixels are scanned. After each change of the value encoded into a pixel, the impact of such change on the image reconstructed is evaluated. If the image is improved, then the flipping is accepted, otherwise rejected. Such process is repeated until no more single pixel changes to produce a better image within given iteration number.

This can ensure the DBS results in a solution with a simple implementation way, however, at the expense of a huge number of flip trials. The main drawback of DBS is that it often converges to a local optimum instead of a globe one.

1.3.2 Simulated Annealing

Simulated Annealing is a probabilistic technique for approximating the global optimum of a given function in a large search space. The name and inspiration come from annealing in metallurgy – a technique involving heating and controlled cooling of a materials to increase the size of its crystals and reduce their defects. One of implementation of SA is quite similar with DBS. As before, changes that lead to a better image quality will be accepted unconditionally. However, it will also probabilistically accept the value inversions of hologram pixels that increase the error of image reconstruction. The probability of accepting these inversions is relatively large, and as the iteration goes, the probability decreases. In this way, SA can avoid stopping when a local optimum is reached, as that in DBS. It will usually give a better result than DBS, but with a slower convergence.

1.3.3 Genetic algorithm

Genetic algorithm is a method based on natural selection, the process that drives biological evolution. It repeatedly modifies a population of individual solutions randomly generated by selecting individuals from the current population to be parents and uses them to produce the children for the next generation via crossover. Mutation is then applied some random change to maintain the genetic diversity of the genes in the children. Over successive generations, the population evolves toward an optimal solution. The selection, crossover and mutation are three main operations in classic GA. For each operation, there are many different ways to implement. Sometimes, elitism strategy will also be introduced, which means the best individual in a population is transmitted to the next generation without crossover and mutation. GA can be also used together with other conventional algorithms, such as direct search, gradient search etc., which was named as hybrid genetic algorithm (HGA). In some cases, HGA can converge much faster than classic GA.

Compared with DBS and SA, which are essentially serial optimization algorithms, parallel genetic algorithm (PGA) can be easily developed to take advantage of modern computer technology.

1.3.4 Iterative Fourier Transform Algorithm

The IFTA was first proposed by Hirsch et al. Gerchberg and Saxton independently dealt with the phase-retrieval problem using a similar algorithm, and thus the IFTA is also called the Gerchberg-Saxton (G-S) algorithm. Fig. 1.4 shows a general flowchart of the IFTA for generating a Fourier hologram. The algorithm consists of the following four simple steps: (1) Fourier transform an estimate of the object; (2) replace the modulus of the resulting computed Fourier transform with the measured Fourier modulus to form an estimate of the Fourier transform; (3) inverse Fourier transform the estimate of the Fourier transform; and (4) replace the modulus of the resulting computed image with the measured object modulus to form a new estimate of the object. The generalized Gerchberg-Saxton algorithm can be used for any problem in which partial constraints (in the form of measured data or information known a priori) are known in each of two domains, usually the object (or image) and Fourier domains.

One simply transforms back and forth between the two domains, satisfying the constraints in one before returning to the other.

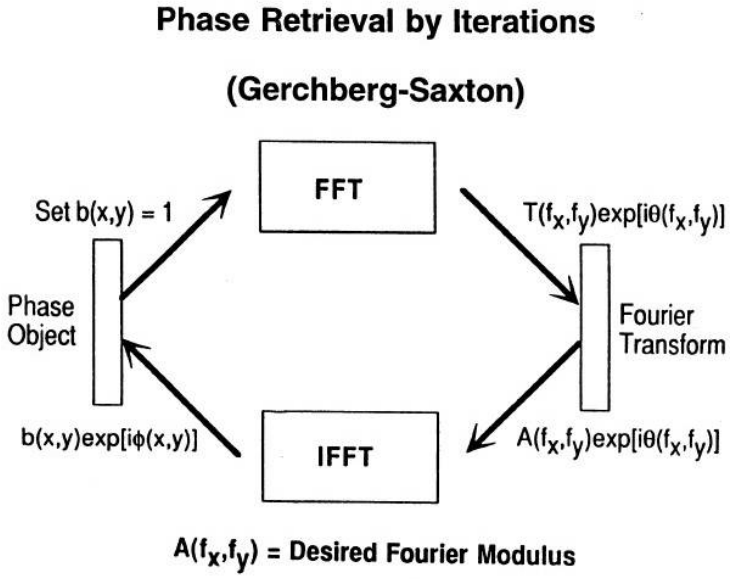


Fig. 1.4 A flowchart of phase retrieval by iterations for generating a Fourier hologram

1.4 Industrial schemes for CGH design

For many industrial applications, holograms have relatively large sizes, i.e. space bandwidth product (SBWP). Meanwhile, the low fabrication cost of hologram is very important for mass manufacture. Therefore, those conventional optimization algorithms listed above are usually inappropriate for design of industrial holograms. Herein, as two widely used industrial schemes, the dot matrix technology (dot matrix hologram) [46-53] and holographic printer [54-64] are introduced.

1.4.1 Dot matrix hologram

Dot matrix hologram [46-53], as shown in Fig. 1.5, is a type of hologram that is composed of thousands of fine diffraction grating dots with different grating constants and orientation, which can be controlled by computer according to the pattern design. These diffraction grating dots can be exposed onto a photoresist plate forming a relief hologram that is used for mass production with embossing technique. Controlling the angle, exposure, size, shape,

and spacing of every grating in the hologram, allows the end-user a wide range of visual effects. When illuminated with white light the gratings split the light into a spectrum of colors and redirect the light at various angles to form a kinetic hologram image. Because of large visual angle, high diffraction efficiency and kinetic visual effect of dot matrix holograms, they are widely used in security printing and anti-counterfeiting.

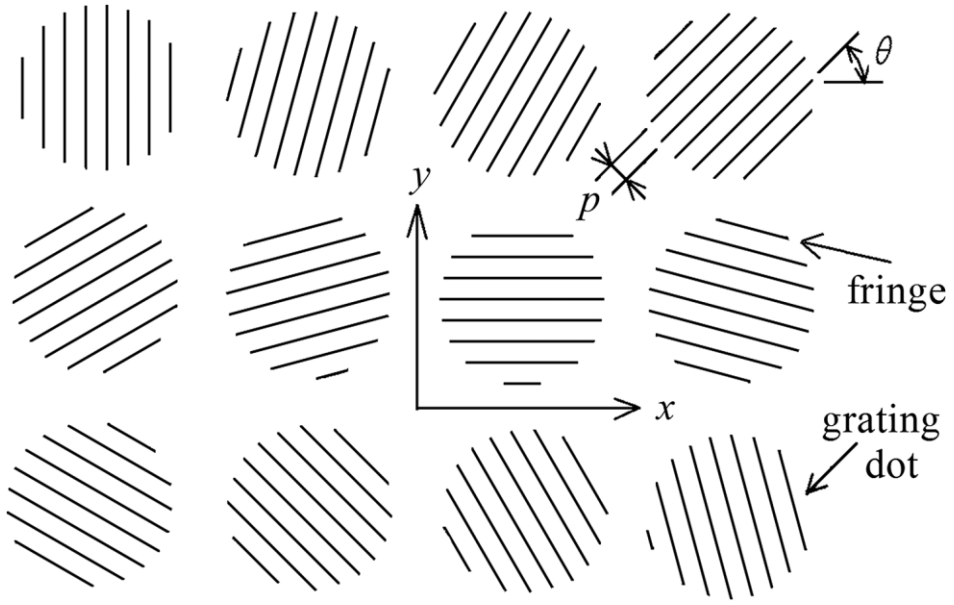


Fig. 1.5 A dot matrix hologram consisting of fine diffraction grating dots with different grating constants and orientation

Electron beam writers, pattern copiers, and two-beam writers can be used to create dot-matrix holograms. For brightness of dot-matrix holograms, holograms created by e-beam writers are brightest and those created by pattern copiers are darkest. For mastering speeds, pattern copiers are fastest and e-beam writers are slowest. For machine prices and running costs, e-beam writers are very expensive, but two-beam writers and pattern copiers are cheap. Because the performance of two-beam writers is always at the balance position for all the above-mentioned features, two-beam writers are more popular than e-beam writers and pattern copiers.

1.4.2 Holographic printer

Holographic printing technology has been largely reported recently [54-64]. A common feature of all holographic printers is division of the printed hologram into a 2D array of

holographic elements (hogels). Information to be recorded in a hogel is displayed on a SLM. Recording of the whole hologram is done by successive exposure of all hogels using a motorized X-Y translation stage, as shown in Fig. 1.6.

Holographic printers can be basically classified into two categories. One is “holographic stereogram printer”. It can be only used for the holographic stereograms, which are synthesized from sequences of closely spaced two-dimensional (2D) perspective views and not faithfully reconstructs wavefront of the recorded object. These series of 2D perspective images are stored in hogels. With an appropriate illumination, each hogel in the holographic stereogram diffracts a fraction of perspective images within a certain viewing angle. The diffracted lights are fractions of the perspective image recorded on the hogels in the holographic stereogram. On the observing plane, relatively large numbers of viewing points are formed by the merged lights diffracted from the hogels. An observer can see a 3D image from these perspective views at the corresponding viewing point. When the observer’s left and right eyes are located at different viewing points, the observer sees stereoscopic images and perceives 3D images.

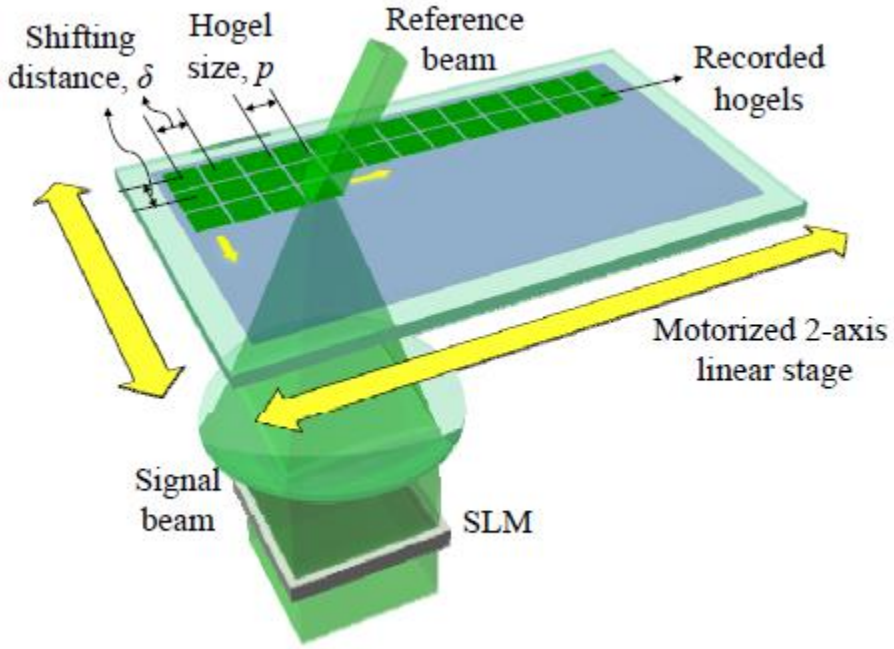


Fig. 1.6 Schematic diagram of a holographic printer

The other one is “holographic printer for CGH”, which can print the hologram that reconstructs wavefront from any kind of CGH and also the computer-generated holographic stereogram. The holographic fringe printer transfers the input CGH into a holographic emulsion forming a thin hologram and thus this printing method lacks color selectivity. The holographic volume printer decodes the 3D object wavefront from the CGH and records it as an analogue volume hologram. Since its wavelength selectivity, it is suitable for while-light reconstruction hologram, particularly for a full-color hologram.

1.5 Fabrication technologies of holograms

1.5.1 Diamond machining

One of the first techniques for fabrication of holograms is diamond machining [65-67], which can generate diffractive microstructures directly through mechanical removal of optical material. It is a process of mechanical machining of precision elements using lathes or derivative machine tools (e.g., turn-mills, rotary transfers) equipped with natural or synthetic diamond-tipped tool bits, as shown in Fig. 1.7.

Diamond machining is a multi-stage process. Initial stages of machining are carried out using a series of computer numerical control (CNC) lathes of increasing accuracy. A diamond-tipped lathe tool is used in the final stages of the manufacturing process to achieve sub-nanometer level surface finishes and sub-micrometer form accuracies. The surface finish quality is measured as the peak-to-valley distance of the grooves left by the lathe.

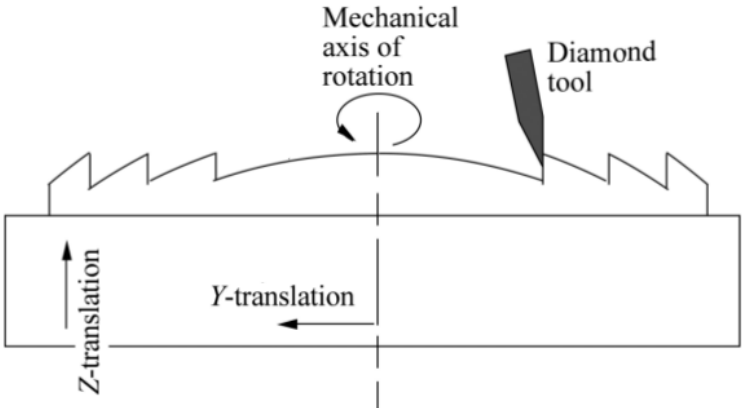


Fig. 1.7 Diamond machining of a micro-lens.

1.5.2 Photolithography

Photolithography, also termed optical lithography or UV lithography, is a process used in microfabrication to pattern parts of a thin film or the bulk of a substrate [68, 69]. It uses light to transfer a geometric pattern from a photomask to a light-sensitive chemical "photoresist" on the substrate. A series of chemical treatments then either engraves the exposure pattern into or enables deposition of a new material in the desired pattern upon the material underneath the photo resist. The process is illustrated in Fig. 1.8.

Firstly, a photoresist layer is deposited on a substrate by spin coating. The thickness of the photoresist layer depends on the viscosity and the spin coating speed. Secondly, a binary mask of alternating transparent and opaque areas is fabricated using some type of pattern generator. The mask is laid on a substrate coated with a thin layer of photoresist, which is exposed to ultraviolet light through the mask. After the resist is developed, a pattern is created in the photoresist layer. The substrate is then etched until the required depth is reached. The photoresist pattern is then removed, resulting in a binary element.

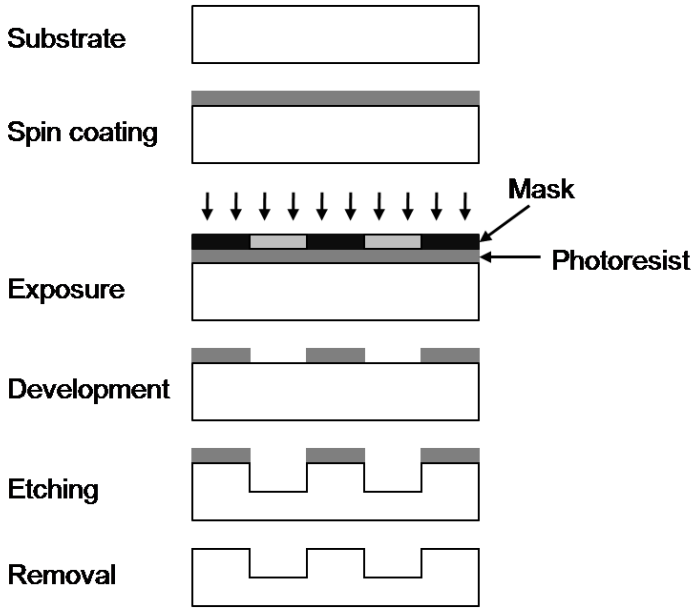


Fig. 1.8 Experimental steps of photolithography: spin coating, exposure, development, etching and removal.

1.5.3 Direct laser writing

Direct laser writing (DLW), also known as Multiphoton lithography, is a three-dimensional (3D) printing technology which allows the construction of readily assembled structures with sub-100 nm resolution [70-72]. Similar to standard photolithography techniques, structuring is accomplished by illuminating photoresists via light of a well-defined wavelength. The fundamental difference, however, is that this method relies on nonlinear photon absorption by photopolymers. The beam of an ultra-fast laser is tightly focused inside the volume of a transparent material, causing it to absorb two or more photons and polymerize locally. Moving the beam according to a path representing a Computer Aided Design (CAD) model, one can fabricate a realistic micromodel of this design.

The experimental procedure for fabricating a 3D structure by DLW is shown in Fig. 1.9. (I) The laser beam is tightly focused into the volume of the material. (II) Either the focused beam or the sample move following a computer-generated pattern. (III) After the laser writing of the structure, the sample is immersed into an appropriate developer. (IV) The freestanding structure is revealed.

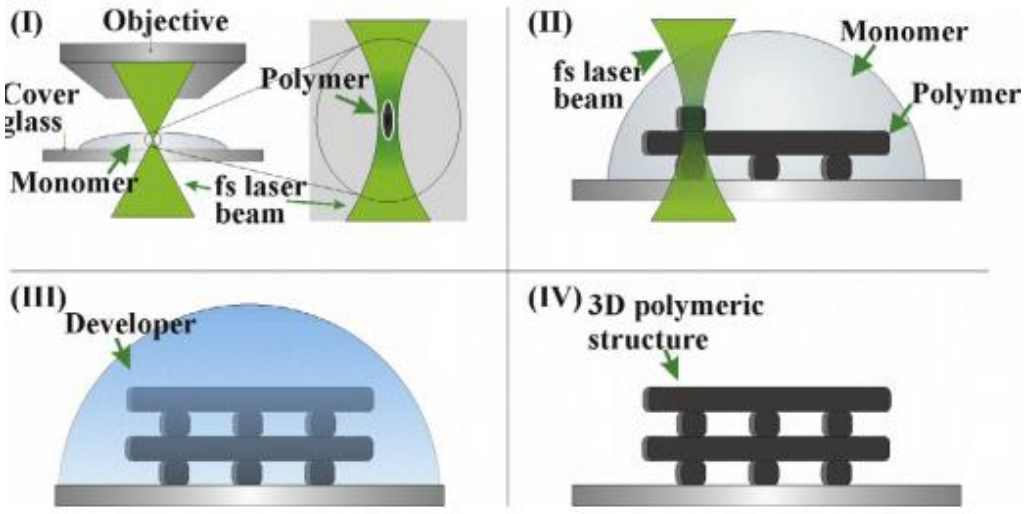


Fig. 1.9 Experimental procedure: (I) beam focusing, (II) laser writing, (III) development, (IV) completed structure.

1.5.4 Electron beam lithography

Electron beam lithography (EBL) is a powerful technique [73, 74] for creating nanostructures that are too small to fabricate with conventional photolithography. State of the art EBL

systems can achieve resolutions of a few nanometres. The technique works by moving a highly focussed electron beam over a sample to write out a pattern designed with suitable CAD tools.

The pattern is recorded in an electron sensitive film (or resist) deposited on the sample before exposure by spin coating. The electron beam induces a change in the molecular structure and solubility of the resist film. Following exposure to the electron beam, the resist is developed in a suitable solvent to selectively dissolve either the exposed or unexposed areas of the resist. After exposing and developing, the resist layer on top of the sample can be used as a mask or template for transferring the pattern into a more useful medium, as shown in Fig. 1.10.

The advantage of e-beam lithography stems from the shorter wavelength of accelerated electrons compared to the wavelength of ultraviolet (UV) light used in photolithography, while the main drawbacks are the high cost and the slow exposure process of EBL system, resulting in a long writing time (several hours) for relatively small areas (usually a few mm^2).

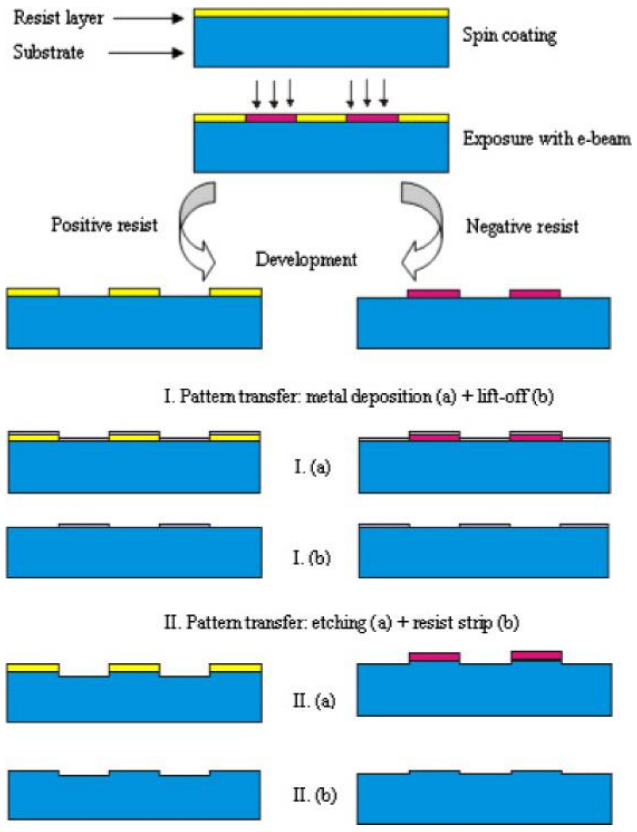


Fig. 1.10 Electron beam lithography

Chapter 2 Design of holograms by direct binary search for structured-light projection system

2.1 Introduction

Three-dimensional (3D) surface imaging [75], as one of the fundamental topics in computer vision, have made tremendous progresses in research, development and commercialization in the recent decades because of its great application demands in a variety of market segments. Some typical examples include industrial inspection, cultural heritage, dental health care or object recognition. One of the most widely used techniques is based on the projection of structured light [76, 77], which is active illumination of the scene with specially designed 2D spatially varying intensity pattern.

As illustrated in Fig. 2.1, a spatially varying 2D structured illumination is generated by a projector. The intensity of each pixel on the structured-light pattern is represented by the digital signal. An imaging sensor, for instance, a camera is used to acquire a 2D image of the scene under the structured-light illumination. If the scene is a planar surface without any 2D surface variation, the pattern shown in the acquired image is similar to that of the projected structured-light pattern. However, when the surface in the scene is nonplanar, the geometric shape of the surface distorts the projected structured-light pattern, as seen from the camera. The principle of structured light 3D surface imaging techniques is to extract the 3D surface shape based on the information from the distortion of the projected structured pattern. As shown in Fig. 2.1, the geometric relationship between an imaging sensor, a structured-light projector and an object surface point can be expressed by the triangulation principle [75]. There are many kinds of structured-light patterns [75, 78] for the surface imaging technique, among which, the binary pattern [79-81] is one of the reliable and simple structured patterns.

In a variety of projection imaging system [82-84], computer-generated holograms (CGHs) are widely used. Such devices can be realized as static structures etched, which are sometimes called diffractive optical elements (DOEs), or displayed on dynamically micro-display devices such as liquid crystal on silicon (LCOS). In both cases, the CGHs can perform the

entire functionality associated with optical assembly containing many components like lens, in the imaging system, leading to more compact, convenient, reliable and low-cost scheme. The diffraction property of CGH potentially allows for projection angles much larger than the conventional projection system which are limited by the necessity for a relatively large projection lens. Although being able to enlarge image, the lens assembly may bring severe aberrations, which can only be handled by the corporation of highly complex and expensive lens system. The conventional lens assembly is not a good choice for projection, particularly in the miniaturized projector [84].

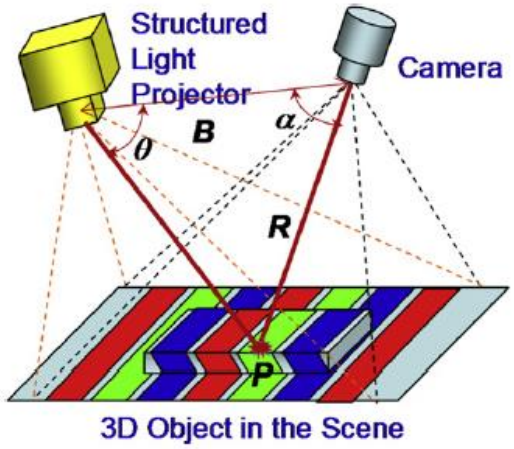


Fig. 2.1 Illustration of structured light.

Moreover, compared with the amplitude modulating (AM) technique used in conventional projector, which selectively block incident light to form the desired image, a holographic projector employing a phase modulating (PM) CGH has a transmission of unity, in which the improvement of diffraction efficiency can be expected. A phase-only holographic projector is also able to exert control over the imaging system with a wide projection angle, so that the project system can be built without residual optical aberrations. Besides, extended depth-of-field in holographic projectors is reported as an advantage as well compared with that in conventional projectors [86]. Basically, the pixel size of CGH employed in a holographic projection system should be as small as possible, so that subsequent lens power in the system to achieve the desired projection angle is minimized [84]. Nevertheless, due to the resolution of lithography usually used, design of CGH, fabrication of DOE and construction of a practical projection system with certain specifications need to be comprehensively considered.

In this chapter, we will design a binary-phase hologram for a structured-light projector with certain specifications, for example, the typical work distance can be ~40 cm, and the depth of field is ±10 cm; the laser wavelength is 632 nm and the pattern angle is around 53 degree. The direct binary search is adopted to optimize the binary CGH first based on the Fourier imaging system. Furthermore, the Fresnel hologram with a reasonable depth of focus (DoF) is synthesized by the DBS for Fresnel imaging system. The projection system with a ~ 53 degree field of view (FoV) is discussed as well. The fabrication process of binary hologram and its diffraction patterns are presented.

2.2 Fourier hologram and efficient DBS

The synthesis problem for CGHs amounts to choosing a binary transmittance for the addressable cells that performs the desired wavefront reconstruction. In this part, we mainly focused on the implementation of an efficient DBS algorithm [37, 38] by considering an optical configuration in which propagation between the hologram and the observation plane is described by a 2D Fourier transform.

2.2.1 Fourier hologram

Herein, the hologram is placed at the front focal plane of a thin positive lens and illuminated with a plane monochromatic wave. The resulting wavefront at the back focal plane of the lens is the 2D Fourier transform of the spatial distribution of the light just behind the hologram. If we let *H(u, v)* represent the transmittance of the hologram, and *h(x, y)* represent the complex amplitude of the wavefront at the back focal plane, then with an appropriate definition of coordinates, we can write

$$h(x,y)=\iint H(\mu,\nu)\exp[-i2\pi(\mu x+\nu y)]dxdy \tag{2.1}$$

The CGH is defined as a MxN array of rectangular cells of dimension R by S with binary transmittance H_{kl}, -M/2<k<M/2 and –N/2<l<N/2 (M and N are even). Therefore,

$$H(\mu,\nu)=\sum_{k=-M/2}^{M/2-1}\sum_{l=-N/2}^{N/2-1}H_{kl}\,rect\left(\frac{\mu-kR}{R},\frac{\nu-lS}{S}\right) \tag{2.2}$$

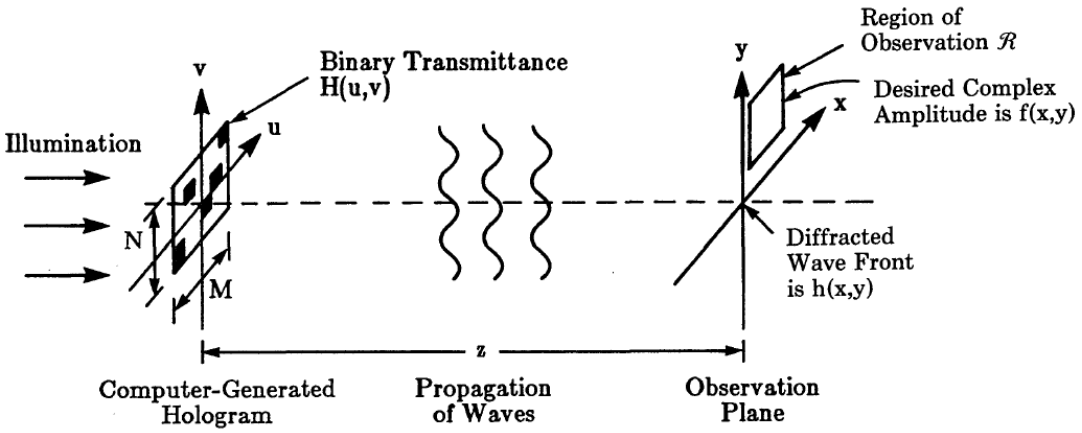


Fig. 2.2 Schematic of wavefront reconstruction by computer-generated hologram

where

$$rect(a,b)=\begin{cases}1, if\ |a|,|b|<1/2\\ 0, else\end{cases} \tag{2.3}$$

The observed complex amplitude *h(x,y)* is then given by

$$h(x,y)=RS\sin c(Rx,Sy)\sum_{k=-M/2}^{M/2-1}\sum_{l=-N/2}^{N/2-1}H_{kl}\exp\left[i2\pi(Rxk+Syl)\right] \tag{2.4}$$

where $\sin c(a,b)=\sin(\pi a)\sin(\pi b)/(\pi^2 ab)$. Sampling this diffraction pattern at the points (x, y) = (mX, nY) corresponding to the Nyquist interval x=1/MR and Y=1/NS yields

$$h(mX,nY)=\left(\frac{RS}{XY}\right)^{1/2}\sin c\left(\frac{m}{M},\frac{n}{N}\right)h_{mn} \tag{2.5}$$

where h_{mn} is the inverse discrete Fourier transform (DFT) of H_{kl},

$$h_{mn}=\frac{1}{\sqrt{MN}}\sum_{k=-M/2}^{M/2-1}\sum_{l=-N/2}^{N/2-1}H_{kl}\exp\left[i2\pi\left(\frac{mk}{M}+\frac{nl}{N}\right)\right]=iDFT(H_{kl}) \tag{2.6}$$

2.2.2 DBS and its efficient implementation

Since the above equation is a discrete summation, a trial (*i*th + 1) reconstruction of h_{mn} can be expressed as a recursive relation consisting of the present (*i*th) reconstruction plus the

amount of change. At each trial inversion, we alter only a single cell $H_{k'l'}$ in the binary CGH. Rather than performing an MxN point fast Fourier transform (FFT) in order to compute the new reconstruction *h'*, we express it in terms of the reconstruction h before the trial inversion:

$$h'_{mn} = h_{mn} \pm \frac{1}{\sqrt{MN}} \exp\left[i2\pi \left(\frac{mk'}{M} + \frac{nl'}{N} \right) \right] \tag{2.7}$$

If the trail inversion at (*k', l'*), e.g. $H_{k'l'}$=1, take "+", if $H_{k'l'}$=-1, take "-". This expression need be evaluated only at the AxB sample points within R, whereas the MxN point DFT would yield the new reconstruction at all MN points in the observation plane.

Assume the target image have AxB addressable points, define the mean squared error *e* between the object and reconstructed image as

$$e = \frac{1}{AB} \| f - \lambda h \|^2 = \frac{1}{AB} \sum_{m=-A/2}^{A/2-1} \sum_{n=-B/2}^{B/2-1} | f_{mn} - \lambda h_{mn} |^2 \tag{2.8}$$

where f_{mn} be the object scaled to have a peak spectral amplitude of unity, the parameter λ is a complex factor that scales the reconstruction for a minimum mean squared error fit to the original object, expressed as

$$\lambda = \frac{\langle f, h \rangle}{\|h\|^2} = \frac{\sum_{m=-A/2}^{A/2-1} \sum_{n=-B/2}^{B/2-1} f_{mn} h^*_{mn}}{\sum_{m=-A/2}^{A/2-1} \sum_{n=-B/2}^{B/2-1} |h_{mn}|^2} \tag{2.9}$$

The * is the complex conjugate operation.

The DBS algorithm is a search -oriented technique that seeks to directly minimize the mean -squared error between the reconstructed and desired wavefronts. The DBS algorithm is initialized by generating a random binary hologram. The reconstructed image g_{mn} is calculated via the inverse FFT. The hologram is then scanned in lexicographic order. For each hologram point, we invert the binary value of its transmittance and compute the resultant reconstructed image. The error between the new reconstruction and the object is then calculated and compared to the previous error. If the error decreases, the altered hologram configuration and the new error value are retained. Otherwise, the hologram point is restored to its original value. Once every addressable point in the hologram has been considered, an

iteration is said to be completed. The algorithm terminates when no inversions are kept during an entire iteration.

The DBS algorithm is demanding computationally for computing 2 x (M x N) FFTs in one round of scan, where (M, N) is the size of the hologram. The computational requirement for the DBS may be substantially reduced by recursively computing h after each trial inversion, and it may be reduced still further by recursively computing the error. However, even for moderate space-bandwidth products, this method is still too computationally intensive for a DBS to be practical. Rather than updating the reconstruction after each inversion and using this for re-computing the mean-squared error, it is possible to evaluate the effect of a trial inversion on the mean-squared error without directly re-computing the reconstruction. Substitute Eq. (2.9) to Eq. (2.8), we have

$$e = \frac{1}{AB}\left(\|f\|^2 - \frac{\left|\langle f, h\rangle\right|^2}{\|h\|^2} \right) \tag{2.10}$$

Since the first term is a constant for a given object f, we only compute the remaining terms in order to determine the new mean-squared error resulting from a trial inversion. Assume h is present reconstruction and h' is the trial reconstruction with the inversion of $H_{k'l'}$,

$$\langle f, h'\rangle = \langle f, h\rangle + \frac{a'}{\sqrt{MN}} \sum_{m=-A/2}^{A/2-1} \sum_{n=-B/2}^{B/2-1} f_{mn} \exp\left[-i2\pi \left(\frac{mk'}{M} + \frac{nl'}{N} \right) \right] = \langle f, h\rangle + a' F_{k'l'} \tag{2.11}$$

$F_{k'l'}$ can be computed and stored, the $\langle f, h \rangle$ may be updated after each trial inversion with a single addition.

For the term $\|h\|^2$ due to the trial inversion of $H_{k'l'}$, we have

$$\begin{aligned} \|h'\|^2 &= \|h\|^2 + \frac{AB}{MN} + 2a' \mathrm{Re}\left\{ \frac{1}{\sqrt{MN}} \sum_{m=-A/2}^{A/2-1} \sum_{n=-B/2}^{B/2-1} h_{mn} \exp\left[-i2\pi \left(\frac{mk'}{M} + \frac{nl'}{N} \right) \right] \right\} \\ &= \|h\|^2 + \frac{AB}{MN} + 2a' \mathrm{Re}\left(\tilde{H}_{k'l'} \right) \end{aligned} \tag{2.12}$$

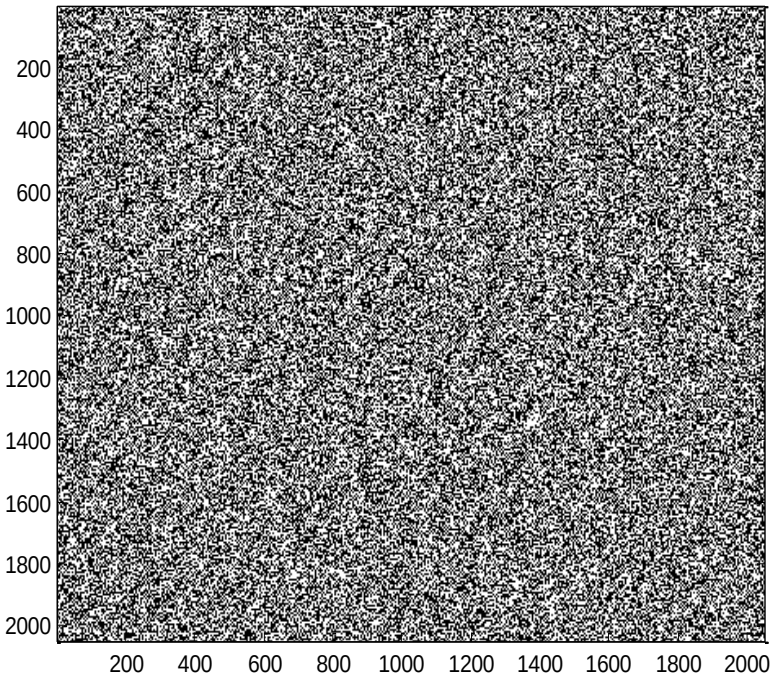
In contrast to $F_{k'l'}$, H_{kl} will change each time an inversion is accepted. For the acceleration of computation, we update H_{kl} by using the FFT at less frequent intervals, which is an accepted approximation.

2.3 Binary holograms by DBS

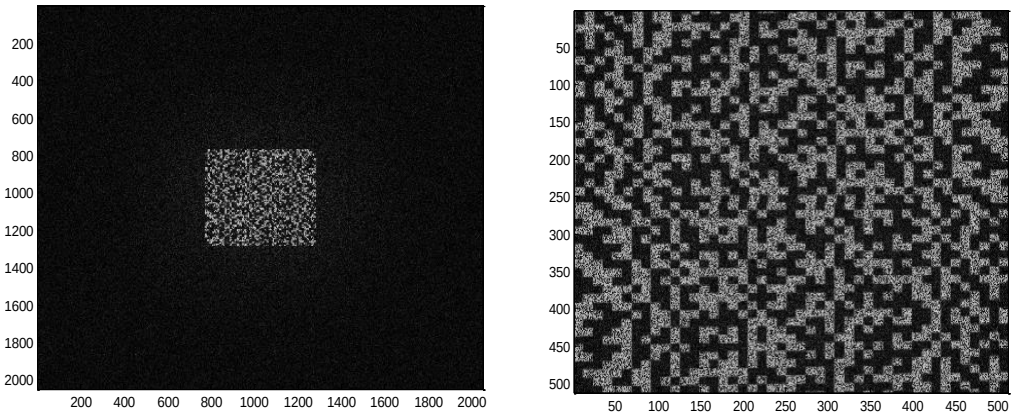
We first computed a part (512 x 512 pixels) of the target pattern with the fast DBS based on Fourier holography. For providing enough degree of freedom, the 512 x 515 image is zero-padded to a window of 2048 x 2048 pixels. The computation of errors is restricted to the 512 x 512 region in which the object is located, usually, the center of the window. The synthesized binary hologram is of 2048 x 2048 pixels shown in Fig. 2.3 (a). Fig. 2.3 (b) shows its reconstructed image, where the target pattern is embedded in a 2048 x 2048 pixels matrix. Fig. 2.3 (c) shows the reconstructed pattern of 512 x 512 pixels. We evaluated the contrast of the reconstructed pattern by

$$C = (\bar{I}_1 - \bar{I}_0)/(\bar{I}_1 + \bar{I}_0) \qquad (2.13)$$

Where $\bar{I}_1$ represents the average intensity in the dark regions of the target pattern and $\bar{I}_0$ represents the average intensity in the bright regions of the target pattern. The contrast was 87%, with the average intensity of 1.12 in the dark regions and of 16.3 in the bright regions of the target pattern.



(a)



(b) (c)

Fig. 2.3 (a) binary hologram designed, (b) reconstructed image by hologram shown in (a), (c) reconstructed target pattern

Unlike reconstruction of image on the focal plane for the case of Fourier holography, for Fresnel holography, the wavefront is specified at some distance behind the CGH, referred to here as the plane of observation, and it must be back propagated numerically to the plane of the CGH. With Fresnel approximations mentioned in Chapter 1, the back propagation may be computed by multiplying the desired wavefront by a quadratic phase function, calculating the discrete Fourier transform and then multiplying the result by an additional quadratic phase function. The fast DBS was applied to synthesize the Fresnel holograms as well. The hologram is of 512 x 512 pixels with pixel size of 2 μm, wavelength λ = 632.8 μm and the propagation distance from the hologram to image was z = 40 cm. The synthesized hologram is shown in Fig. 2.4 (a), and its entire reconstructed image at a distance of 40 cm is presented in Fig. 2.4 (b). The reconstructed target images from Fresnel hologram are also shown at a distance of 30 cm, 40 cm, 50 cm and 1 km, such as in Fig. 2.5. In fact, the size of the hologram is very small compared with the working distance of 400 mm. Although the designed projector has a high field of view of 53°, the numerical aperture of the hologram is very small. The size of hologram contains only a slightly larger area than one Fresnel zone. This is in fact the plane wave approximately. Thus, the Fourier hologram can be designed in the projector, and the depth of focus will be good when adjust power and position of lens used behind hologram.

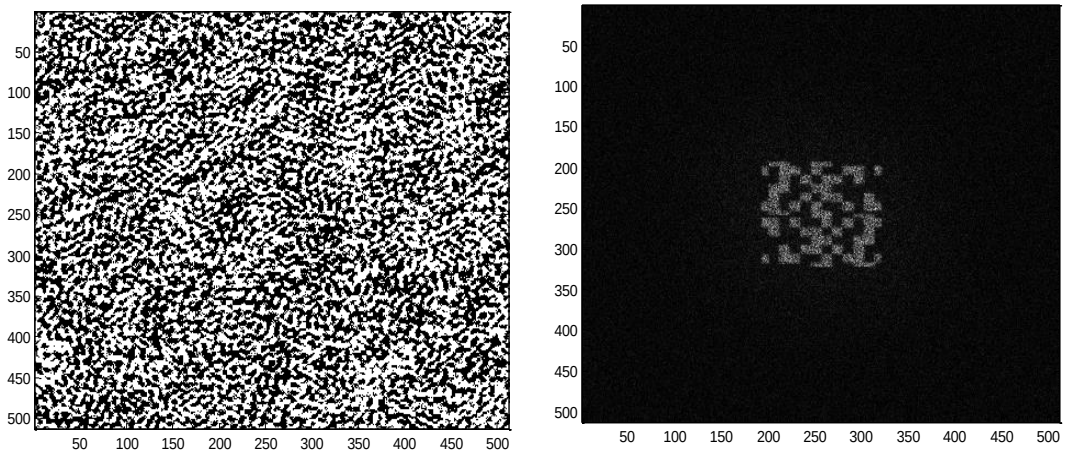


Fig. 2.4 (a) Binary Fresnel hologram of 512 x 512 pixels, (b) Reconstructed image from the binary hologram shown in (a)

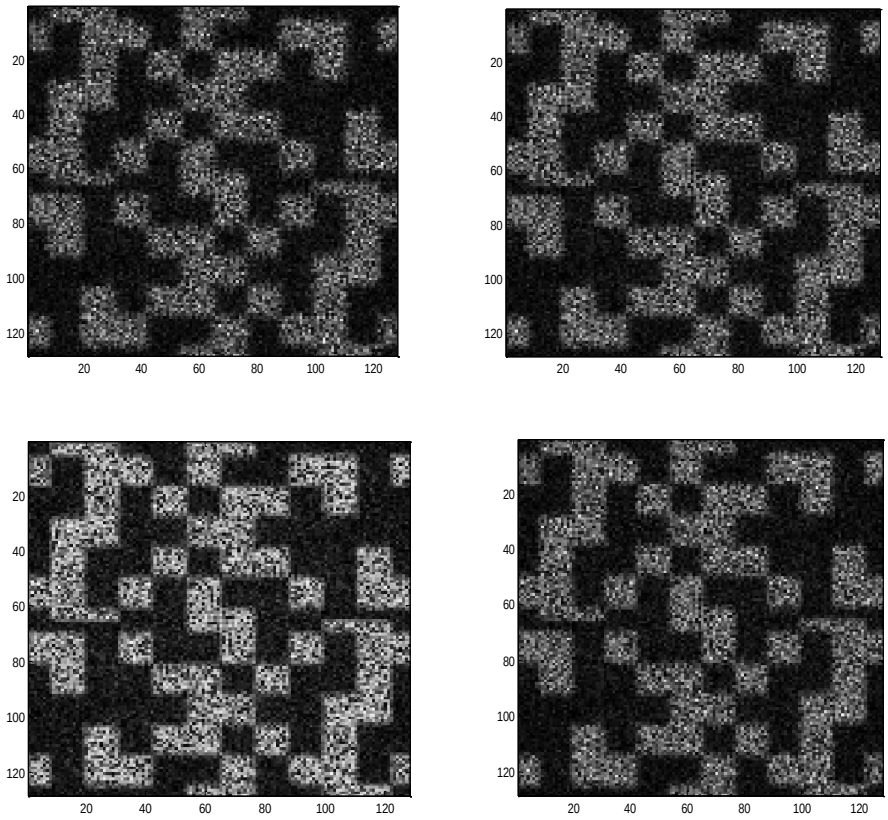


Fig. 2.5 Reconstructed pattern from Fresnel hologram shown in Fig. 2.4 (a) at distance from the hologram (clockwise) : 30 cm, 40 cm, 50 cm and 1 km

2.4 Projection imaging system

The pixel size is usually chosen, in conventional projection system, to represent a compromise between maintaining an adequate aperture ratio whilst minimizing diffractive effects. However, such a restriction does not apply for a projection system which exploits diffraction of CGH. The diffraction angle from a hologram pattern of pixel size Δ placed behind a lens and illuminated with coherent collimated light of wavelength λ, is given [85] by

$$\theta = \arctan\left(\frac{\lambda}{2\Delta}\right) \qquad (2.14)$$

This inverse relationship between diffraction angle and feature size suggests that the pixel size employed in a holographic projection system should be as small as possible, so that subsequent lens power to achieve the desired projection angle is minimized.

For a typical projector, suppose a diffraction pattern of size 40 cm x 40 cm (or 50 cm x 50 cm) is required at a distance of 40 cm (or 50 cm), which indicates the diffraction angle is around 53.1 degree. According to the Eq. (2.14), the pixel size will be 0.237 micron for wavelength of 632 nm. However, the resolution of lithography we will use can only achieve to 2 micron. The optical architecture proposed by Buckley [84] can be adopted to overcome this problem, as shown in Fig. 2.6. The lens pair of L1 and L2 form a telescope, which expands the laser beam to capture the entire hologram pattern. The reverse arrangement is used for the lens pair of L3 and L4, which acts to demagnify the hologram pixels and consequently increase the diffraction angle. The demagnification D is set by the ratio of focal lengths f3 to f4 and, due to the properties of Fraunhofer diffraction, the images remain in focus at all distances from L4. The demagnification D is:

$$D = \frac{f_3}{f_4} = \frac{2}{0.237} \approx 8.4 \qquad (2.15)$$

Therefore, the required demagnification by lens pair L3 and L4 to obtain a diffraction angle of 53.1 ° must be 8.43X.

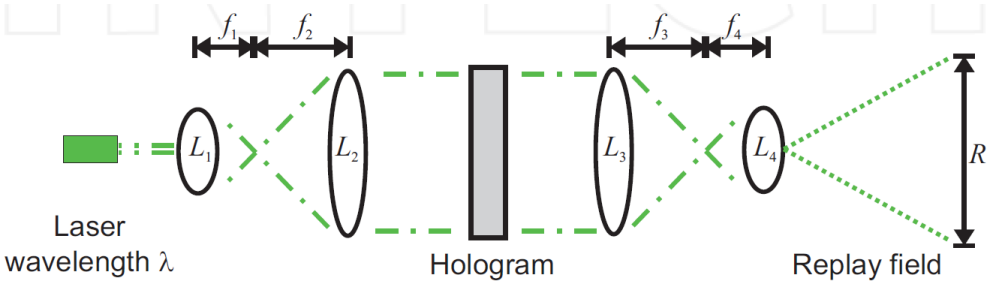


Fig. 2.6 Optical design for a holographic projector, in which beam expansion is performed by lenses L1 and L2, and demagnification by lenses L3 and L4.

2.5 Experimental results

The final desired object window is of 2048x2048 pixels and placed in the centre of a window of 4096x4096 pixels after "zero-padding", which can provide more degrees of freedom to reduce reconstruction errors in the optimization of the phase mask and can also increase the contrast in the final diffraction pattern. The resulting diffraction pattern is shown in Fig 2.7.

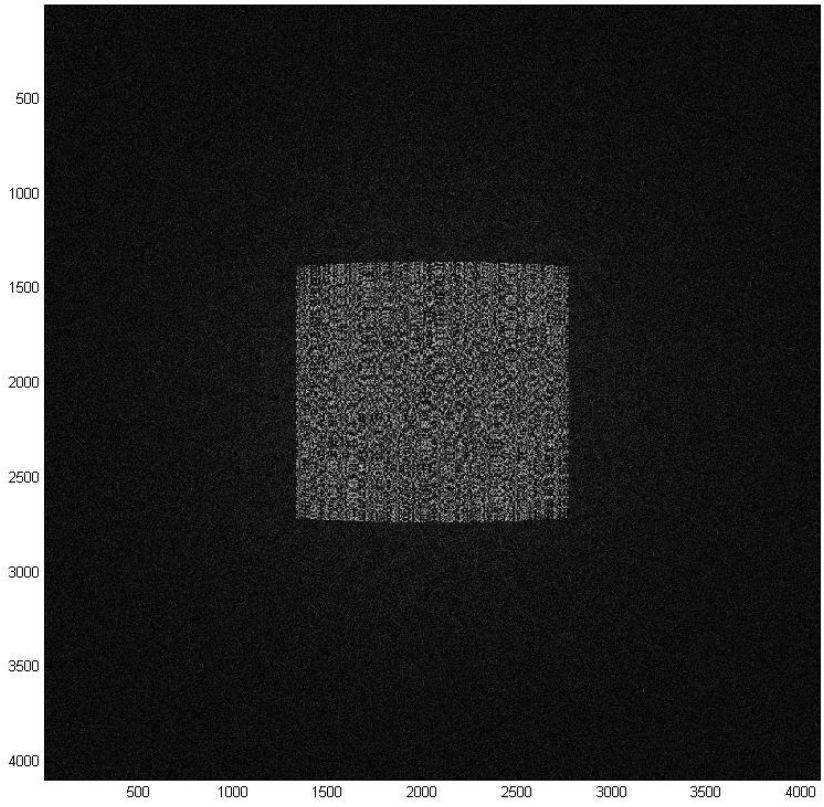


Fig. 2.7 Reconstructed image of 4096x4096 pixels.

The diffractive optical element has been manufactured by photolithography. This method requires the use of a chromed "photomask" whose pattern composed of pixels with or without chromium and will be reproduced in a photoresist by UV exposure. Once developed, the photoresist remains where there has been exposure to UV in the case of a positive photoresist. The fused silica substrate (fused silica) is then etched by reactive ion etching (RIE) where there is no photoresist, as shown in Fig. 2.8. This gives a two-level glass pattern.

The photomask was fabricated by electron beam (e-beam) etching by the Laboratory of Micro and Nanofabrication INRS-Varennes. In this method, the pattern is etched directly on a resin and once it developed, chromium is removed chemically. The pixel size of the photomask and therefore the phase mask is limited to 2x2 microns because it is the minimum pixel size reproducible with the COPL photolithography apparatus. The pattern is of 4096x4096 pixels, therefor, the etched portion of the phase mask is around 8.2x8.2 mm on a 4x4 inch silica substrate.

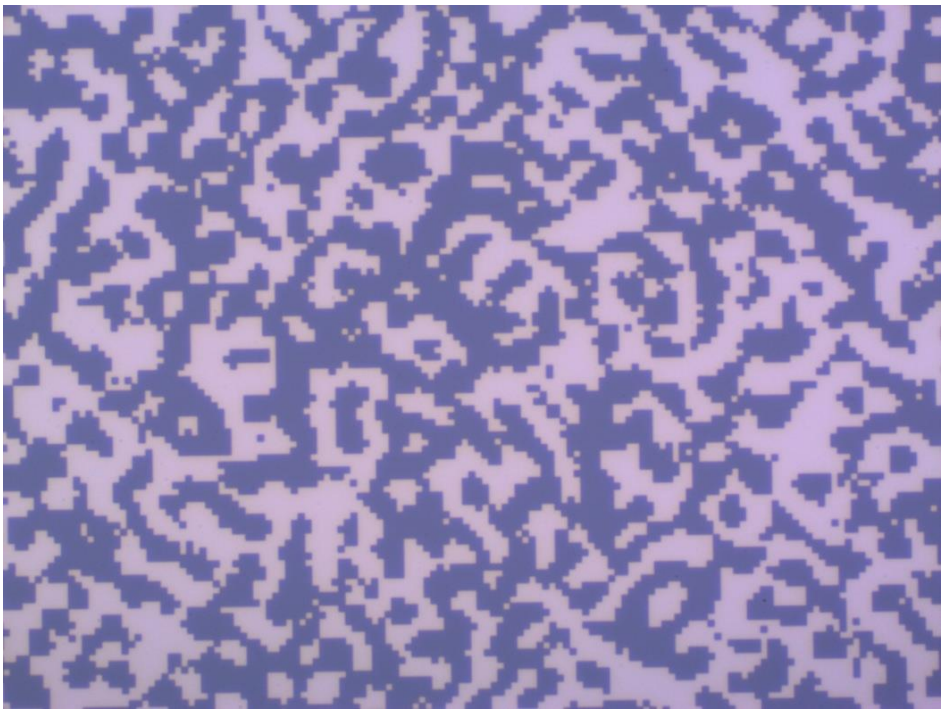
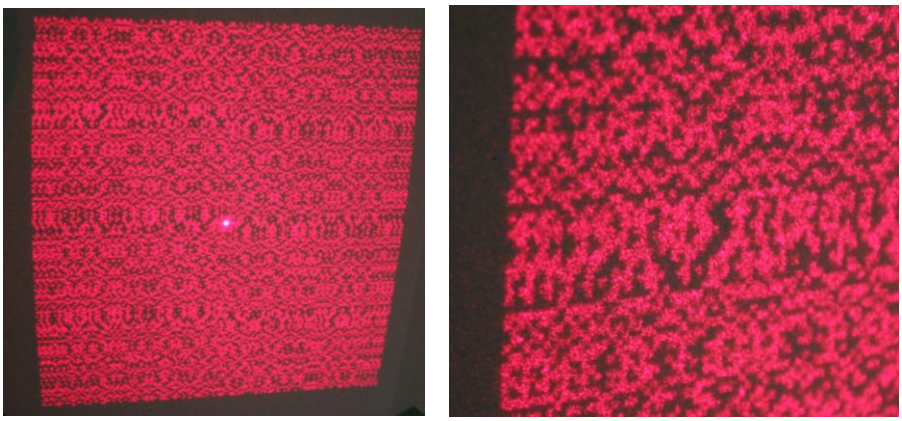


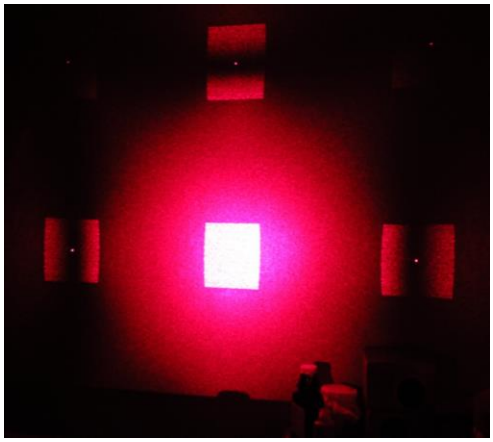
Fig. 2.8 Microscopic image of etched chrome photomask

A fiber laser diode (Blue Sky Research FIBERTEC II, 25mW, 658 ± 5nm) was used to project the pattern. Its light was collimated by an achromatic lens of 50mm focal length. We

tested the magnification system to increase the diffraction angle of the pattern. Our tests have shown that the diffraction angle can be increased in proportion to the inverse of the magnification factor of the zoom system placed at the exit of the diffractive element.



(a) (b)



(c)

Fig. 2.9 (a) Projection of the final pattern with the phase mask reproduced on a fused sillica substrate, (b) Enlarged part of left-top part of projection image (a), (c) the side orders of projection diffraction.

As shown in Fig. 2.9 (a), the target pattern was reproduced with a good enough contrast. An enlarged part of left-top corner of projection image in Fig 2.9 (a) was present in Fig. 2.9 (b). However, since the reproduction of the photomask pattern is never perfectly reproduced on the silica substrate, it introduces additional noise as well as higher diffraction orders as shown in Fig. 2.9 (c). These secondary orders, however, was much smaller in intensity than the central order. This replication also caused the appearance of the undiffracted zero order in

the center of the pattern. The fabrication of DOE and optical reconstruction of pattern were done by Prof. Simon Thibault's group at Ulaval.

2.6 Conclusion

We considered a practical projection image system with certain working specification, e.g. working distance of 40 cm, depth of field of 10 cm and a diffraction angle of 53 degree for 632 nm working wavelength, and then designed and optimized a binary-phase hologram with 2048x2048 pixels by direct binary search for this image system. The hologram was fabricated by E-beam lithography with the resolution of 2 μm. To achieve the required diffraction angle, we discussed the optical architecture in holographic projection image system. The designed CGH and holographic projection image system were validated experimentally by optical reconstruction.

Chapter 3 Computer-generated binary hologram of very large space-bandwidth product for laser projector

3.1 Résumé

Une nouvelle approche pour concevoir l'hologramme à phase binaire avec des ouvertures polygonales de forme arbitraire est proposée. La diffraction de l'ouverture est calculée avec la transformation analytique d'Abbe. Explorer le très grand nombre de degrés de liberté disponibles dans l'hologramme généré par ordinateur (HGO) de très grand nombre de pixels, une optimisation en deux étapes comprenant l'algorithme génétique avec la recherche locale de codage des phases binaires des ouvertures suivie par la recherche directe de co-sommets flottants des ouvertures triangulaires élémentaires, est développé pour des performances élevées et une faible erreur de reconstruction d'image du HGO binaire.

3.2 Abstract

A new approach for designing the binary phase hologram with arbitrary-shaped polygonal apertures is proposed. The diffraction of the aperture is computed with the analytical Abbe transform. To explore the very high number of degrees of freedom available in the computer-generated hologram (CGH) of very large pixel count, a two-step optimization, including the genetic algorithm with local search for encoding binary phases of apertures followed by the direct search for floating covertices of the elementary triangular apertures, is developed for high performance and low image-reconstruction error of the binary CGH.

3.3 Introduction

Optical projector is one of the major products for optical display. The holographic laser projectors are currently under active development because of its advantages for wide color gamut, compact size, light weight, low power consumption, low cost and potential for 3D

display [83]. Commercial holographic laser projectors are widely used for laser show, stage lighting and immersive environment in virtual reality [87-88]. One active application is the virtual keyboard and touch screen, where the holographic laser projector produces a keyboard image on the desk and a trackpad senses the finger's position. The computer-generated hologram (CGH) is also used for generating structural illuminating patterns for 3-D object profile sensing and optical gesture sensing in the Microsoft Kinect^{TM} system [89, 90]. Recent development includes zoomable portable holographic multiple projector [91] and high optical efficiency holographic laser projector for automotive head up displays [92].

In the dynamic holographic laser projector, the key device is the spatial light modulators (SLM) of various types [93]. The phase-only modulation SLM is available [27] and full-range complex amplitude modulation has been implemented by two conjugated couple–mode SLMs [29]. The complex modulation was achieved also by synthesizing two phase functions masked by two complementary checkboard patterns displayed on a single phase-only SLM [94], or two position-shifted amplitude functions displayed on an amplitude SLM and coupled via a sinusoidal grating in the Fourier plane [95]. SLMs usually provide binary amplitude or binary phase modulation, such as the digital micromirror devices (DMD) and the ferroelectric liquid crystal device. Most current SLMs have a resolution of 1920x1080 pixels (HDTV) or 1280x768 pixels (WXGA), which are high resolutions for the image display. However, such a space-bandwidth product (SBWP), e.g. pixel count, is much too small for encoding the interference fringes. Most CGHs are oversampled to provide a sufficient number of degrees of freedom in the CGH design. Many efforts have been devoted to improving the reconstruction quality of the binary CGH in the SLMs. Techniques as the shape adaptive down-sampling, the block truncation coding [96, 97], the breaking down object structure into typical elements [98], the phase shifting [99] and the binary CGH converted from the optical scanning holograms [100] have been proposed recently.

Static CGH is widely useful in the holographic laser projector for 3D display and structured illumination [101]. In contrast with the dynamic CGH using the SLM, the static CGH has a very large pixel count. For instance, with the industrial e-beam lithography the minimum feature size of the e-beam writing can reach to 50 *nm*, and a hologram of one square inch size can be written within 6-8 hours [102]. With a pixel size of 50 *nm*, a CGH of 8 x 8 *mm* (1/3 x

1/3 inch) typical physical size for the laser projector can have 160,000 x 160,000 = 25.6 gigapixels. The number of degrees of freedom can be higher by a factor of 6.25 when defining the e-beam focal spot size of 2 *nm* size, blurred by photoresist diffusion to 20 *nm* as the pixel size. In this chapter, we focused on the design of binary CGHs, which are easy to fabricate and duplicate. The binary amplitude (0,1) and binary phase (-1,1) CGHs consist of binary apertures. Within the scalar diffraction limit, the apertures of subwavelength size must be excluded in the CGH, so that the very high pixel count of the CGH will serve for the high number of degrees of freedom in the CGH design and for the high definition of the aperture shapes. With the conventional laser lithography, a typical CGH can be of 65,000 x 65,000 = 4.3 gigapixels with 1 μm pixel size [103], which is also a large pixel count. Therefore, the challenge is how to manage such a high number of degrees of freedom in the CGH design for image reconstruction of high quality with low error.

The CGH technology has made important progress recently in industrial scale with the dot matrix hologram and holographic prints. The former consists of arrays of local gratings of different specific pitches, orientations and modulation depths with one local grating recording and reconstructing one image point [104]. The latter consists of array of hogels, which is based on a spatial and spectral discretization of computed fringe pattern of the hologram [105]. With the advantages for easy design, low request for the coherence of illuminating light and possibility of colorful display, the dot matrix holograms and holographic prints are widely used for 3-D display and security applications. However, the discrete sampling in the object and hologram domains creates blurring and loss of fidelity in the image. The dot matrix hologram and the holographic prints are not capable of wavefront reconstruction. Moreover, in the fabrication, the local gratings and hogels are displayed on the SLM and recorded optically as grayscale fringes, so that the dot matrix hologram and holographic prints are not binary CGHs.

This chapter is involved with the binary CGHs to explore their very high SBWP for the best performance. Conventional design of the binary hologram with the Gerchberg-Saxton iterative algorithm does not produce satisfactory result as the hard-cut of the continuous phase CGH to binary phase CGH after the iterative design can result in high reconstruction error, while the hard-cut within the iteration loops can lead to stagnation of the iterations.

The binary CGHs are better designed with the Direct Binary Search (DBS) method. The DBS applies the metropolis process to scan the CGH. Starting from a random binary matrix the sign of the bit in a pixel is inverted one-by- one. The inversion, which reduces the error, is retained. Otherwise, the inversion is rejected. The DBS is essentially a local search approach. The genetic algorithm (GA) has a higher probability to achieve a global optimization. In addition, the digital halftone and error diffusion techniques are also useful in the design of the binary CGH. However, all the above methods become unpractical for the computation of the CGH of high pixel count.

As the binary CGH consists of binary polygonal apertures of arbitrary shapes, the latter may be designed and optimized directly. The analytical Abbe transform is used to compute the apertures diffraction as a coherent addition of the diffracted fields from all the straight edges of the polygonal apertures. The two-step optimization algorithm consisting of the hybrid genetic algorithm (GA) for encoding the binary phases of the apertures, followed by the direct search for optimizing positions of the floating co-vertices of the apertures was developed. An example design for a binary CGH of 7744 x 7744 pixel size reconstructing a grayscale level image with the reconstruction error of 2% is shown.

3.4 Abbe transform

In this section, we briefly introduce the Abbe transform, which was used for computing the diffraction fields of 2D objects, as stated by Straubel in his 1888 dissertation and, more recently, by Komrska in 1982, who applied the Abbe transform to the diffraction of polygonal apertures [106].

The Fraunhofer diffraction of an aperture can be computed by the Abbe transform. When an aperture $\Omega(\xi_1, \xi_2)$ of unit transmission is illuminated by a plane wave, the complex- valued diffracted field is computed as

$$U(\vec{p}) = \iint_{\Omega} \exp(-i2\pi \vec{p} \cdot \vec{\xi}) d\vec{\xi} \qquad (3.1)$$

where $\vec{\xi}(\xi_1, \xi_2)$ and $\vec{p}(m,n)$ is the position vector in the aperture and the Fraunhofer diffraction plane, respectively. Eq. (3.1) represents the Fourier transform of the aperture $\Omega(\xi_1, \xi_2)$. If the aperture is periodically repeated in the plane (ξ_1, ξ_2), the diffraction pattern consists of punctual diffraction orders located by the integer coordinates (m, n).

The integrand $\psi(\vec{\xi}) = \exp(-i2\pi\vec{p}\cdot\vec{\xi})$ are the plane waves satisfies the Helmholtz equation $\nabla^2\psi + k^2\psi = 0$ with the wave vector $\vec{k} = 2\pi\vec{p}$, so by using of the 2D Gauss theorem, the Eq. (3.1) can be rewritten as

$$U(\vec{p}) = \frac{-1}{(2\pi p)^2}\iint_{\Omega}\nabla^2\psi\, d\vec{\xi} = \frac{-1}{(2\pi p)^2}\oint_C \nabla\psi\cdot\hat{n}dl \qquad (3.2)$$

where C is the boundary of the aperture Ω, $\hat{n}$ donates the outer normal to the contour, and dl is a vector of differential length along C. Since $\nabla\psi = -i2\pi\psi\vec{p}$, the 2D Abbe transform of the surface integral in Eq. (3.2) is a contour integral computed as

$$U(\vec{p}) = \frac{i}{2\pi p^2}\oint_C \exp(-i2\pi\vec{p}\cdot\vec{\xi})\vec{p}\cdot\hat{n}dl \qquad (3.3)$$

The transformation of the surface integral in Eq. (3.1) into the contour integral along the boundary C of the aperture is referred as the 2D Abbe transform.

Furthermore, if the boundary C is polygonal, the contour integration in Eq. (3.3) becomes a summation of the contributions from its straight edges as

$$U(\vec{p}) = \frac{i}{2\pi p^2}\sum_{q=1}^{Q}\vec{p}\cdot\hat{n}_q L_q \sin c\left(\vec{p}\cdot\hat{t}_q L_q\right)\exp\left(-i2\pi\vec{p}\cdot\hat{\xi}_{M_q}\right) \qquad (3.4)$$

where Q is the number of the straight edges in the polygon, $\vec{L}_q = \vec{\xi}_{q+1} - \vec{\xi}_q$ with $\vec{\xi}_q$ denoting position vector of the vertex, and $\vec{\xi}_{M_q}$ denoting the midpoint of the edge q, which has the length L_q, the unit external normal vector $\hat{n}_q$ and the unit tangent vector $\hat{t}_q = \vec{L}_q / L_q$, as shown in Fig. 3.1, and $\sin c(x) = \sin(\pi x)/(\pi x)$. From Eq. (3.4), the diffraction of a straight edge *q*, shows maximum amplitude on a line normal to the edge and passing through the midpoint of the edge. On this line $\vec{p}\cdot\hat{n}$ is maximum and $\sin c(\vec{p}\cdot\hat{t}_q L_q) = 1$. The diffraction amplitude

decreases fast from the midpoint when the position vector $\vec{p}$ leaves the direction of the edge normal. The phase of the diffraction pattern of edge q is expressed as $\exp(i2\pi\vec{p}\cdot\vec{\xi}_{M_q})$. The diffraction pattern of the polygonal aperture is then the coherent summation of the diffraction patterns of all its edges.

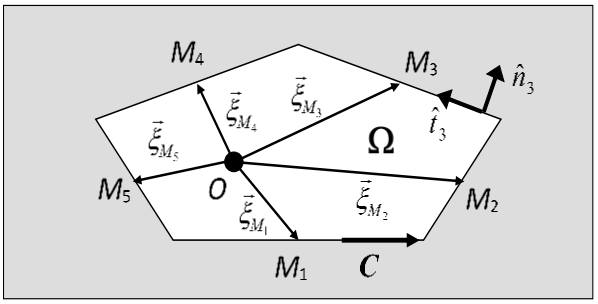


Fig. 3.1 Schematic diagram of polygonal aperture Ω with Q=5 straight edges.

3.5 Design of binary CGH

The binary amplitude-only CGH was first invented by Lohmann in 1967. The design of the binary CGH using the iterative phase retrieval algorithm does not result in good performance. Moreover, the iterations with the fast Fourier transform (FFT) of the CGH of 8,000 x 8,000 pixel size is not effective, and the FFT of the potential CGH of 160,000 x 160,000 pixels could be prohibited by the high computational cost. The direct binary search (DBS) in general results in a good performance [37]. The DBS can be accelerated by using the analytical expressions with some acceptable approximations to compute the cost function directly without computing the reconstructed image with the FFT at each inversion of the bit [38]. As the search progresses, the pixels assigned with the same binary values are clustered to form polygonal apertures of arbitrary shapes. In the DBS after the first round of scanning the apertures in the CGH are basically formed. The subsequent rounds by the DBS can reduce the error further. However, for binary CGH with very high SBWP the conventional design methods would fail by the prohibited computational cost.

3.5.1 Triangle-based layout

As we mentioned earlier, the binary CGH consists of polygonal apertures. Since any arbitrary-shaped polygonal aperture can be divided into arbitrary-shaped elementary triangular apertures, the design for the binary CGH can be based on the component triangles. Each period of the binary CGH is divided into *M*x*N* rectangular cells. Each cell of *P*x*Q* pixels is further divided into four triangles, labeled with the index *s* = 1-4. The co-vertex of the 4 triangles in each cell is not fixed at the center of the cell but can be floating within a predefined region, so that the component triangles can have arbitrary shapes, as shown in Fig. 3.2. After the triangles are assigned with binary values, the neighbouring triangles of the same binary value are clustered to form polygonal apertures in the CGH. In the proposed design, the number *M*x*N* of the cells are equal to the number of pixels in the reconstructed image. The size *P*x*Q* of the cells is decided by the available SBWP of the CGH with consideration of two facts: First, the size *P*x*Q* does not affect the computational cost, as the diffractions are computed with the analytical Abbe transform. Second, the physical size of the cell must be larger than the wavelength, as the design uses the scalar diffraction theory.

3.5.2 Diffraction of binary CGH

Consider now a binary CGH with the period divided into *M*x*N* rectangular cells. Each cell of *P*x*Q* pixels is divided into four triangles. In all the cell (*j*, *r*) with *j*=0, 1, 2, …, *M*-1 and *r*=0, 1, 2, …, *N*-1, the 4 triangles are indexed by *s*=1…4 in the same manner, as shown in Fig. 3.2. The co-vertex of the 4 triangles in each cell is at an arbitrary location in the cell. Referring to Fig. 3.2 (a) and applying the Abbe transform Eq. (3.4) to a particular triangular aperture of index *s* located in the cell (*j*, *r*), carefully writing down the position vector $\vec{p}(m,n)$ and the normal, tangent and midpoint position vectors, $\hat{n},\hat{t},\hat{\xi}_M$ for each edge *q*, we obtained the diffracted field of the aperture *s*, as a coherent summation of the contributions from its three edges as

$$U_s(m,n,j,r)=\left\{\Theta_s(m,n,j,r)\exp\left[-i2\pi\left(\frac{m}{M}j+\frac{n}{N}r\right)\right]\right\}\times\exp\left[i\varphi_s(j,r)\right] \qquad (3.5)$$

where $\varphi_s(j,r)$ is the binary phase value of the aperture *s*, and $\Theta_s(m,n,j,r)$ is from the Abbe transform.

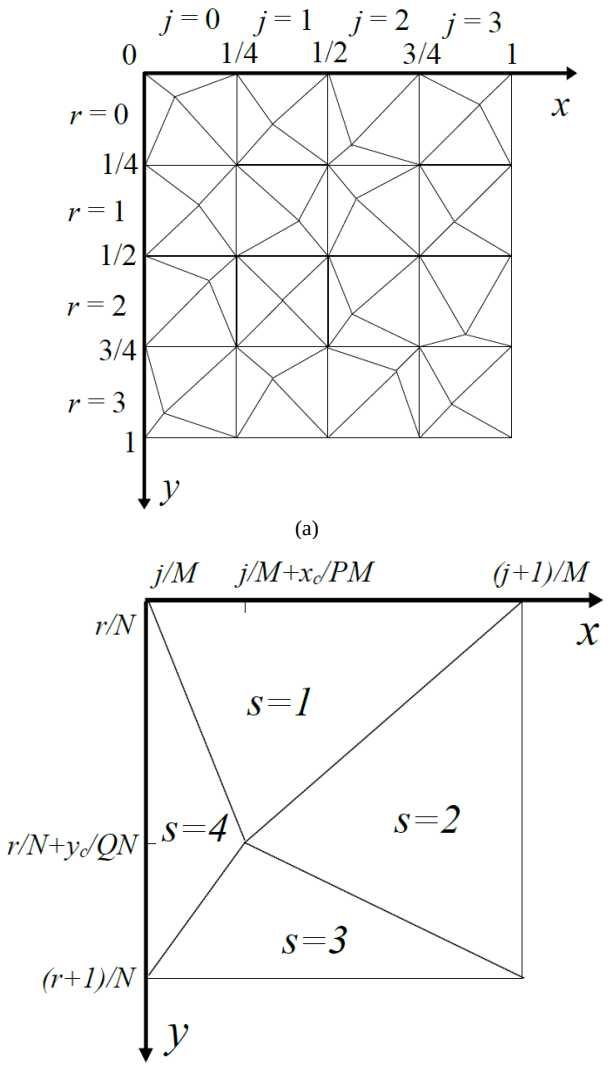


Fig. 3.2 CGH layout (a) One period of CGH with rectangular cells divided to four triangular apertures; (b) arbitrary-shaped triangles in a cell.

For a triangular aperture s =1 in cell (j, r), $\Theta_1(m,n,j,r)$ is expressed as

$$\Theta_1(m,n,j,r)=\frac{i}{2\pi(m^2+n^2)}\left\{\left[\left(-\frac{n}{M}\right)\sin c\left(\frac{m}{M}\right)\exp\left(-i\pi\frac{m}{M}\right)\right]+\left[\left(\frac{m}{N}\cdot\frac{y_c(j,r)}{Q}+\frac{n}{M}\cdot\left(1-\frac{x_c(j,r)}{P}\right)\right)\times\right.$$
$$\left.\sin c\left(\frac{m}{M}\cdot\left(\frac{x_c(j,r)}{P}-1\right)+\frac{n}{N}\cdot\frac{y_c(j,r)}{Q}\right)\times\exp\left(-i\pi\left(\frac{m}{M}\cdot\left(1+\frac{x_c(j,r)}{P}\right)+\frac{n}{N}\cdot\frac{y_c(j,r)}{Q}\right)\right)\right]$$
$$+\left[\left(\frac{m}{N}\cdot\left(-\frac{y_c(j,r)}{Q}\right)+\frac{n}{M}\cdot\frac{x_c(j,r)}{P}\right)\times\sin c\left(\frac{m}{M}\cdot\left(-\frac{x_c(j,r)}{P}\right)+\frac{n}{N}\cdot\left(-\frac{y_c(j,r)}{Q}\right)\right)\times\right.$$
$$\left.\left.\exp\left(-i\pi\left(\frac{m}{M}\cdot\frac{x_c(j,r)}{P}+\frac{n}{N}\cdot\frac{y_c(j,r)}{Q}\right)\right)\right]\right\} \qquad (3.6)$$

which depends only on the coordinates of the 3 vertices: two vertices are at the two corners of the cell, and the third one at (x_c, y_c) is the co-vertex of the 4 triangles in the cell (j, r).

In the same way, the complex-valued function $\Theta_2(m,n,j,r)$ is given by

$$\begin{aligned}
\Theta_2(m,n,j,r) &= \frac{i}{2\pi(\mathrm{m}^2+\mathrm{n}^2)}\left\{\left(\frac{m}{N}\right)\sin c\left(\frac{n}{N}\right)\exp\left[-i\pi\left(\frac{2m}{M}+\frac{n}{N}\right)\right]\right. \\
&+\left(\frac{m}{N}\cdot\left(\frac{y_c(j,r)}{Q}-1\right)+\frac{n}{M}\cdot\left(1-\frac{x_c(j,r)}{P}\right)\right)\sin c\left(\frac{m}{M}\cdot\left(\frac{x_c(j,r)}{P}-1\right)+\frac{n}{N}\cdot\left(\frac{y_c(j,r)}{Q}-1\right)\right) \\
&\exp\left(-i\pi\left(\frac{m}{M}\cdot\left(\frac{x_c(j,r)}{P}+1\right)+\frac{n}{N}\cdot\left(\frac{y_c(j,r)}{Q}+1\right)\right)\right)+\left(\frac{m}{N}\cdot\left(-\frac{y_c(j,r)}{Q}\right)+\frac{n}{M}\cdot\left(\frac{x_c(j,r)}{P}-1\right)\right) \\
&\left.\sin c\left(\frac{m}{M}\cdot\left(1-\frac{x_c(j,r)}{P}\right)+\frac{n}{N}\cdot\left(-\frac{y_c(j,r)}{Q}\right)\right)\exp\left(-i\pi\left(\frac{m}{M}\cdot\left(\frac{x_c(j,r)}{P}+1\right)+\frac{n}{N}\cdot\frac{y_c(j,r)}{Q}\right)\right)\right\}
\end{aligned} \tag{3.7}$$

Also, $\Theta_3(m,n,j,r)$ and $\Theta_4(m,n,j,r)$ are expressed respectively as

$$\begin{aligned}
\Theta_3(m,n,j,r) &= \frac{i}{2\pi(\mathrm{m}^2+\mathrm{n}^2)}\left\{\left(\frac{n}{M}\right)\sin c\left(-\frac{m}{M}\right)\exp\left[-i\pi\left(\frac{m}{M}+\frac{2n}{N}\right)\right]\right. \\
&+\left(\frac{m}{N}\cdot\left(\frac{y_c(j,r)}{Q}-1\right)+\frac{n}{M}\cdot\left(-\frac{x_c(j,r)}{P}\right)\right)\sin c\left(\frac{m}{M}\cdot\left(\frac{x_c(j,r)}{P}\right)+\frac{n}{N}\cdot\left(\frac{y_c(j,r)}{Q}-1\right)\right) \\
&\exp\left(-i\pi\left(\frac{m}{M}\cdot\frac{x_c(j,r)}{P}+\frac{n}{N}\cdot\left(\frac{y_c(j,r)}{Q}+1\right)\right)\right)+\left(\frac{m}{N}\cdot\left(1-\frac{y_c(j,r)}{Q}\right)+\frac{n}{M}\cdot\left(\frac{x_c(j,r)}{P}-1\right)\right) \\
&\left.\sin c\left(\frac{m}{M}\cdot\left(1-\frac{x_c(j,r)}{P}\right)+\frac{n}{N}\cdot\left(1-\frac{y_c(j,r)}{Q}\right)\right)\exp\left(-i\pi\left(\frac{m}{M}\cdot\left(\frac{x_c(j,r)}{P}+1\right)+\frac{n}{N}\cdot\left(\frac{y_c(j,r)}{Q}+1\right)\right)\right)\right\}
\end{aligned} \tag{3.8}$$

$$\begin{aligned}
\Theta_4(m,n,j,r) &= \frac{i}{2\pi(\mathrm{m}^2+\mathrm{n}^2)}\left\{\left(-\frac{m}{N}\right)\sin c\left(-\frac{n}{N}\right)\exp\left[-i\pi\left(\frac{n}{N}\right)\right]\right. \\
&+\left(\frac{m}{N}\cdot\left(\frac{y_c(j,r)}{Q}\right)+\frac{n}{M}\cdot\left(-\frac{x_c(j,r)}{P}\right)\right)\sin c\left(\frac{m}{M}\cdot\left(\frac{x_c(j,r)}{P}\right)+\frac{n}{N}\cdot\left(\frac{y_c(j,r)}{Q}\right)\right) \\
&\exp\left(-i\pi\left(\frac{m}{M}\cdot\left(\frac{x_c(j,r)}{P}\right)+\frac{n}{N}\cdot\left(\frac{y_c(j,r)}{Q}\right)\right)\right)+\left(\frac{m}{N}\cdot\left(1-\frac{y_c(j,r)}{Q}\right)+\frac{n}{M}\cdot\left(\frac{x_c(j,r)}{P}\right)\right) \\
&\left.\sin c\left(\frac{m}{M}\cdot\left(-\frac{x_c(j,r)}{P}\right)+\frac{n}{N}\cdot\left(1-\frac{y_c(j,r)}{Q}\right)\right)\exp\left(-i\pi\left(\frac{m}{M}\cdot\frac{x_c(j,r)}{P}+\frac{n}{N}\cdot\left(\frac{y_c(j,r)}{Q}+1\right)\right)\right)\right\}
\end{aligned} \tag{3.9}$$

If the co-vertices are in the middle of the cells, for the Equations above, we will know that $\Theta_3(m,n,j,r)=\Theta_1^*(m,n,j,r)$, $\Theta_4(m,n,j,r)=\Theta_2^*(m,n,j,r)$, which is understandable since the triangle s=3 (s=4) is centrosymmetric to the triangle s=1 (s=2) with respect to the center of the cell (j, r).

The complex amplitude of the diffraction pattern of the entire CGH is a coherent addition of all the diffracted fields from all the triangular apertures, which is given by

$$\begin{aligned} T(m,n) &= \sum_{s=1}^{s=4}\sum_{j=0}^{j=M-1}\sum_{r=0}^{r=N-1} U_s(m,n,j,r) \\ &= \sum_{s=1}^{s=4}\sum_{j=0}^{j=M-1}\sum_{r=0}^{r=N-1}\left\{\Theta_s(m,n,j,r)\exp\left[-i2\pi\left(\frac{m}{M}j+\frac{n}{N}r\right)\right]\exp\left[i\varphi_s(j,r)\right]\right\} \end{aligned} \tag{3.10}$$

Note that when two neighboring triangles assigned with the same binary value were merged in a polygon, the contributions of their shared edge to the diffraction pattern will be computed twice on the same line but in the opposite directions and will be therefore canceled.

If the locations of the co-vertices in each cell are the same: $x_c(j,r)\equiv\bar{x}, \mathrm{y}_c(j,r)\equiv\bar{y}$, it also can be seen that, $\Theta_s(m,n,j,r)=\Theta_s(m,n)$ is independent of the cell location (j, r) and the phase value $\varphi_s(j,r)$ of the aperture. In contrast to $\Theta_s(m,n)$, the remaining factor in Eq. (3.10) depend only on the cell position (j, r) and phase value $\varphi_s(j,r)$ of the aperture and are independent of the aperture layout in the cell. In such a case, the whole diffraction pattern could be rewritten as

$$\begin{aligned} &T(m,n) \\ &= \sum_{s=1}^{s=4}\left\{\Theta_s(m,n)\sum_{j=0}^{j=M-1}\sum_{r=0}^{r=N-1}\left\{\exp\left[-i2\pi\left(\frac{m}{M}j+\frac{n}{N}r\right)\right]\exp\left[i\varphi_s(j,r)\right]\right\}\right\} \\ &= \sum_{s=1}^{s=4}\left\{\Theta_s(m,n)\cdot\mathrm{DFT}\left\{\exp\left[i\varphi_s(j,r)\right]\right\}\right\} \end{aligned} \tag{3.11}$$

For the Eq. (3.11), two factors contribute to the diffraction patterns of triangle-based CGH: the position (x_c, y_c) of the co-vertex in each cell and the phase value φ_s of each triangular aperture, which provide a large number of degree of freedom to be exploited. Basically, the position (x_c, y_c) of the co-vertex could be take any values as long as $1<x_c<P$ and $1<y_c<Q$. For the binary CGH, the phase value φ_s could only be 0 or π.

Form the Abbe transform, we know that the diffraction of a polygonal aperture with straight edges may be computed by the analytical expressions, which depend only on the positions of its vertices. This conclusion could also be understood by Sommerfeld's qualitative discussion [107]: a straight edge of the diffracting object can act as an infinitely narrow slit, which sends a light fan in the direction perpendicular to the edge, and the fan width is inversely proportional to the edge length. The diffraction pattern is a coherent summation of all the

light fans from all the edges of different orientations and lengths in the CGH. Thus, the basic physical elements in the binary CGH for the diffraction are the straight edges of the apertures. Optimization of the binary CGH by varying the orientations and length of all the straight edges, i.e. the arbitrary shapes of the apertures, could be efficient, as will be shown later.

3.6 Design algorithms

To manage the high number of degrees of freedom, the proposed design for the binary CGH is performed in two steps. The first step is to assign the binary phase values to the triangular apertures. This step uses the hybrid GA. The second step is to optimize the positions of the floating co-vertices of the 4 triangles in the cells. That optimizes the aperture shapes and the lengths and orientations of the edges. The second step uses the direct local search algorithm.

3.6.1 Hybrid genetic algorithm for assigning binary phases

The CGH consists of *M*x*N* rectangular cells with each cell of *P*x*Q* pixels divided into 4 triangular apertures. In the first step, the binary phase values in the component triangular apertures are assigned with the hybrid GA. All the co-vertices of the 4 triangles in the cells are at the cell centers, such that a square cell consists of 4 right isosceles triangles at different positions associated with the index s = 1- 4. The shapes of the triangles will not be altered in the first step. The chromosomes are the vectors of 4*M*x*N* bits. Each bit is encoded with values 1 or -1, representing the binary phase value of the corresponding triangle in the binary CGH. The decoding procedure from the bits of a chromosome to the binary value of a CGH is: For a randomly generated chromosome X with 4MN genes represented by 1 or -1, the first four value X_1, X_2, X_3, X_4 of the genes in a chromosome X is assigned to the four triangular s=1, 2, 3, 4 in the cell (j=0, r=0), and then the second four value X_5, X_6, X_7, X_8 to the four triangular s=1, 2, 3, 4 in the (j=1, r=0), and so on, till the last four value X_{4MN-3}, X_{4MN-2}, X_{4MN-1}, X_{4MN} to these in the cell (j=M-1, r=N-1). In the GA, an initial population of N_0+2 chromosomes are generated by a random bit generator with a uniform probability distribution at the beginning. The value of N is usually between 60~120. A large number of population would cost too much time to convergence while a small number would make it hard to convergence

a good enough solution. In the classical GA, we use the rank-based selection, the nonlinear cumulative normal distribution fitness, the multiple point crossover and the elitism. Moreover, a local search is applied in the process of GA, which could accelerate the speed of convergence. The CGHs represented by the chromosomes are evaluated by the cost function, *e.g.* root-mean-square (RMS) errors of their reconstructed images, which defined as

$$e = \sqrt{\sum_{a=-M/2}^{M/2} \sum_{b=-N/2}^{N/2} \left(\left| f(a,b) \right|^2 - \left| f_t(a,b) \right|^2 \right)^2} \qquad (3.12)$$

where the integers a and b define the locations of pixels, $f(a, b)$ and $f_t(a, b)$ are the diffracted amplitude of the reconstructed image and the ideal diffracted amplitude of the target image, respectively.

The selection based on the error values often leads to a premature convergence. Thus, the chromosomes in the population are first ranked according to their errors in an ascending order. The best chromosomes with the least errors are ranked at the top rank and the worst one is ranked at the bottom. Moreover, the two worst chromosomes are discarded from the population for the reproduction. A roulette wheel selection is performed N_0 times for matting the chromosomes, based on the fitness, which is in fact the the selection probability defined as

$$\Phi(p) = 1 - C_g(p) = 1 - \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{p} \exp\left[\frac{(y-\mu)^2}{\sigma^2} \right] dy \qquad (3.13)$$

which is then normalized as

$$\Phi'(p) = \Phi(p) / \left[\sum_{p'}^{N_0} \Phi(p') \right] \qquad (3.14)$$

where $C_g(p)$ is the cumulative normal distribution, p is the rank of the chromosomes, and $\mu = N_0/2$ is the mean value and $\sigma = N_0/6$ is the standard deviation of the normal distribution, respectively.

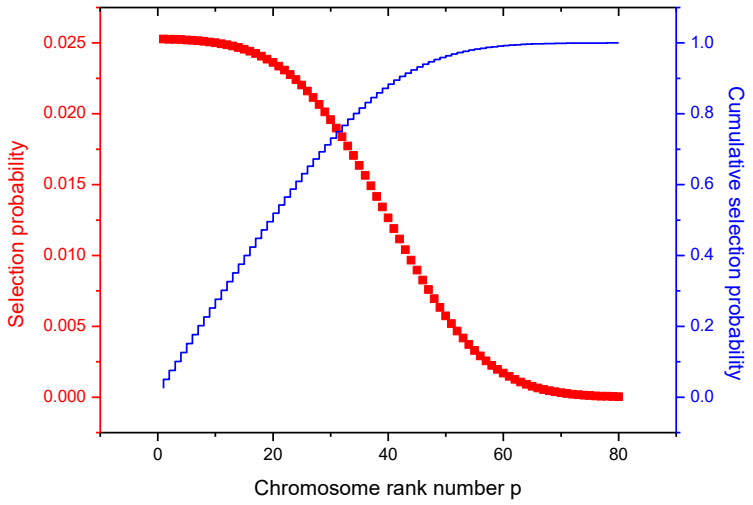


Fig. 3.3 The selection probability and cumulative selection probability distribution for all ranked chromosome.

In the roulette wheel process, a random number r ranged from 0 to 1 is generated by the random number generator with the uniform probability distribution. If $\Phi'(p-1) \le r \le \Phi'(p)$, then the chromosome at rank p is selected. For example, in Fig. 3.4, if r=0.84, the chromosome ranked at p=37 has been selection. With this high nonlinear fitness function, the better individuals with lower errors have higher probability to survive and to produce the offspring of next generations.

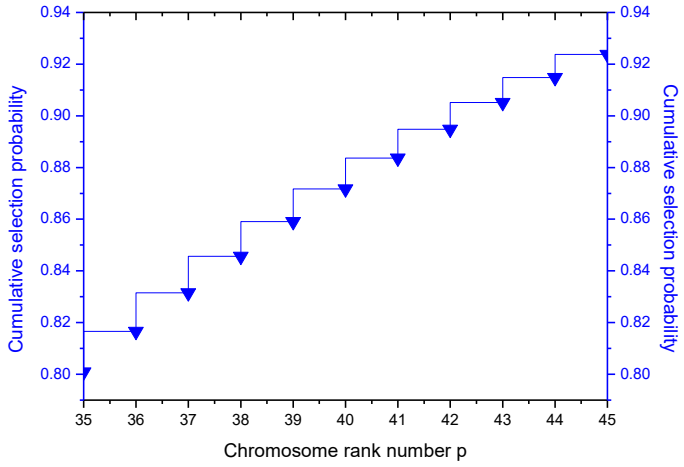


Fig. 3.4 The selection probability and cumulative selection probability distribution for all ranked chromosome.

After N_0 times selection, a mating population is formed, then two genetic operators, crossover and mutation, are applied to the chromosomes in the mating population. In the double-point crossover, the position of the first exchange point was randomly chosen between 1 to 4*MN*-1. The position of the second exchange point was randomly chosen between the first point position and 4*MN*. For randomly two paired chromosomes, the bits between these two points are exchanged and thus two new chromosomes are produced. The double-point crossover is performed with a high probability Pc>0.6 and If the crossover does not occur, the paired chromosomes enter the population directly for mutation.

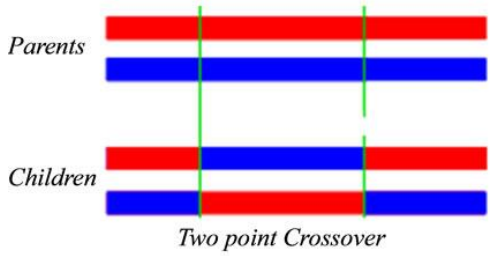


Fig. 3.5 Schematic diagram of two-point crossover

After the crossover, the mutation for maintaining the genetic diversity of the genes is performed, which flips each bit in all the chromosome with an exponentially decreasing mutation probability P_m defined as $P_m(k+1) = \alpha P_m(k) = \alpha^k P_m(1)$ with $\alpha \leq 1$. Thus $P_m(k)$ decreases with the generation number k. For every bit in the chromosomes a random number *R* ranged in (0, 1) is generated. If $R < P_m(k)$, then the bit is flipped. Otherwise it is not flipped. At the begin of the mutation, a relatively large number of bits are inversed, while close to the end of the mutation, only a few number of bits are flipped as $P_m(k)$ decreases with k.

Furthermore, the elitism was applied to prevent the best 2 individuals in the family from deterioration. The elites were directly entered to the next generation without crossover and mutation. In the same time, the 2 worst members are discarded in each generation.

The hybrid GA is a combination of the GA with the local search. As the classical GA performs long jumps in the solution space to look for the global minimum, introduction of the local search process within the GA can improve the optimization significantly [15]. Before sending the population to next generation, a local search is performed. The direct binary search (DBS) is applied to a randomly chosen chromosome in the population by scanning and flipping the binary phase value of every bit in the chromosome and retained

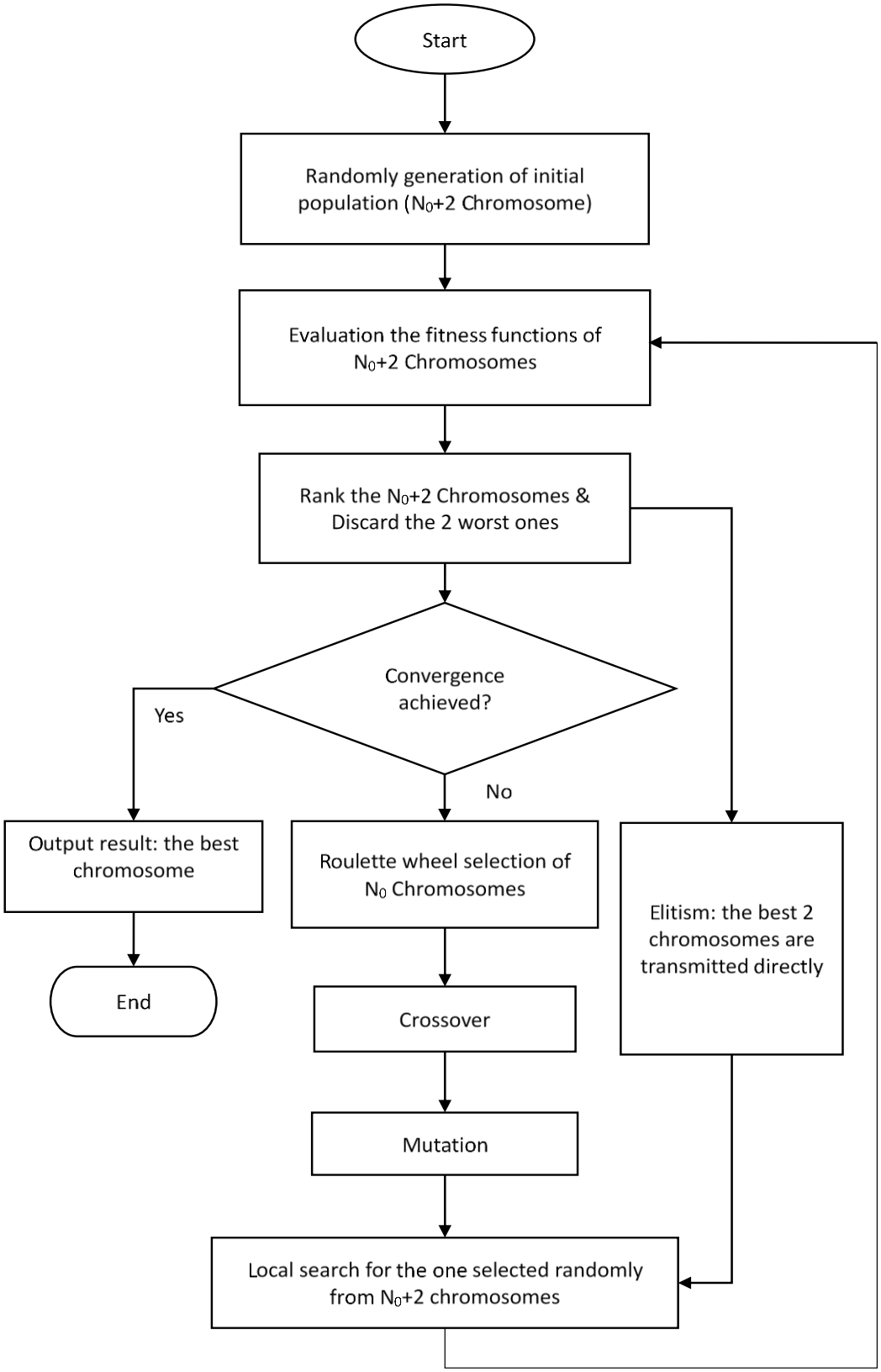


Fig. 3.6 Flowchart of the hybrid GA for designing CGH of polygonal apertures.

only the change which reduces the error. The new diffraction pattern T'(m, n) after flipping phase value of triangle s in cell (j,r) could be computed as

$$T'(m,n) = T(m,n) - U_s(m,n,j,r) + U'_s(m,n,j,r) \quad (3.15)$$

, where U and U' is the diffraction pattern of the triangular aperture s in cell (j,r) before and after flipping the phase value, respectively. The hybrid GA is iterated until the convergence to a final solution. Fig. 3.6 shows the flowchart of the hybrid GA with the local search including the elitism.

3.6.2 Direct search for optimization of the floating co-vertices

After the hybrid GA for optimizing the binary phase values of all the triangular apertures, the best individual in the population is selected for further optimization. In the first step of design all the triangular apertures are right isosceles triangles. In the second step of the design, the orientation of the straight edges and the shape of the polygonal apertures are optimized by optimizing the position of the floating co-vertex of the 4 triangles in each cell using the direct search method, resulting in immediate changes of the lengths and orientations of 4 straight edges and the shape of the 4 triangles. The floating of the co-vertices of the 4 triangles creates additional degrees of freedom and can optimize the binary CGH efficiently. The direct search algorithm moves each co-vertex step-by-step within a window around the center of the cell. The move that reduces the cost function was retained. Otherwise, the move is rejected. The relative small window located in the middle of the cell was used to prevent the potential large digitization error since the co-vertex move too close to edges of the cell. The new diffraction pattern T'(m, n) after move the co-vertex in the cell (j, r) could be computed as

$$T'(m,n) = T(m,n) - \sum_{s=1}^{s=4} U_s(m,n,j,r) + \sum_{s=1}^{s=4} U'_s(m,n,j,r) \quad (3.16)$$

where U and U' is the diffraction pattern of the triangular aperture (j, r, s) before and after moving the position (x_c, y_c), respectively. For sake of computation time, the move of the co-vertex was stepwise with a discrete step of 4 pixels. The window, in which the co-vertex was floated, was located in the middle of the cell to prevent from the co-vertex too close to boundary of the cell, resulting in the too narrow triangular apertures. The direct search scans

the CGH cell-by-cell in several rounds. The direct search for optimal co-vertex positions could also be performed within the hybrid GA loops for improving the best chromosome in each generation. However, this algorithm can take too much time but did not always produce better performance.

3.7 Experimental results

The binary CGHs with arbitrary-shaped polygonal apertures were designed by the hybrid GA followed by the optimization of the co-vertex positions. A design example was for a grayscale image of the notepad keyboard. The image size was of 242 x 242 pixels with an object window of 240 x 240 pixels. The number of the cells per period in the CGH was set as $M = N = 242$. Each cell was chosen to have a size of PxQ = 32x32 pixels. Assume a pixel size of 50 *nm* by the e-beam lithography, the cells of 32x32 pixels have a physical size of 1.6 x 1.6 μm, which is larger than the illuminating wavelength. The CGH was of 7744 x7744 pixels with the physical dimension of 0.4 x 0.4 *mm*.

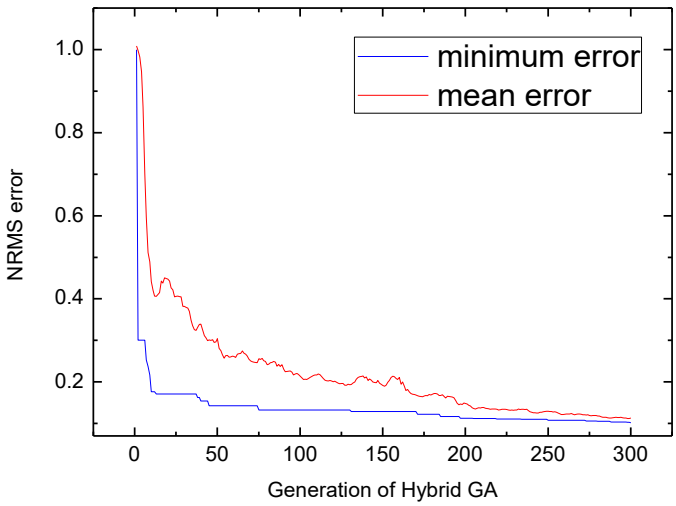
3.7.1 First step of hybrid GA

In the first step of design with the hybrid GA, the initial population of 80+2 chromosomes were generated at random. Each chromosome was a vector of $4MN$ = 234,256 bit long. After evaluating the image reconstruction errors and discarding the two worst solutions, the 80 chromosomes were selected with a roulette wheel game for reproduction of next generation. The selection probability was based on the rank of the chromosomes as given in Eq. (3.14). The roulette wheel games were performed 80 times, mating 40 pairs of parent chromosomes. Each pair generated 2 descendants with the double-point crossover, in which the crossover probability was P_c=0.9. The mutation with the exponentially decreasing probability $P_m(k+1) = \alpha P_m(k) = \alpha^k P_m(1)$ was performed with $P_m(1) = 0.1$ and α=0.985 and k is the number of generation.

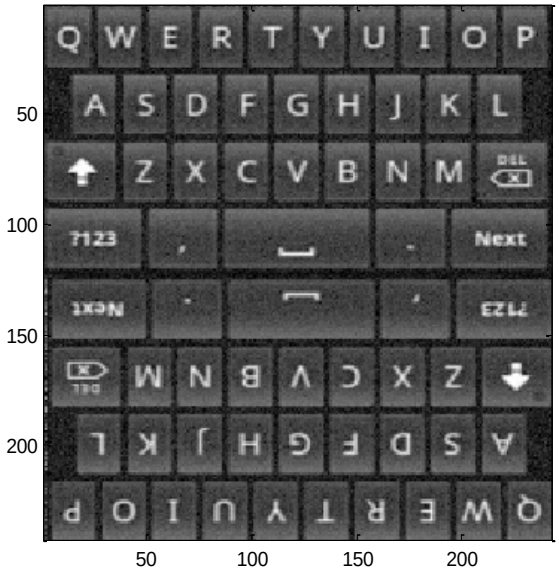
After the mutation, 80 new offspring chromosomes were produced in addition to the 2 elite chromosomes, which were transmitted directly without crossover and mutation. Then the local search was performed for one chromosome randomly chosen in the new population.

The local search scanned and flipped the binary phase value of every bit in the chosen chromosome and retained only the changes which reduced the error. Finally, the new population of 82 chromosomes went to the same process for the next population.

The minimum normalized RMS error of the best chromosome in the population and the mean error of the population are plotted in Fig. 3.7 (a) as a function of the number of generations. The mean error of the whole population showed some fluctuations before converged to 0.12 after 300 generations. The reason is that the random selection and crossover can not guarantee



(a)



(b)

Fig. 3.7 Results from the hybrid GA for assigning binary phases: (a) Normalized RMS errors as a function of generation; (b) Grayscale image of keyboard reconstructed by the CGH.

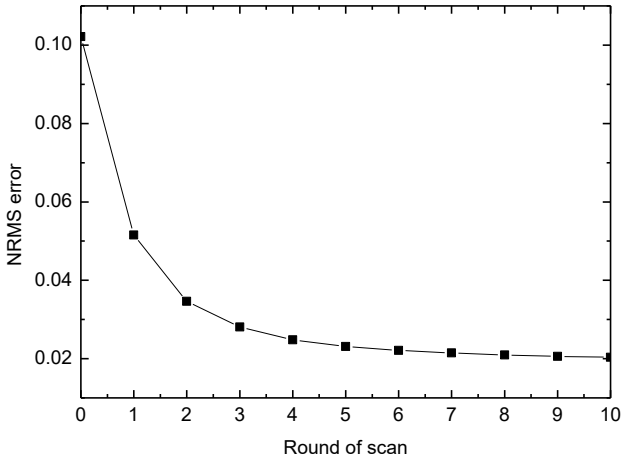
the chromosomes get better after each generation. But the whole population was well evolved in general as the generation goes. The minimum error of the best solution dropped drastically in few first generations. This can attribute to the local search incorporated in the GA. After the first generation of hybrid GA the error was decreased to 0.3. At this level, the binary apertures in the CGH were basically formed. Then, the minimum error decreased monotonically thank to the elitism, as can be seen in Fig. 3.7 (a). The minimum error was reduced to 0.14 after the first 50 generations and reached 0.1 after 300 generations. In the last 100 generations, the minimum error decreased from 0.11 to 0.10. This was still a good convergence. The reconstructed image with the normalized RMS error of 0.10 from the best chromosome after 300 generations is shown in Fig. 3.7 (b), in which the noise on the image is still visible.

3.7.2 Second Step of floating co-vertices

In the hybrid GA the binary phase values were assigned to all the right isosceles triangular apertures. Then, the best chromosome in the population was chosen for further optimization. In this step all the co-vertices in the cells were permitted to float and a direct position search for the optimal co-vertex positions was performed. This step changed the shapes of all the triangles, and also changed the lengths and orientations of all the straight edges of the apertures. This step was critical to improve the quality of the reconstructed images, according the qualitative description by Sommerfeld for the Abbe transform.

The co-vertex was floating in a window of 8 x 8 pixels located at the center of the cell of 32x32 pixels size to avoid the generations of too small triangles. The co-vertex in each cell was floated step-by-step with a discrete step of 4 pixels in row and column, respectively, resulting in 9 possible positions to be chosen. The co-vertex position which reduces the error was kept, otherwise it was reject. The direct search was performed within a cell and then cell-by-cell and scanned the CGH for 10 rounds.

The error was decreased from 0.10 by the hybrid GA to 0.051 after first round of direct position search. It was further reduced to 0.02 after 10 rounds, as shown in Fig. 3.8 (a). Although only 9 positions in each cell could be chosen for a co-vertex, the direct search showed a good performance. The reconstructed image with the normalized RMS error of 2.03% is shown in Fig. 3.8 (b). The random noise level in Fig. 3.8 (b) was reduced significantly, compared with that shown in Fig. 3.7 (b).



(a)

(b)

Fig. 3.8 Results by Direct Search for optimal co-vertex positions. (a) Normalized RMS errors after each round of the search, which converges to 2%; (b) Grayscale image of keyboard reconstructed from the designed CGH.

The designed binary CGH to give the reconstructed image in Fig. 3.8 (a) is shown in Fig. 3.9 (a) and an enlarged part of the CGH is shown in Fig. 3.9 (b), in which we can see clearly the irregular polygonal apertures formed by the optimized binary valued triangles. As the design used the Abbe transform to compute the reconstructed image, and the resulting Fig. 3.8(b) was computed with the Abbe transform.

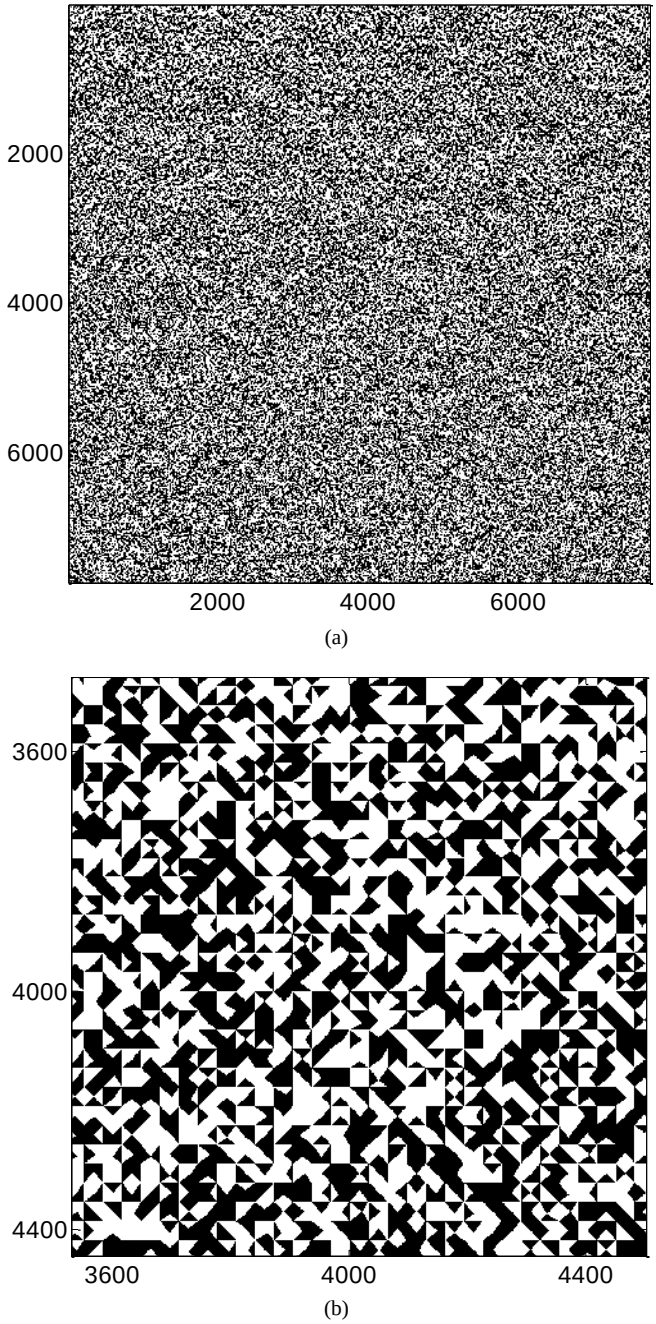


Fig. 3.9 (a) Designed binary CGH; (b) 8X enlarged part of CGH

For the experimental validation, the reconstructed image was computed with the fast Fourier transform (FFT) of the binary CGH of 7744 x7744 pixel size, but not with the Abbe transform. This FFT simulation is widely accepted as reliably close to optical demonstration. A central part of the 8X enlarged image is shown in Fig. 3.10, where the noise was dispersed to the surrounding of the object window. The error in the reconstructed image computed by the FFT was 15.12%. The main source of the error in the FFT results might come from the fact that the aperture edges, which are not parallel to horizontal and vertical axes, were discretized to the stepwise forms in the FFT, while they were all straight lines in the Abbe transform. This error may diminish with the fabrication process, which could smooth out the subwavelength stepwise features in the CGH.

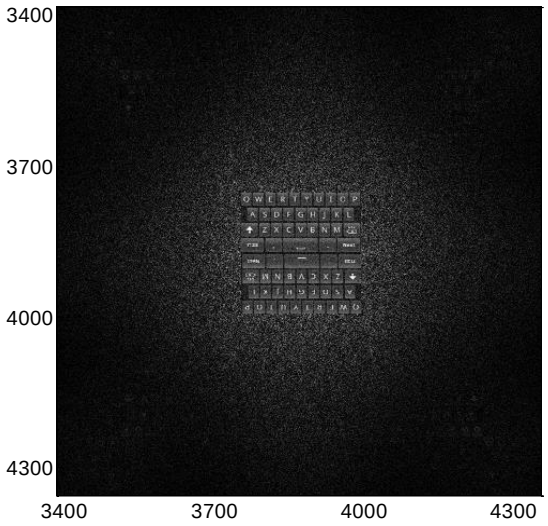


Fig. 3.10 8X enlarged part of the image reconstructed by FFT from the designed CGH.

It took 50 hours for the total design and simulation, based on a Matlab script by a desk computer with a processor of intel (R) Core (TM) i7-4770 CPU operated on a 64-bit Window 7 system. The hybrid GA took 32 hours for the 300 generations. The direct search took 18 hours for the 10 rounds. This computation cost for design a static CGH was acceptable, compared with its fabrication cost.

3.8 Conclusion

The static computer-generated hologram (CGH) has a very high space bandwidth product (SBWP), which can be explored for the high performance of the CGH. For the binary CGH the arbitrary-shaped polygonal apertures were optimized with the hybrid genetic algorithm for the binary phase values of the apertures and the direct search algorithm for the shapes of the apertures. The Abbe transform permits to express the diffraction pattern of a binary CGH of polygonal apertures with the analytic expressions, which can be computed independently of the size of the CGH. The reconstructed image can be expressed as a coherence addition of the diffraction patterns from all the straight edges of different orientations and lengths. The proposed design strategy permitted handling the binary CGH of very high SBWP and resulting in high performance with low reconstruction error of 2% for a grayscale image. The nanometer scale resolution of the e-beam writing has been used for providing the very high number of degrees of freedom in the CGH design for the high performance.

Due to the limitation of the computation time, the present design example showed a CGH of a physical size of 0.4 x 0.4 mm only. For a realistic CGH of 8 x 8 mm size (1/3 inch) in the laser projectors, a binary CGH of 154,800 x 154,800 pixels would be computed for a reconstructed image of 5,120 x 5,120 pixel size. The research is still under ways for designing large size binary CGH using the parallel computation technology on the supercomputers.

Chapter 4 Design quadrilateral apertures in binary computer-generated holograms of large space bandwidth product

4.1 Résumé

Une nouvelle approche pour concevoir l'hologramme binaire généré par ordinateur d'un très grand nombre de pixels est proposée. La diffraction des ouvertures HGO est calculée par la transformation analytique d'Abbe et en considérant les bords de l'ouverture comme les éléments de diffraction de base. Le coût de calcul est indépendant de la taille de HGO. Les ouvertures polygonales de forme arbitraire dans le CGH se composent d'ouvertures quadrilatérales, qui sont conçus en assignant les phases binaires en utilisant l'algorithme génétique parallèle avec une recherche locale, suivi de l'optimisation des emplacements des co-sommets avec une recherche directe. La conception se traduit par des performances élevées avec une faible erreur de reconstruction d'image.

4.2 Abstract

A new approach for designing the binary computer-generated hologram of a very large number of pixels is proposed. Diffraction of the CGH apertures is computed by the analytical Abbe transform and by considering the aperture edges as the basic diffracting elements. The computation cost is independent of the CGH size. The arbitrary-shaped polygonal apertures in the CGH consist of quadrilateral apertures, which are designed by assigning the binary phases using the parallel genetic algorithm with a local search, followed by optimizing the locations of the co-vertices with a direct search. The design results in high performance with low image reconstruction error.

4.3 Introduction

Computer-generated hologram (CGH) has gained great development and success since its invention by Lohmann and Paris in 1967 [8] for the applications in wavefront engineering, structured illumination, 3D display, immersion entertainment etc. The dynamic CGH encoded on the spatial light modulators (SLM) can have a resolution of 1920 x 1080 pixel, as that for the high definition television (HDTV) images. However, such a space bandwidth product (SBWP) is still too low for encoding a CGH. The research in the design of dynamic CGH mainly focuses on the fast synthesizing of CGHs in real time. On the other hand, the static CGH can have a very high SBWP, thanks to the advanced fabrication technology. For instance, with the industrial e-beam lithography, the minimum feature size of the e-beam writing can reach to 50 *nm*, and a static CGH of 1-inch square (25.4 x 25.4 mm^2) size can be written within 6-8 hours. With a pixel size of 50 *nm*, a small CGH of 1-inch square size can have 508,000 x 508,000 = 258 gigapixels. This SBWP can be still 6.25 times higher when using the e-beam of 20 *nm* resolution. The challenge in the design of static CGH is then how to manage such a large number of degrees of freedom to achieve the best performance.

In the practical holography industry, the static holograms can have a size much larger than 25.4 x 25.4 mm^2. The very high SBWP is beneficial for the techniques such as the dot matrix hologram [104] and the holographic print [105], which are widely used in the security, anti-counterfeiting applications and in the 3D display. The holograms consist, respectively, of a large 2D array of local diffraction gratings with different grating orientations, pitches and depth, or of "hogels", which are the spatially and spectrally discretized computed local fringe patterns. A wide range of visual effects can be achieved by controlling the angles, exposure, size, shape, and spacing of the local gratings, or hogles. In the dot matrix hologram and the holographic print, one local grating, or one hogel generates one image spot of a given intensity. However, the discrete segmentation in the hologram and/or object domain causes blurring and loss of fidelity in the image.

In this chapter, we focused on the binary CGH, which is easy to fabricate and duplicate on the substrate. For the binary CGH with high SBWP, the conventional Gerchberg-Saxton iterative algorithm [43] becomes inefficient since the fast Fourier transform (FFT) is slow for the CGH with a huge number of pixels. Moreover, the hard banalization leads to high errors and stagnation of the iterations. The direct binary search (DBS), which invert the sign

of each bit scanned, and thus retains the inversion if the error reduced, or rejects it otherwise, can generate acceptable performance. It can be accelerated by computing the image reconstruction error with analytical expressions directly without performing FFT after each inversion of the bit [38]. However, these methods become unpractical due to the huge computational burden for CGH with high SBWP.

A direct design of the polygonal apertures of arbitrary shapes in the binary CGH was demonstrated. The Abbe transform [106] was used to compute the diffraction of a polygonal aperture of arbitrary shape, as a coherent addition of the diffracted fields from its edges with analytical expressions. The computation time of the analytical Abbe transform is independent of the size of the CGHs. In this paper, we design the CGH with binary phases and arbitrary-shaped apertures composed of elementary quadrilateral apertures with floating vertices. The quadrilateral apertures layout provides more degrees of freedom for optimizing the aperture shapes than that in the triangle aperture layout presented in Chapter 3. In the former layout up to 12 edges in a cell can be modified, while in the latter layout, only the co-vertices of the 4 triangles is floating and only 4 edges can be modified in a cell.

There is recently a large body of research works on the so-called polygon-based methods for synthesizing the CGH in the holographic 3D display applications [108, 109]. This technique models the 3D object surface by plural triangular mesh facets, as that used in computer graphics, and computes the angular spectrum of the elementary facet, which is the 2D facet but at an arbitrary orientation in the 3D space, to synthesize the CGH. Thus, the diffraction is computed for the triangular facets on the object. However, the CGH of polygonal aperture in this paper is designed by computing the diffraction of the elementary arbitrary quadrilateral apertures on the CGH. In this paper, the CGH with arbitrary polygonal apertures are optimized with a two-step optimization algorithm, which is the parallel Genetic Algorithm (GA) with local search for encoding the binary phases of the apertures, followed by the direct search for optimizing positions of the floating vertices of the apertures. An example design for a binary CGH of 8192 x 8192 pixel size reconstructing a grayscale level image with the reconstruction error of 3.8% is shown.

4.4 Diffractions of binary CGH

In the proposed CGH layout one period of the CGH is divided into M x N rectangular cells of the same size of P x Q pixels. For a cell (j, r) with the 4 corners at A(j/M, r/N), B((j+1)/M, r/N), C((j+1)/M, (r+1)/N), and D(j/M, (r+1)/N), there are 4 vertices G_1, G_2, G_3, G_4 floating along the corresponding straight edges. The 4 vertices with a co-vertex G_0 floating in the internal of the cell divide the cell into 4 quadrilateral apertures, indexed by s = 1-4, as shown in Fig. 4.1(a). In the local coordinate system (x, y) with the origin at the corner A of the cell (j,r), the coordinates of the vertices are $G_1(x'(j, r)/P, 0)$, $G_2(1, y'(j, r)/Q)$, $G_3(x''(j, r)/P, 1)$, $G_4(0, y''(j, r)/Q)$ and $G_0(x_c(j, r)/P, y_c(j, r)/Q)$, as shown in Fig. 4.1(b).

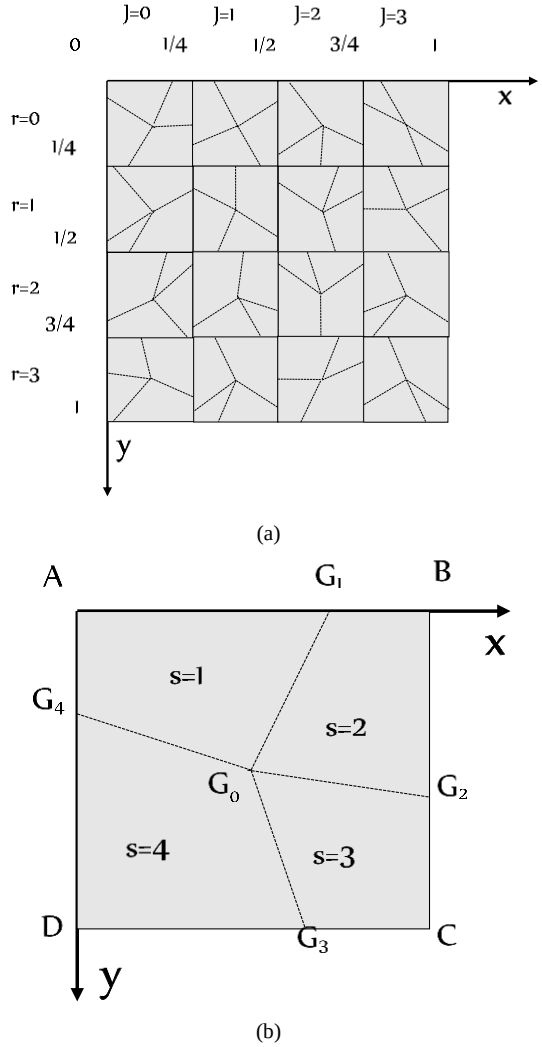


Fig. 4.1 CGH layout, (a) One period of CGH with rectangular cells divided to four quadrilateral apertures; (b) arbitrary-shaped quadrilateral in a cell.

In Chapter 3, the CGH's polygonal apertures of arbitrary shapes were composed of elementary triangular apertures. The cell was divided into 4 triangles whose shapes were optimized by moving the central co-vertex G_0 with the 4 vertices G_1, G_2, G_3, G_4 fixed at the corners of the cell. The new layout, proposed in this Chapter, provides a greater number of degrees of freedom, as the 4 vertices of the 4 quadrilateral apertures can move independently along the respective edges of the cell. In the triangle aperture layout, only the co-vertices G_0 in the cell is floating, while in the quadrilateral aperture layout, up to 12 edges could be modified by moving the G_0, G_1, G_2, G_3, G_4 respectively. Arbitrary shapes of the apertures could be achieved by moving the co-vertices of the quadrilateral apertures.

The Fraunhofer diffraction of the quadrilateral apertures can be computed by the Abbe transform, which reduces the surface integral over an aperture to a line integral along the closed boundary of the aperture. In the case of polygonal aperture, the line integral along the boundary becomes an algebraic summation as

$$U(\vec{p}) = \frac{i}{2\pi p^2} \sum_{q=1}^{Q} \vec{p} \cdot \hat{n}_q L_q \sin c\left(\vec{p} \cdot \hat{t}_q L_q\right) \exp\left(-i2\pi \vec{p} \cdot \hat{\xi}_{M_q}\right) \tag{4.1}$$

where Q is the number of the edges in the polygon, $\vec{\xi}(\xi_1, \xi_2)$ and $\vec{p}(m,n)$ denote the position vector in the aperture plane and the Fraunhofer diffraction plane, respectively, and $\vec{\xi}_{M_q}$ denotes the midpoint of the *q*-th edge, which has the length L_q, the unit external normal vector $\hat{n}_q$ and the unit tangent vector $\hat{t}_q = \vec{L}_q / L_q$. Eq. (4.1) is deduced in Chapter 3. Thus, the complex-valued diffraction pattern of a quadrilateral aperture of index *s* in the cell (*j*, *r*) with a phase value $\varphi_s(j,r)$ as shown in Fig. 4.1(b), can be written as

$$U_s(m,n,j,r) = \left\{\Theta_s(m,n,j,r) \exp\left[-i2\pi\left(\frac{m}{M} j + \frac{n}{N} r\right)\right]\right\} \exp\left[i\varphi_s(j,r)\right] \tag{4.2}$$

where $\Theta_s(m,n,j,r)$ is the diffraction pattern of the aperture *s* in the cell (*j*, *r*) as computed from Eq. (4.1). With the vertices defined in the local coordinate system in a given cell (*j*, *r*), $\Theta_s(m,n,j,r)$ depends only on the shape of the aperture and the local coordinates of its vertices.

Thus, the diffraction pattern of the cell is a summation of U_s with s = 1 – 4. The complex amplitude of the diffraction pattern of the entire CGH is a coherent summation of the diffracted patterns of all the cells for j = 0, 1, 2, … M-1 and r = 0, 1, 2, … N-1. For the 4 quadrilateral apertures in the cell, $\Theta_s(m,n,j,r)$ can be expressed as

$$\begin{aligned}
\Theta_1 &= \frac{i}{2\pi\left(m^2+n^2\right)}\left(\bar{H}_1+\bar{H}_9-\bar{H}_{12}+\bar{H}_8\right)\\
\Theta_2 &= \frac{i}{2\pi\left(m^2+n^2\right)}\left(\bar{H}_3+\bar{H}_{10}-\bar{H}_9+\bar{H}_2\right)\\
\Theta_3 &= \frac{i}{2\pi\left(m^2+n^2\right)}\left(\bar{H}_5+\bar{H}_{11}-\bar{H}_{10}+\bar{H}_4\right)\\
\Theta_4 &= \frac{i}{2\pi\left(m^2+n^2\right)}\left(\bar{H}_7+\bar{H}_{12}-\bar{H}_{11}+\bar{H}_6\right)
\end{aligned} \tag{4.3}$$

where $\bar{H}_1 \sim \bar{H}_{12}$ denote the line integrals along the cell edges AG_1, G_1B, BG_2, G_2C, CG_3, G_3D, DG_4, G_4A, and the lines within the cell G_1G_0, G_2G_0, G_3G_0 and G_4G_0, respectively. For instance, Θ_1 are computed with

$$\begin{aligned}
\bar{H}_1(m,n,j,r) &= \left[-\frac{x'(j,r)}{P}\frac{n}{M}\right]\cdot\sin c\left[\frac{x'(j,r)}{P}\frac{m}{M}\right]\cdot\exp\left[-i\pi\frac{x'(j,r)}{P}\frac{m}{M}\right]\\
\bar{H}_8(m,n,j,r) &= \left[-\frac{y''(j,r)}{Q}\frac{m}{N}\right]\cdot\sin c\left[-\frac{y''(j,r)}{Q}\frac{n}{N}\right]\cdot\exp\left[-i\pi\frac{y''(j,r)}{Q}\frac{n}{N}\right]\\
\bar{H}_9(m,n,j,r) &= \left[\frac{y_c(j,r)}{Q}\frac{m}{N}+\frac{x'(j,r)-x_c(j,r)}{P}\frac{n}{M}\right]\cdot\sin c\left[\frac{x_c(j,r)-x'(j,r)}{P}\frac{m}{M}+\frac{y_c(j,r)}{Q}\frac{n}{N}\right]\\
&\quad\cdot\exp\left[-i\pi\left(\frac{x_c(j,r)+x'(j,r)}{P}\frac{m}{M}+\frac{y_c(j,r)}{Q}\frac{n}{N}\right)\right]\\
\bar{H}_{12}(m,n,j,r) &= \left[\frac{y_c(j,r)-y''(j,r)}{Q}\frac{m}{N}+\left(-\frac{x_c(j,r)}{P}\right)\frac{n}{M}\right]\cdot\sin c\left[\frac{x_c(j,r)}{P}\frac{m}{M}+\frac{y_c(j,r)-y''(j,r)}{Q}\frac{n}{N}\right]\\
&\quad\cdot\exp\left[-i\pi\left(\frac{x_c(j,r)}{P}\frac{m}{M}+\frac{y_c(j,r)+y''(j,r)}{Q}\frac{n}{N}\right)\right]
\end{aligned} \tag{4.4}$$

The line integrals on the apertures' boundaries are all in the clockwise direction, so that when the two neighboring apertures are encoded with the same binary phase value and are therefore merged into a larger polygonal aperture, the contributions of their shared edge to the diffraction pattern will be computed twice on the same segment, but in the opposite directions and will be canceled. For instance, if the apertures s=1 and s=2 in Fig. 4.1(b) are assigned to

same binary phase value, and therefore merged to the polygonal aperture $ABG_2G_0G_4$, the diffraction from the shared edge G_1G_0 actually vanishes because of the line integrals $+\bar{H}_9$ in Θ_1 and $-\bar{H}_9$ in Θ_2 in the summation for s = 1 – 4, as shown Eq. (4.3).

According to the Abbe transform, the basic physical elements for the diffraction in the binary CGHs are the straight edges of the apertures. As described by Sommerfeld, a straight edge diffracts as an infinitely narrow slit, which sends a light fan in the direction perpendicular to the edge with the fan width inversely proportional to the edge length. The diffraction pattern of a binary CGH is a coherent summation of that from all the edges of all possible locations, orientations and lengths.

If the locations of points G1, G2, G3, G4, G0 are the same in each cell, all the quadrilateral apertures with the same s in all the cells have the same shape and area and the same $\Theta_s(m,n,j,r)$, independently on the location of the cell (j, r), so that the diffraction pattern of the CGH can be computed as

$$\begin{aligned} T(m,n) &= \sum_{s=1}^{s=4}\sum_{j=0}^{j=M-1}\sum_{r=0}^{r=N-1}\left\{\Theta_s(m,n,j,r)\cdot\exp\left[-i2\pi\left(\frac{m}{M}j+\frac{n}{N}r\right)\right]\cdot\exp\left[i\varphi_s(j,r)\right]\right\} \\ &= \sum_{s=1}^{s=4}\left\{\Theta_s(m,n)\left\{\sum_{j=0}^{j=M-1}\sum_{r=0}^{r=N-1}\left\{\exp\left[-i2\pi\left(\frac{m}{M}j+\frac{n}{N}r\right)\right]\cdot\exp\left[i\varphi_s(j,r)\right]\right\}\right\}\right\} \\ &= \sum_{s=1}^{s=4}\left\{\Theta_s(m,n)\cdot\mathrm{DFT}\left\{\exp\left[i\varphi_s(j,r)\right]\right\}\right\} \end{aligned} \qquad (4.5)$$

where DFT denotes the Discrete Fourier Transform.

4.5 Design algorithms

To take advantage of the high number of degrees of freedom, the proposed design for the binary CGH is performed in two steps. The first step is to assign the binary phase values to the quadrilateral apertures, which is implemented by the coarse-grained parallel GA with local search [110]. The second step is to optimize the positions of the points G1, G2, G3, G4 along the boundaries of the cells which is achieved by the direct search algorithm.

4.5.1 Coarse-grained parallel GA with a local search for optimizing binary phases

The CGH consists of M x N rectangular cells with each cell of P x Q pixels divided into 4 quadrilateral apertures. The initial locations of the vertices on the boundaries of the cell are the same for all the cells in this step. In the local coordinate system (x, y) with the origin at the corner A, the initial setting for $G_1(x'/P, 0)$, $G_2(1, y'/Q)$, $G_3(x''/P, 1)$, $G_4(0, y''/Q)$ and $G_0(x_c/P, y_c/Q)$ was x'=P/4, y'=Q/4, x''=3P/4, y''=3Q/4, $x_c=P/2$, $y_c=Q/2$. All the quadrilateral apertures with the same s in all the cells have the same shape and area. By using Eq. (4.5) to compute the diffraction pattern, the size of the reconstructed image is limited to be equal to the number of the cells M x N in the CGH.

Parallel genetic algorithm (PGA) with a local search is used to assign binary phase to each aperture. Parallel genetic algorithm (PGA) can be classified into three different models: master-slave PGA, coarse-grained PGA and fine-grained PGA. Coarse-grained PGA is also called distributed model or island-based model, which divides the whole population into several subpopulations to be computed on separate processors. Each processor executes GA on its own subpopulation, meanwhile, the processors would occasionally exchange chromosomes with a certain probability so that the good genes could spread to other subpopulations.

The procedures using coarse-grained parallel genetic algorithm to assign the binary phases are outlined below:

1. Initialization

The binary phase values for all the 4 x M x N quadrilateral apertures of a CGH are encoded as an individual chromosome, which is a string of 4 x M x N bits. An initial subpopulation of N_{sub}+1 chromosomes is generated randomly for each processor with a uniform probability distribution. The extra one chromosome, as a transit chromosome for replacing the worst one by the best one in each generation, will be discarded later.

2. Evaluation

The N_{sub}+1 chromosomes in each subpopulation are evaluated by the cost function, e.g. mean squared errors of the image reconstructed from the CGH presented by corresponding chromosome, which defined as

$$e = \sqrt{\sum_{m=-M/2}^{M/2} \sum_{n=-N/2}^{N/2} \left(\left| f(m,n) \right|^2 - \left| f_t(m,n) \right|^2 \right)^2} \tag{4.6}$$

where the integers *m* and *n* define the locations of pixels, $f(m, n)$ and $f_t(m, n)$ are the amplitude of the reconstructed image and the target image, respectively. The term of diffraction efficiency is not included in Eq. (4.6) for keeping the computational cost low in the optimization steps by computing the reconstructed image only within the range $-M/2 \le m \le M/2$ and $-N/2 \le n \le N/2$ instead of in the window of full size. Since in this step, the locations of points G_1, G_2, G_3, G_4, G_0 are the same in each cell, the image reconstructed from the CGH could be obtained by Eq. (4.5).

3. Ranking & Elitism

The N_{sub}+1 chromosomes in each subpopulation are ranked according to their ascending errors. The best individuals with the least errors are ranked at the top and the worst ones are ranked at the bottom. In this step, the worst one will be discarded, meanwhile, the best one will be kept as the elite, and entered directly to the pool for local search later without any crossover and mutation. The elitism thus prevents the best individual in the subpopulation from deterioration.

4. Selection

A stochastic universal sampling selection is performed in each subpopulation to produce a mating pool, based on the fitness values for the chromosomes. The fitness is defined for the chromosome of rank *p* as

$$\phi(p) = \frac{N_{sub} - p}{N_{sub} - 1} \tag{4.7}$$

The normalized fitness values are then

$$\Phi(p) = \frac{\phi(p)}{\sum_{p'=1}^{N_{sub}} \phi(p')} \tag{4.8}$$

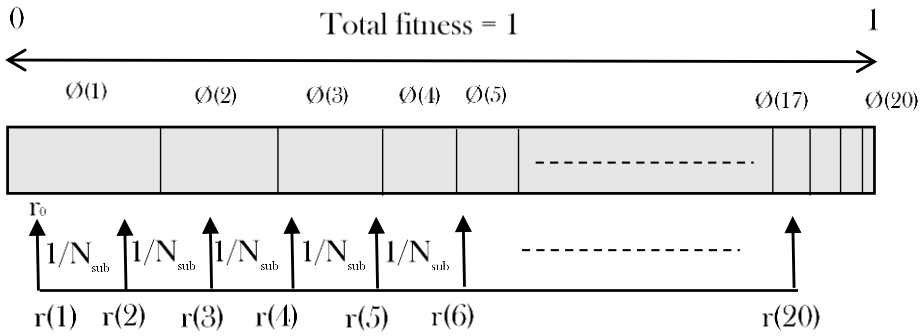


Fig. 4.2 Illustration of stochastic universal sampling selection.

In the selection process for mating, a random number $r_0 \in (0, 1/N_{sub})$ is generated, then we define a number *r*

$$r(N_s) = r_0 + \frac{N_s - 1}{N_{sub}} \tag{4.9}$$

where $N_s = 1, 2, \ldots N_{sub}$ is the index of the selection for mating, as shown in Fig. 4.2. In each selection N_s, if $C_\Phi(p-1) < r(N_s) \le C_\Phi(p)$, where $C_\Phi(p)$ is the cumulative probability of $\Phi(p)$, then the chromosome at rank *p* is selected. With the stochastic universal sampling selection, the better individuals have a higher probability to survive, while the worse ones would disappear eventually. Especially, the chromosome at top rank *p*=1 is always selected for mating with N_1=1. While in the conventional roulette wheel selection, it is probable that the best chromosomes *p* = 1, 2 could be not selected.

5. Crossover

The genes of the randomly paired chromosomes after the selection are exchanged so that offsprings could be produced. The exchanges of genes are implemented by a uniform crossover, also known as scattering crossover, enabling the parent chromosomes to contribute on the gene level rather than on the segment level as that done in conventional single/double point crossover.

For each pair of chromosomes, a random binary crossover mask of the same length is generated first. Where there is a 1 in the mask, the gene of the 1st offspring is copied from

the 1st parent, and the gene of the 2nd offspring is copied from the 2nd parent. If there is a 0 in the mask, the gene from the 2nd parent is assigned to the 1st offspring, and the gene from the 1st parent is assigned to the 2nd offspring, as shown in Fig. 4.3. The offspring has approximately half of the genes from each parent, although cross over points are randomly chosen. Uniform crossover usually occurs with a given high crossover probability $Pc > 0.6$.

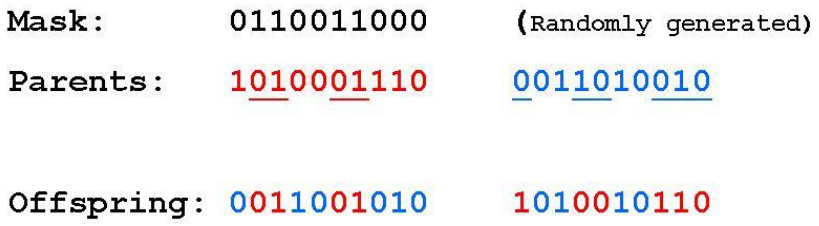


Fig. 4.3 Illustration of uniform crossover with a random binary mask.

6. Mutation

After the crossover, the mutation for maintaining the genetic diversity of the genes is performed, which flips each bit in each chromosome with an exponentially decreasing mutation probability P_m defined as $P_m(k+1) = \alpha P_m(k) = \alpha^k P_m(1)$ with $\alpha \leq 1$. Thus $P_m(k)$ decreases with the generation number k. For every bit in the chromosomes a random number R ranged in (0, 1) is generated. If $R < P_m(k)$, then the bit is flipped. Otherwise it is not flipped. At the begin of the mutation, a relative large number of bits are inversed, while close to the end of the mutation, only a few number of bits are flipped as $P_m(k)$ decreases with k.

7. Local search

Introduction of the local search within a GA can significantly improve the GA convergence. Before sending the subpopulation to the next generation, the direct binary search (DBS) is performed as the local search and applied to a randomly chosen chromosome from the subpopulation by scanning and flipping the binary phase value of every bit in the chromosome. The flip, which reduces the error, is retained. Otherwise, the flip is rejected. The new diffraction pattern T'(m,n) after flipping phase value of quadrilateral apertures s in cell (j,r) can be computed as

$$T'(m,n) = T(m,n) - U_s(m,n,j,r) + U'_s(m,n,j,r) \quad (4.10)$$

where U_s and U'_s is the diffraction pattern of aperture s in cell (j,r) before and after flipping the binary phase value, respectively. The GA with local search for each subpopulation is iterated on each processor till the convergence is achieved.

8. Migration

Basically, each subpopulation evolves independently following the steps above on separating computing processor, which would result in the low diversity of the population since there is no genes exchanging between different subpopulations. Introduction of migration operation in coarse grained PGA could increase the diversity by exchanging the chromosomes between different processors. There are two key parameters: migration rate, which indicates the execution frequency of the migration, and migration number, which determine the chromosome number for migration. The migration topology used is the one-way ring structure as shown in Fig. 4.4, which indicates that, for instance, on one hand, the best chromosomes in subpopulation SP2 is chosen to migrate to replace the worst chromosome in subpopulation SP3, on another hand, the worst chromosome in subpopulation SP2 is taken the place by the best chromosome migrated from subpopulation SP1.

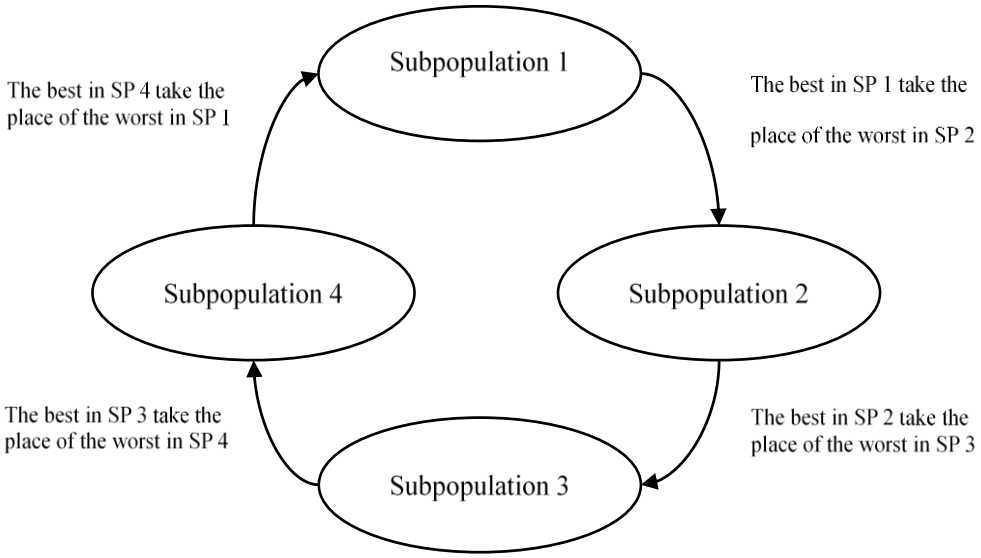


Fig. 4.4 Migration topology for 4 subpopulations: one-way ring structure.

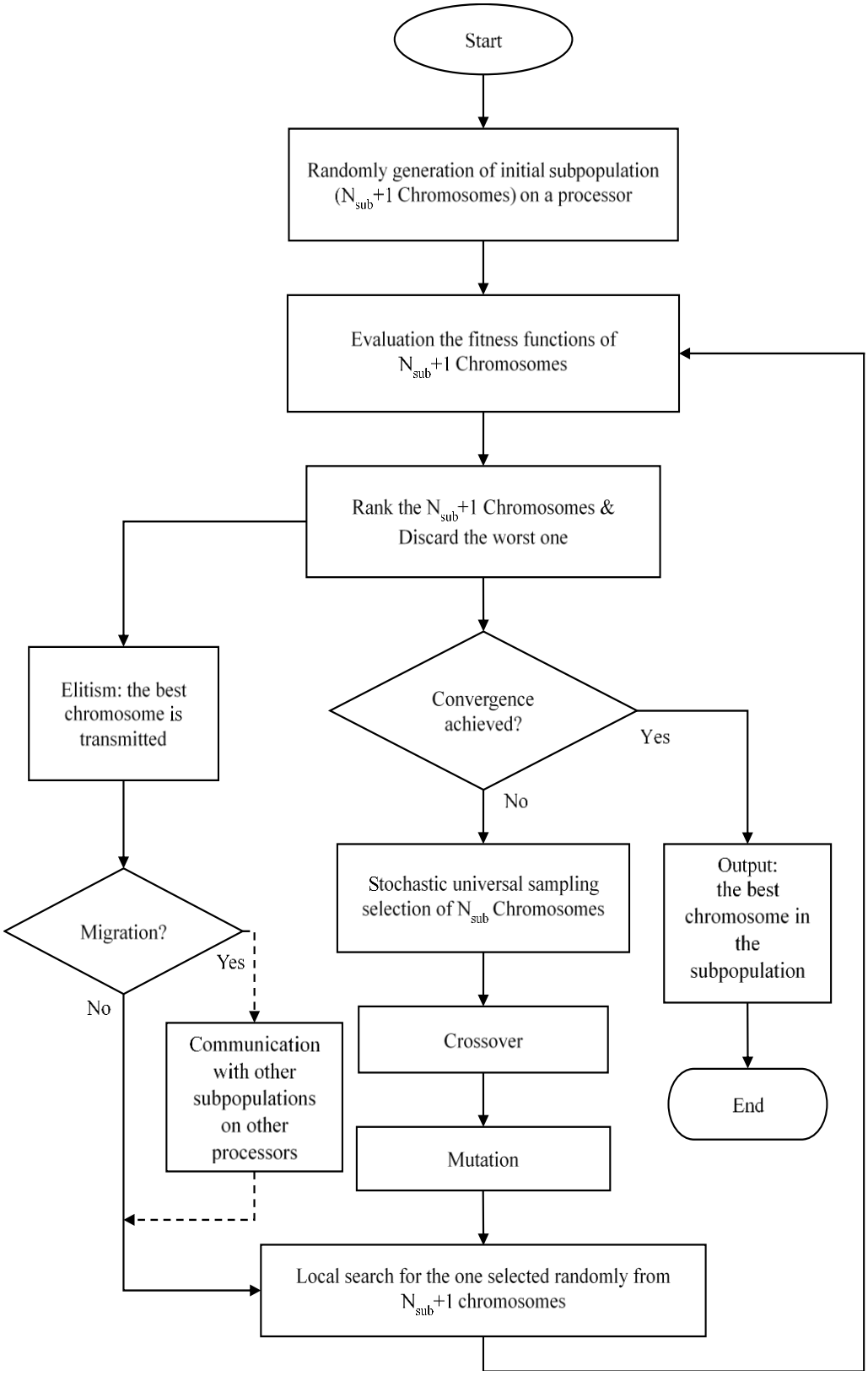


Fig. 4.5 Flowchart of the PGA with local search on each processor for designing CGH of polygonal apertures.

9. Iteration and output

The coarse-grained parallel GA with local search is executed in an iteration till the solution achieve to convergence. In each subpopulation, a best chromosome is obtained on each processor as the output. The best chromosome in the whole population is the best one among all the output chromosomes from all the processors. Fig. 4.5 shows the flowchart of the coarse-grained parallel GA with local search on each processor for designing CGHs.

4.5.2 Direct search for optimal aperture shapes

After optimization of the binary phase values for all the quadrilateral apertures by parallel GA with local search, the best individual in the whole population is selected for further optimization. Unlike in the first step of design, where the 4 quadrilateral apertures in all the cells are identical, in the second design step, the shape of the polygonal apertures are optimized by moving the positions of the vertices G_1, G_2, G_3, G_4 in the cells along their respective boundary using the direct search method, which results in immediate changes of the lengths of the edges on the boundary and the orientations and lengths of the straight edges in the internal of the cells, and thus, the shape of all the quadrilateral apertures. The move of the co-vertices creates additional degrees of freedom and can improve the performance of the binary CGH efficiently. As according to the Abbe transform the diffraction pattern of a binary CGH is the coherent summation of all the diffraction patterns of all the edges of different orientations and lengths in the CGH.

In the direct search point G_1 first moves along the cell boundary AB, and then point G_2, G_3, G_4 moves along the corresponding cell boundary BC, CD, DA in sequence. One move of the vertex G_1 changes the shapes of both quadrilateral apertures, s_1 and s_2, the whole diffraction pattern could then be computed as

$$T'(m,n) = T(m,n) - \sum_{s=1}^{s=2} U_s(m,n,j,r) + \sum_{s=1}^{s=2} U'_s(m,n,j,r) \qquad (4.11)$$

where U_s and U'_s are the diffraction patterns of the quadrilateral apertures (j, r, s) before and after moving of the vertex G_1. The diffraction patterns after the moves of vertices G_2, G_3 and G_4 are computed similarly. One move of vertices G_2, G_3 and G_4 gives, respectively, the whole diffraction pattern

$$T'(m,n) = T(m,n) - \sum_{s=2}^{s=3} U_s(m,n,j,r) + \sum_{s=2}^{s=3} U'_s(m,n,j,r) \quad (4.12)$$

$$T'(m,n) = T(m,n) - \sum_{s=3}^{s=4} U_s(m,n,j,r) + \sum_{s=3}^{s=4} U'_s(m,n,j,r) \quad (4.13)$$

$$T'(m,n) = T(m,n) - \sum_{s=4}^{s=1} U_s(m,n,j,r) + \sum_{s=4}^{s=1} U'_s(m,n,j,r) \quad (4.14)$$

For the sake of computation time, the moves of G_1, G_2, G_3, G_4 are stepwise with given step sizes. The moves which reduce the cost function are retained, and otherwise rejected. The direct search scans the CGH cell by cell in several rounds. The G_0 located at the center of the cell can also be floating for further optimization. However, the moves of G_0 are restrained to prevent it from being too close to the cell boundary and resulting in too small quadrilateral apertures.

4.6 Experimental design

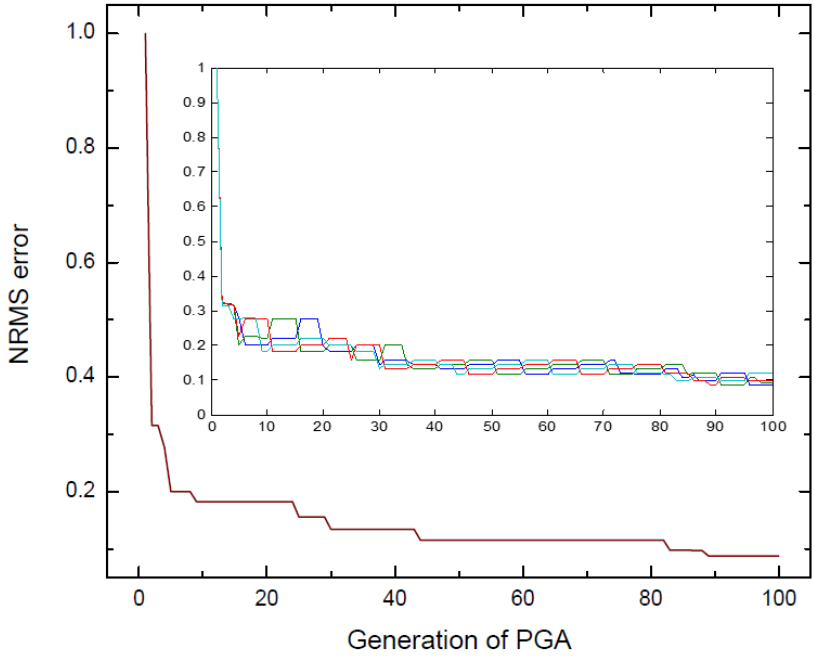
The proposed approach was adopted to synthesize binary CGHs of arbitrary-shaped polygonal apertures. A grayscale LOGO image of "International year of light 2015" with 256 x 256 pixels was used as the target image for reconstruction. The number of the cells per period in the CGH was set as $M = N = 256$. Each cell was chosen to have a size of PxQ = 32x32 pixels. Assume a pixel size of 50 *nm* by the e-beam lithography, the cells of 32x32 pixels have a physical size of 1.6 x 1.6 μm, which is larger than the illuminating wavelength. Thus the CGH was of 8192 x 8192 pixels with the physical dimension of 0.41 x 0.41 *mm* for one period. The design was performed on a desk computer with Intel (R) Core (TM) i7-4770 CPU, which provides 4 Cores facilitating for the implementation of the parallel GA.

4.6.1 PGA with a local search for assigning binary phases

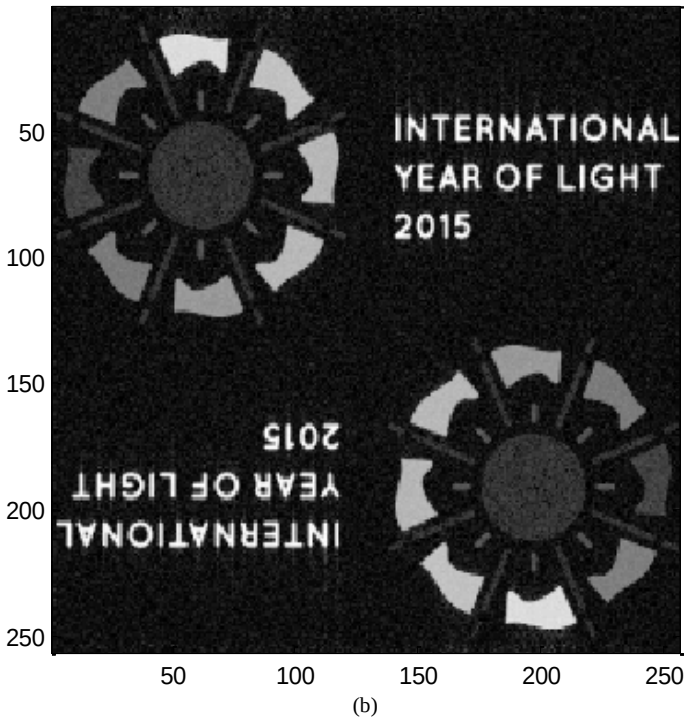
In the first step, parallel GA with local search was used for assigning binary phases to each quadrilateral aperture. A randomly generated population of 80 chromosomes was divided into 4 subpopulations operated on the 4 different Cores. The initial subpopulation of 20+1 chromosomes was generated at random for each Core. The chromosome was a string of $4MN$ = 262,144 bit long. After evaluating the image reconstruction errors and discarding the worst solution, the 20 chromosomes were selected with a stochastic universal sampling method for reproduction of next generation. The selection process for mating was based on the rank of the chromosomes with the linear fitness value given by Eq. (4.8). The stochastic universal sampling was run 20 times for mating 10 pairs of parent chromosomes in each subpopulation. Each pair of the parent chromosomes generated 2 descendants with the uniform crossover, in which the crossover probability was $P_c = 0.8$. Then, the mutation with the exponentially decreasing probability $P_m(k+1) = \alpha P_m(k) = \alpha^k P_m(1)$ was performed with $P_m(1) = 0.1$ and $\alpha=0.97$, where k is the number of generation. After the mutation, 20 new offspring chromosomes were produced in addition to the 1 elite chromosome, which was transmitted directly without crossover and mutation. The migration is executed in every 5 generations. The elite was sent to the subpopulation on a different Core, and meanwhile, an elite from subpopulation on another Core was received, as shown in Fig. 4.4. Then, the local search was performed for one chromosome randomly chosen in the new subpopulation of 20+1 chromosomes. Finally, the new population of 21 chromosomes went to the same process for the next population. In each subpopulation, there was a best chromosome. After evaluation of the 4 best chromosomes on 4 different Cores, the optimized one in the whole population was obtained in each generation.

After 100 generations of parallel GA in the first step, the normalized root mean square (RMS) errors of the best chromosome in the whole population are plotted in Fig. 4.6(a) as a function of the number of generations. The minimum error dropped to 0.32 only after one generation in the parallel GA. This attributes to the local search for one randomly selected chromosome in each subpopulation. Only after 10 generations, the error decreased to 0.18. Furthermore, it reduced to 0.12 after 50 generations and reach to 0.09 after 100 generations, as shown in Fig.4.6 (a). The minimum error of the best chromosome in the whole population decreased monotonically thanks to the elitism. The RMS errors of the best chromosome in each

subpopulation are also plotted in four different colors in the subsection of Fig. 4.6 (a) as a function of the number of generations.



(a)



(b)

Fig. 4.6 Results from the parallel GA with local search for assigning binary phases: (a) Normalized RMS errors as a function of generation; (b) Grayscale image reconstructed by the CGH.

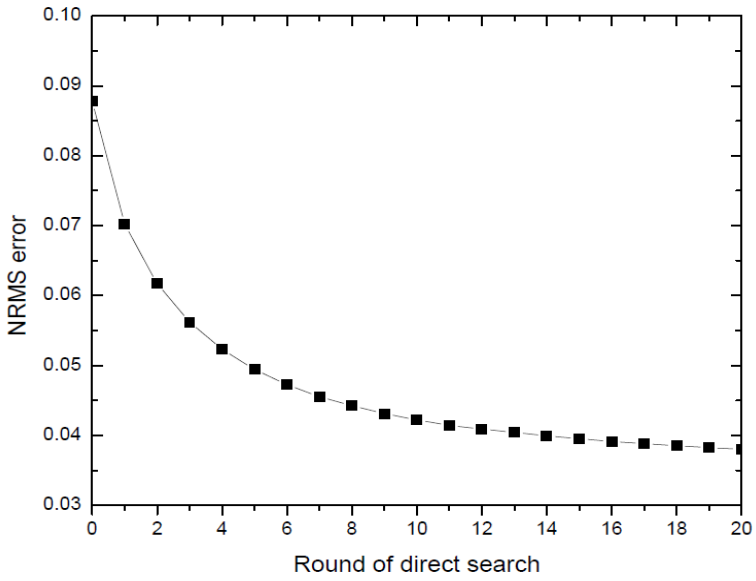
It can be seen that, minimum errors of the best chromosome in each subpopulation didn't decrease monotonically due to the migration of the elites between subpopulations every 5 generations. Although minimum errors in each subpopulation show some fluctuations, the best chromosome in the whole population evolved well. The reconstructed image with the normalized RMS error of 0.09 for the best chromosome in the whole population after 100 generations is shown in Fig. 4.6 (b), in which the noise on the image is still visible.

4.6.2 Direct search for optimized shapes of apertures

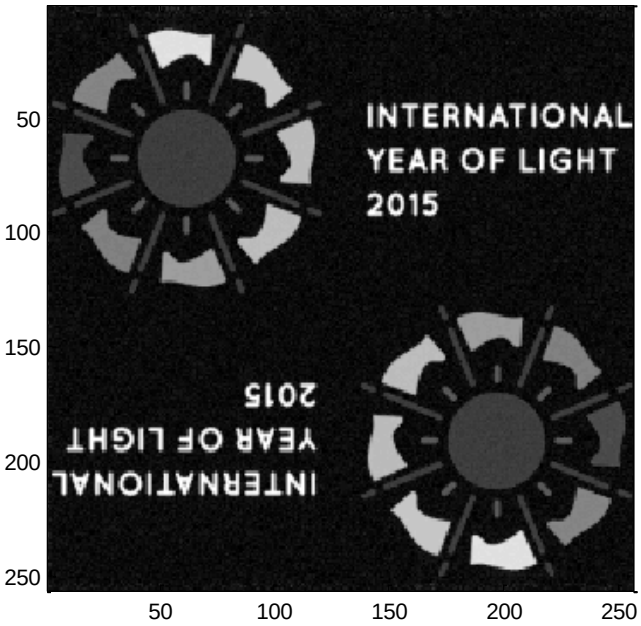
In the second step, further optimization for the best chromosome in the whole population is performed by changing the position of the vertex one-by-one along each straight edge boundary of the cell. This created additional number of degrees of freedom and arbitrary shapes of apertures as well. Considering that the sizes of apertures should be larger than the wavelength in the scope of scalar diffraction theory, the co-vertex G_0 was kept still in the center of the cell, and only 2 possible positions of the vertices were chosen (P/4, 3P/4 for G_1, G_3, or Q/4, 3Q/4 for G_2, G_4) for each point of G_1, G_2, G_3, G_4 along each boundary of the cell. The vertex position which reduces the error was kept, otherwise rejected. The direct search was performed within a cell and then cell-by-cell. The direct search process was run for 20 rounds. The RMS error was decreased from 0.088 by the parallel GA to 0.07 after the first round of direct position search. It was further reduced to 0.038 after 20 rounds, as shown in Fig. 4.7(a). The reconstructed image with the error of 0.038 is shown in Fig. 4.7(b), which was of much better quality than that in Fig. 4.6(b).

A part of the binary CGH designed by the two-step approach to give the reconstructed image in Fig. 4.7(b) is shown in Fig. 4.8, in which the irregular polygonal apertures were clearly presented. The image reconstructed in Fig. 4.6(b) and 4.7(b) were computed with the Abbe transform. Finally, for the validation of the design the fast Fourier transform of the 8192 x 8192 pixels binary CGH was performed, which produced the expected target image. The 8X central part of the reconstructed image by direct FFT was shown in Fig. 4.9, from which we can see the expected target image in the 256 x 256 pixel window with the noise mainly diffused around the object window. The error of the reconstruction image by FFT was

13.63%, which was higher than the designed error value, probably because the apertures edges were considered as straight lines when computing the diffraction in the Abbe transform, while in the FFT the aperture edges were the staircase lines. Note that when this CGH will be fabricated with the e-beam machine, the aperture edges would be in between the straight lines and the clearly cut staircase lines.



(a)



(b)

Fig. 4.7 Results by Direct Search for optimal co-vertex positions. (a) Normalized RMS errors after each round of the search; (b) Grayscale image reconstructed from the designed CGH.

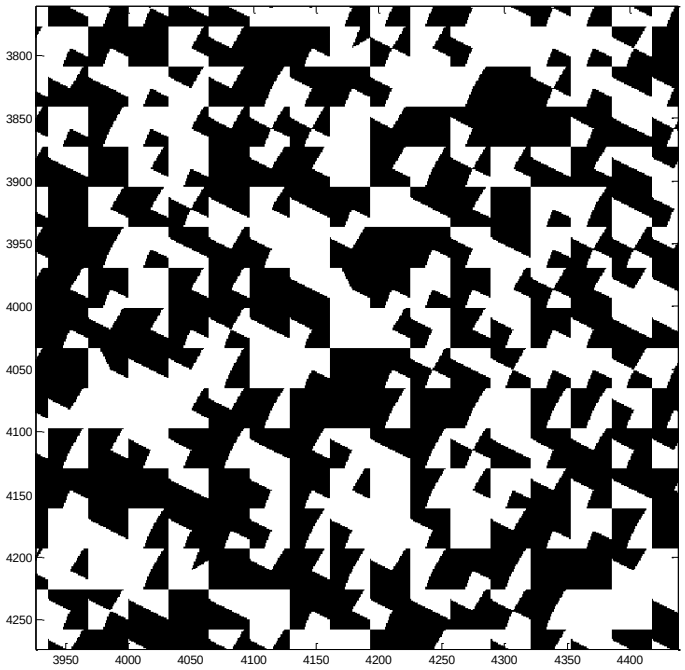


Fig. 4.8 Sixteen times enlarged part of the designed polygonal CGH

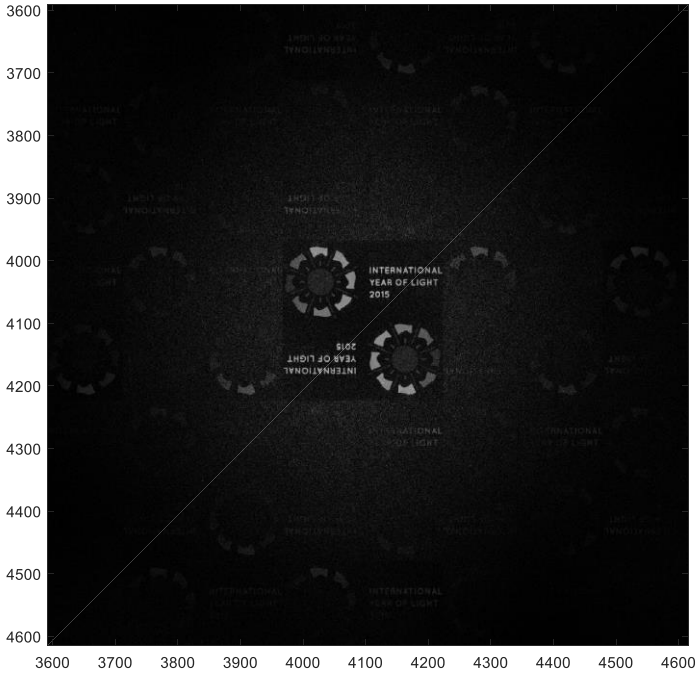


Fig. 4.9 Part of enlarged image reconstructed by FFT from the designed CGH

4.6.3 Computational cost

The parallel GA with local search produced good results in the first step for assigning the binary phase values of the CGH apertures, but at high computational cost. As can be seen from Tab. 4.1, it took 2657 minutes for the parallel GA with 100 generations in the first step, and 983 minutes for direct search of 20 rounds in the second step to achieve the final error of 0.038. This is the Scheme 1 of optimization.

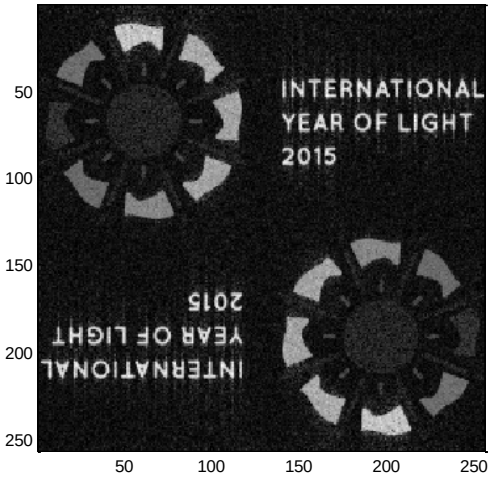
As the second step optimization for the shapes of the CGH apertures by the direct search was effective, we tested Scheme 2 with only 10 generations of parallel GA in the first step for optimizing the binary phase values of the apertures, which took 269 minutes for the error reaching around 0.18, followed by 20 rounds of direct search for optimizing the aperture shapes for the error reduced to 0.044, as shown in Fig. 4.10. The Scheme 2 generated an error slightly higher than 0.038, but needed only 1253 minutes of computation time, which is 3 times less than scheme 1.

In the other extremity, we performed scheme 3, in which only one local search optimization for a randomly selected chromosome, instead of PGA, was performed. It only took 9 minutes for the error to be reduced to 0.35. Then, the error was decreased to 0.047 after 20 rounds of direct search in the second step, as shown in Fig. 4.11. This scheme took a total time of 992 minutes for the computation.

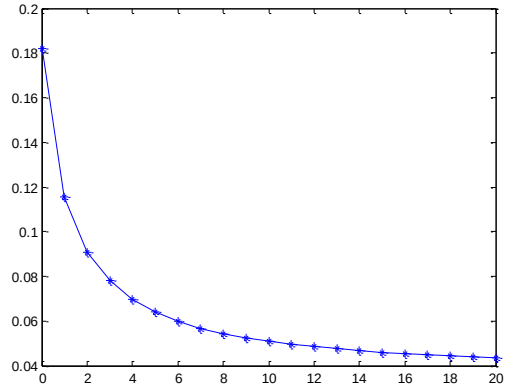
Scheme 1 (100 generations of PGA with local search + 20 rounds of direct search) provided the best solution with the error of 0.038 with the highest time cost; scheme 3 (1 local search + 20 rounds of direct search) achieved a good solution with the error of 0.047 in the shortest time; and scheme 2 (10 generations of PGA with local search + 20 rounds of direct search) reached the error of 0.044 at an accepted time cost. This two-step approach gives us a flexible way to balance between the time cost and optimal results.

	Scheme 1		Scheme 2		Scheme 3	
	Step 1: 100 generations	Step 2: 20 rounds	Step 1: 10 generations	Step 2: 20 rounds	Step 1: 1 DBS	Step 2: 20 rounds
Time (min.)	2657	983	269	931	9	897
Normalized error	8.8%	3.8%	18.2%	4.2%	35%	4.7%

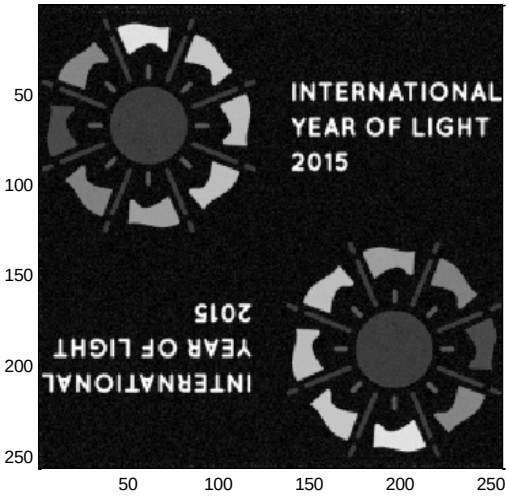
Tab. 4.1 The time cost and normalized error after step 1 and then step 2 for three different schemes



(a)

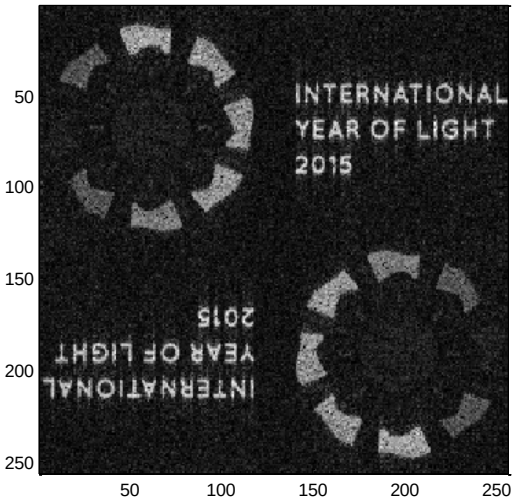


(b)

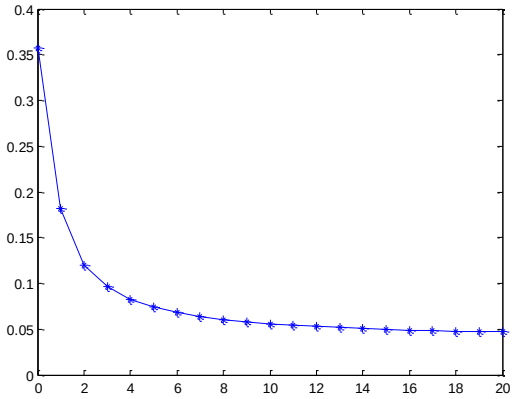


(c)

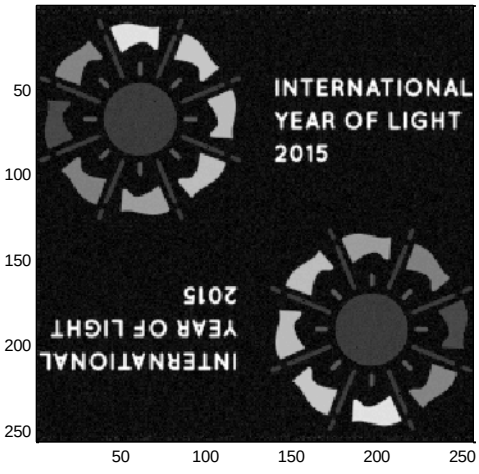
Fig. 4.10 Results by scheme 2: (a) image reconstructed after 10 generations of PGA with local search in step 1 (b) Normalized RMS errors after each round of direct search; (c) Grayscale image reconstructed by the CGH after step 2.



(a)



(b)



(c)

Fig. 4.11 Results by scheme 3: (a) image reconstructed after one local search in step 1 (b) Normalized RMS errors after each round of direct search; (c) Grayscale image reconstructed by the CGH after step 2.

4.7 Conclusion

The computer-generated hologram (CGH) has a very high space bandwidth product (SBWP), which can be explored for the high performance of the CGH. The Abbe transform expresses the diffraction pattern of a binary CGH of polygonal apertures as a coherence addition of the diffraction patterns from all the straight edges of different orientations and lengths. In the binary CGH, the arbitrary-shaped polygonal apertures were composed with the elementary quadrilateral apertures. This layout provides a high degree of freedom for the design optimization. The binary CGH was designed directly by parallel genetic algorithm with local search and then the direct search algorithm. The proposed design strategy permitted handling the binary CGH of very high SBWP and resulting in high performance with low reconstruction error of 3.8% for a grayscale image.

Conclusion

Computer-generated holograms (CGHs) are increasingly being used for a broad range of applications, such as beam shapers/splitters, optical disc read-heads, pattern generators and anti-fraud protection, optical display and imaging, etc. The objective of this thesis, as stated in the General Introduction, was to investigate on modelling, design and optimization of CGHs applied for practical projection system, and of CGHs with large space bandwidth product provided by advanced fabrication technology. This chapter summarizes main results and contributions of our research work presented in this thesis, and indicates several potential improvements and future work as well.

Achieved work

We considered certain working specification of a typical projection image system, e.g. working distance of 40 cm, depth of field of 10 cm and diffraction angle of ~50 degree with 632 nm working wavelength, and affordable fabrication ability, on which based, the Fourier type CGH with 2048 x 2048 pixels was designed and optimized by direct binary search (DBS) and then fabricated by E-beam lithography with the resolution of 2 um. The reconstrued image of 512 x 512 pixels showed a contrast of 87%. We also designed and discussed Fresnel type CGH. To achieve the required diffraction angle – 53 degree, we further discussed the optical architecture in holographic projection image system to “demagnify” the hologram pixels. The designed CGH and holographic projection image system were validated by optical reconstruction in the lab.

During the process of DBS, it had been noticed that pixels would eventually cluster to form polygonal apertures in hologram. Meanwhile, due to advancement of fabrication technology, CGHs of large space bandwidth product can contain huge number of pixels. So, we proposed to directly design arbitrary-shaped polygonal aperture based on triangular layout, and then performed a two-step optimization algorithm to fully exploit high number of degrees of freedom available. During the design, the diffraction of aperture was calculated by analytical Abbe transform, which can avoid of huge number of Fourier transforms over matrix with tremendous pixels in conventional algorithms. The reconstructed image can be expressed as a coherence addition of the diffraction patterns from all the straight edges of different

orientations and lengths. In the first step of optimization, the binary phases to triangular apertures was encoded by serial hybrid genetic algorithm (HGA) – genetic algorithm (GA) with a local search. Roulette wheel selection, double-point crossover, mutation with exponentially decreasing probability and elitism strategy are adopted in GA. Then direct search for floating covertices of elementary triangular apertures was used to further improve the results. A binary CGH of ~8000 x 8000 pixels was designed by our approach to reconstruct a gray-scale image with 2% reconstruction error.

We further analyzed the CGHs based on triangular layout and proposed quadrilateral aperture-based layout, which provides more degrees of freedom and forms more diverse polygonal aperture in CGHs. As the same manner, the diffraction of the quadrilateral aperture and the whole CGH was computed by Abbe transform. Although we still used two-step algorithm for optimization of binary phases in the first step and then orientation and lengths of edges in the second step, the parallel genetic algorithm (PGA) was developed, on a desk computer, based on a migration topology of one-way ring structure. Compared with those for the triangular aperture-based layout previously, more advanced selection and crossover operations are implemented, i.e. stochastic universal sampling selection and uniform crossover. Furthermore, we discussed three different schemes for this two-step algorithm, which provide a flexible way to balance the optimization performance and time cost: scheme 1 (100 generations of PGA with local search + 20 rounds of direct search) achieved the best solution with error of 3.8% in 3640 minutes; scheme 2 (10 generations of PGA with local search + 20 rounds of direct search) reached the error of 4.2% in 1200 minutes; scheme 3 (1 local search + 20 rounds of direct search) gave the error of 4.7% in 906 minutes.

Future work

As described in General Introduction, current research of CGH usually refers to the following 4 topics: object data acquisition, object wave representation, hologram encoding and image reconstruction. We made some progressive results on these two specific issues, i.e. modelling, design and optimization of CGHs applied for practical projection system, and of CGHs with large space bandwidth product. The first issue is related to the 4th topic and the second issue can be regarded as one in the 3rd topic. However, there are some limitations which can be further improved and studied in future work.

For the first issue, the wide diffractive angle was achieved by using two lenses after CGH/DOE in imaging system, and the CGH/DOE used was designed with paraxial approximation in the scope of scalar diffraction theory. A possible alternative is to directly design wide diffractive angle CGH/DOE, without using lenses after CGH/DOE to get desirable angles in imaging system. To model and design CGH/DOE with wide diffractive angles, non-paraxial scalar diffraction theory [111-113] may be used. In ref. [113], by presenting a projection step in combination with the Harvey model [111], the scalar diffraction theory is applied to estimate the non-paraxial diffraction pattern of a DOE at an observation plane in the far field. Vector diffraction theory can also be used instead in this case. Related electromagnetic algorithms usually refer to Finite-Difference Time-Domain (FDTD), Rigorous Coupled Wave Analysis (RCWA), etc. RCWA is usually limited to periodic structures, while FDTD can be used for almost all kinds of structures. Since it is computationally expensive for 3D FDTD simulations, the near field information is usually obtained by FDTD and far fields are then calculated by so-called "near-field to far-field transformation" [114].

As for the second issue, in our layout, CGH was divided to many cells and then each cell was further divided to four triangular or quadrilateral apertures, which would then merge into polygonal apertures later. Essentially, it is not a necessity to predefine these cells. Although provides simpler calculation, it reduces degree of freedom since the orientations of cell boundaries are fixed. One way to improve is to optimize vertices' positions of cells as well in the second step. This is straightforward and easy to implement. While, ideally, we just directly generate all kind of triangular or quadrilateral apertures in the first step without any predefined cells, which, however, will greatly enhance the complexity of diffraction computation by Abbe transform and optimization algorithms followed. Secondly, it will be interesting to implement parallel genetic algorithm using CUDA in graphical processing unit (GPU) in the first step, especially after complexity of computation and optimization are enhanced. Moreover, it is also worthy to adopt a more powerful algorithm in the second step since the used direct search of position is neither a global search nor a parallel algorithm. Lastly, although worked well in our design, the two-step algorithm is not the best choice from the perspective of optimization. How to optimize binary phases and edges' orientation and

lengths simultaneously can be further studied. Those potential improvements mentioned above are mainly for the approach itself.

From a more general perspective, synthesis of polygon-based CGH for 3D object/display (especially, in real-time) by the approach proposed in Chapter 3 and 4 can be studied. This may involve all the 4 research topics mentioned in Introduction.

Bibliography

[1] D. Gabor, “A New Microscopic Principle,” Nature, Vol. 161, pp. 777-778, (1948).

[2] D. Gabor, “Microscopy by Reconstructed Wave-Fronts," Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, Vol. 197, pp. 454-487, (1949).

[3] https://www.nobelprize.org/nobel_prizes/physics/laureates/1971/

[4] A. Lohmann, “Optical single-sideband transmission applied to the Gabor microscope,” Opt. Acta, Vol. 3, pp. 97-99 (1956).

[5] Maiman, T. H. “Stimulated optical radiation in ruby”. Nature. Vol. 187, pp. 493-494, (1960).

[6] E. N. Leith and J. Upatnieks, “Reconstructed wavefronts and communication theory,” Journal of the Optical Society of America, Vol. 52, pp. 1123-1130, (1962).

[7] B. R. Brown and A. W. Lohmann, “Complex spatial filtering with binary masks,” Applied optics, vol. 5, pp. 967-969, (1966).

[8] A. W. Lohmann and D. P. Paris, “Binary fraunhofer holograms, generated by computer,” Applied optics, vol. 6, pp. 1739-1748, (1967).

[9] B. R. Brown and A. W. Lohmann, “Computer-generated Binary Holograms," IBM Journal of Research and Development, vol. 13, pp. 160-168, (1969).

[10] U. Schnars and W. Jüptner, "Direct recording of holograms by a CCD target and numerical reconstruction," Appl. Opt. Vol. 33, pp. 179-181 (1994)

[11] G. Tricoles, "Computer generated holograms: an historical review," Appl. Opt. Vol. 26, pp. 4351-4360 (1987)

[12] G. A. Cirino etc. “Digital Holography: Computer-Generated Holograms and Diffractive Optics in Scalar Diffraction Domain”, NTECH Open Access Publisher, (2011)

[13] G. T. Nehmetallah etc. “Analog and digital holography with MATLAB”, SPIE Press, (2016)

[14] T. Poon, “Digital Holography and Three-Dimensional Display: Principles and Applications,” Springer US, (2016)

[15] T. Poon and J. Liu, “Introduction to Modern Digital Holography: With Matlab”, Cambridge University Press, (2014)

[16] R. V. Pole, “3-D imagery and holograms of objects illuminated in white light,” Appl. Phys. Lett. 10(1), 20–22 (1967).

[17] D. J. DeBitetto, “Holographic panoramic stereograms synthesized from white light recordings,” Appl. Opt. 8(8), 1740–1741 (1969).

[18] S.-C. Kim, etc., “Computer-generated holograms of a real three-dimensional object based on stereoscopic video images,” Appl. Opt. 45, 5669–5676 (2006).

[19] M. Bayraktar and M. Özcan. Method to calculate the far field of three-dimensional objects for computer-generated holography. Applied Optics, 49(24):4647–4654, Aug. 2010.

[20] Y. Zhao, L. Cao, H. Zhang, D. Kong, and G. Jin. Accurate calculation of computer generated holograms using angular-spectrum layer-oriented method. Optics Express, 23(20):25440, Oct. 2015.

[21] D. Leseberg and C. Frère. Computer-generated holograms of 3-D objects composed of tilted planar segments. Applied Optics, 27(14):3020–3024, July 1988.

[22] A. W. Lohmann. Three-dimensional properties of wave-fields. Optik, 51:105–107, 1978.

[23] K. Matsushima, M. Nakamura, and S. Nakahara. Silhouette method for hidden surface removal in computer holography and its acceleration using the switch-back technique. Optics Express, 22(20):24450–24465, Oct. 2014.

[24] B. T. Phong. Illumination for Computer Generated Pictures. Communications of the ACM, 18(6):311–317, June 1975.

[25] H. Sakata and Y. Sakamoto. Fast computation method for a Fresnel hologram using three-dimensional affine transformations in real space. Applied Optics, 48(34):H212– H221, Dec. 2009.

[26] H. Yoshikawa, T. Yamaguchi, and R. Kitayama. Real-Time Generation of Full Color Image Hologram with Compact Distance Look-up Table. In Advances in Imaging, OSA Technical Digest, page DWC4, Apr. 2009.

[27] N. Collings, J. L. Christmas, D. Masiyano, and W. A. Crossland, "Realtime phase-only spatial light modulators for 2D holographic display," J. Disp. Technol. vol. 11, no. 3, pp. 278–284, Mar. 2015.

[28] M. E. Lucente. Interactive computation of holograms using a look-up table. Journal of Electronic Imaging, 2(1):28–34, Jan. 1993.

[29] L. G. Neto, D. Roberge, and Y. Sheng, "Programmable optical phasemostly holograms with coupled-mode modulation liquid crystal television," Appl. Opt., vol. 34, no. 11, pp. 1944–1950, 1995.

[30] F. Dufaux, Y. Xing, B. Pesquet-Popescu, and P. Schelkens. Compression of digital holographic data: an overview. In Applications of Digital Image Processing XXXVIII, volume Proc. SPIE 9599, pages 95990I–95990I–11, Sept. 2015.

[31] Y. Xing, M. Kaaniche, B. Pesquet-Popescu, and F. Dufaux. Digital Holographic Data Representation and Compression. Academic Press, Oct. 2015.

[32] K. Yamamoto, T. Senoh, R. Oi, and T. Kurita. 8k4k-size computer generated hologram for 3-D visual system using rendering technology. In Universal Communication Symposium (IUCS), 2010 4th International, pages 193–196, Oct. 2010.

[33] F. Yaras, H. Kang, and L. Onural. State of the Art in Holographic Displays: A Survey. Journal of Display Technology, 6(10):443–454, Oct. 2010.

[34] J. D. Johnson, Classical Electrodynamics Third Edition, Wiley, (1998)

[35] M. Born and E. Wolf, Principles of Optics, 7th Edition, Cambridge University Press, 1999

[36] J. Goodman, Introduction to Fourier optics, Third Edition, Roberts and Company Publishers, 2004

[37] Michael A. Seldowitz, Jan P. Allebach, and Donald W. Sweeney, "Synthesis of digital holograms by direct binary search," Appl. Opt. 26, 2788-2798 (1987)

[38] Brian K. Jennison, Jan P. Allebach, and Donald W. Sweeney, "Efficient design of direct-binary-search computer-generated holograms," J. Opt. Soc. Am. A 8, 652-660 (1991)

[39] Andrew G.KirkTrevor J.Hall, Design of binary computer generated holograms by simulated annealing: coding density and reconstruction error, Optics Communications, Vol. 94, pp. 491-496 (1992)

[40] Masaki Taniguchi, Katsunori Matsuoka, and Yoshiki Ichioka, "Computer-generated multiple-object discriminant correlation filters: design by simulated annealing," Appl. Opt. 34, 1379-1385 (1995)

[41] Nobukazu Yoshikawa, Masahide Itoh, Toyohiko Yatagai, "Use of generic algorithm for computer-generated holograms", Proc. SPIE 2577, International Conference on Applications of Optical Holography, (27 July 1995);

[42] J.-N. Gillet and Y. Sheng, “Multiplexed computer-generated holograms with polygonal-aperture layouts optimized by genetic algorithm,” Appl. Opt. 42(20), 4156–4165 (2003).

[43] R. W. Gerchberg and W. O. Saxton, “A practical algorithm for the determination of phase from image and diffraction plane pictures,” Optik 35(2), 237–246 (1972).

[44] J. R. Fienup, “Phase retrieval algorithms: a comparison,” Appl. Opt. 21(15), 2758–2769 (1982).

[45] J. R. Fienup, “Phase retrieval algorithms: a personal tour,” Appl. Opt. 52(1), 45–56 (2013).

[46] Ying Tsung Lu, Huang Yen Lin, Sien Chi, "Nonoptical method for making dot-matrix hologram", Proc. SPIE 3956, Practical Holography XIV and Holographic Materials VI, (17 March 2000)

[47] S. L. Yeh and S. T. Lin, “Anticounterfeiting method for a dot-matrix hologram composed of grating dots with different fringe orientations,” Opt. Eng. 54(11), 113106 (2015).

[48] Craig D. Newswanger, "Dot matrix technique for generating diffraction grating patterns", Proc. SPIE 2333, Fifth International Symposium on Display Holography, (17 February 1995);

[49] S. L. Yeh, Shyh Tsong Lin, "Dot-matrix hologram with hidden image," Optical Engineering 41(2), (1 February 2002).

[50] S. L. Yeh, "Hiding techniques to enhance anticounterfeiting capacity of dot-matrix holograms," Optical Engineering 44(8), 087001 (1 August 2005)

[51] B. Zarkov, D. Gruji´c, and D. Panteli´c, “High-resolution dot-matrix hologram generation,” Phys. Scr., vol. T149, p. 014021, 2012.

[52] Yaotang Li, Tianji Wang, Shining Yang, Shichao Zhang, Shaowu Fan, Huanrong Wen, "Theoretical and experimental study of dot matrix hologram", Proc. SPIE 3559, Holographic Displays and Optical Elements II, (4 August 1998);

[53] Sheng Lih Yeh, "Using random features of dot-matrix holograms for anticounterfeiting," Appl. Opt. 45, 3698-3703 (2006)

[54] Alexander V. Morozov, Andrey N. Putilin, Sergey S. Kopenkin, Yuriy P. Borodin, Vladislav V. Druzhin, Sergey E. Dubynin, and German B. Dubinin, "3D holographic printer: Fast printing approach," Opt. Express 22, 2193-2206 (2014)

[55] H. Kang etc. Color Holographic Wavefront Printing Technique for Realistic Representation, IEEE Transactions on Industrial Informatics, Vol. 12, pp. 1590-1598 (2016)

[56] Fei Yang, Yuri Murakami, and Masahiro Yamaguchi, "Digital color management in full-color holographic three-dimensional printer," Appl. Opt. 51, 4343-4352 (2012)

[57] Masahiro Yamaguchi, Nagaaki Ohyama, and Toshio Honda, "Holographic three-dimensional printer: new method," Appl. Opt. 31, 217-222 (1992)

[58] Hao Zhang, Yan Zhao, Liangcai Cao, and Guofan Jin, "Layered holographic stereogram based on inverse Fresnel diffraction," Appl. Opt. 55, A154-A159 (2016)

[59] Changwon Jang, Chang-Kun Lee, Jinsoo Jeong, Gang Li, Seungjae Lee, Jiwoon Yeom, Keehoon Hong, and Byoungho Lee, "Recent progress in see-through three-dimensional displays using holographic optical elements [Invited]," Appl. Opt. 55, A71-A85 (2016)

[60] Keehoon Hong, Soon-gi Park, Jiwoon Yeom, Jonghyun Kim, Ni Chen, Kyungsuk Pyun, Chilsung Choi, Sunil Kim, Jungkwuen An, Hong-Seok Lee, U-in Chung, and Byoungho Lee, "Resolution enhancement of holographic printer using a hogel overlapping method," Opt. Express 21, 14047-14055 (2013)

[61] Hiroshi Yoshikawa and Takeshi Yamaguchi, Review of Holographic Printers for Computer-Generated Holograms, IEEE Transactions on Industrial Informatics, Vol. 12, pp. 1584-1589 (2016)

[62] Youngmin Kim, Elena Stoykova, Hoonjong Kang, Sunghee Hong, Joosup Park, Jiyong Park, and Jisoo Hong, "Seamless full color holographic printing method based on spatial partitioning of SLM," Opt. Express 23, 172-182 (2015)

[63] Gang Li, Jinsoo Jeong, Dukho Lee, Jiwoon Yeom, Changwon Jang, Seungjae Lee, and Byoungho Lee, "Space bandwidth product enhancement of holographic display using high-order diffraction guided by holographic optical element," Opt. Express 23, 33170-33183 (2015)

[64] Liangcai Cao, Zheng Wang, Hao Zhang, Guofan Jin, and Claire Gu, "Volume holographic printing using unconventional angular multiplexing for three-dimensional display," Appl. Opt. 55, 6046-6051 (2016)

[65] B. C. Kress and P. Meyrueis, Applied Digital Optics: From Micro-optics to Nanophotonics. Chichester, UK: John Wiley & Sons, Ltd, Oct. 2009.

[66] D. C. O'Shea, T. J. Suleski, A. D. Kathman, and D. W. Prather, Diffractive Optics: Design, Fabrication, and Test. SPIE Press, 2004.

[67] P. P. Clark and C. Londono, “Production of kinoforms by single point diamond machining,” Optics News, vol. 15, p. 39 1, Dec. 1989.

[68] Lin, B. J., "Optical Lithography", SPIE Press, Bellingham, WA, 2009

[69] K. Jain ; C.G. Willson ; B.J. Lin, Ultrafast deep UV Lithography with excimer lasers, IEEE Electron Device Letters, Vol. 3, pp. 53-55, (1982)

[70] Michael T. Gale, Karl Knop, "The Fabrication Of Fine Lens Arrays By Laser Beam Writing", Proc. SPIE 0398, Industrial Applications of Laser Technology, (26 October 1983);

[71] M. V. Kessels, M. El Bouz, R. Pagan, and K. Heggarty, “Versatile stepper based maskless microlithography using a liquid crystal display for direct write of binary and multilevel microstructures,” Journal of Micro/Nanolithography, MEMS and MOEMS, vol. 6, p. 033002, July 2007.

[72] Selimis A, Mironov V, Farsari M. “Direct laser writing: principles and materials for scaffold 3D printing,” Microelectron. Eng. 2015; 132:83–9.

[73] S. M. Arnold, “Electron Beam Fabrication of Computer-Generated Holograms," Optical Engineering,” vol. 24, p. 245803, Aug. 1985.

[74] Yifang Chen, “Nanofabrication by electron beam lithography and its applications: A review,” Microelectronic Engineering, Volume 135, Pages 57-72, 5 March 2015

[75] Jason Geng, "Structured-light 3D surface imaging: a tutorial," Adv. Opt. Photon. 3, 128-160 (2011)

[76] I. Ishii, K. Yamamoto, K. Doi, and T. Tsuji, “High-speed 3D image acquisition using coded structured light projection,” in IEEE/RSJ International Conference on Intelligent Robots and Systems, 2007. IROS 2007 (IEEE, 2007), pp. 925–930.

[77] J. L. Posdamer and M. D. Altschuler, “Surface measurement by space-encoded projected beam systems,” Comput. Graph. Image Processing 18(1), 1–17 (1982).

[78] Ralf Vandenhouten, Andreas Hermerschmidt, Richard Fiebelkorn, "Design and quality metrics of point patterns for coded structured light illumination with diffractive optical elements in optical 3D sensors", Proc. SPIE 10335, Digital Optical Technologies 2017, 1033518 (26 June 2017)

[79] K. Sato and S. Inokuchi, “Range-imaging system utilizing nematic liquid crystal mask,” in Proceedings of International Conference on Computer Vision (IEEE Computer Society Press, 1987), pp. 657–661.

[80] R. J. Valkenburg and A. M. McIvor, “Accurate 3D measurement using a structured light system,” Image Vision Comput. 16(2), 99–110 (1998).

[81] A. Wiegmann, H. Wagner, and R. Kowarschik, "Human face measurement by projecting bandlimited random patterns," Opt. Express 14, 7692-7698 (2006)

[82] Edward Buckley, Adrian J. Cable, Timothy D. Wilkinson, "Precision measurement system using binary phase computer-generated holograms," Optical Engineering 50(9), 091308 (1 September 2011).

[83] Edward Buckley, "Holographic Laser Projection," J. Display Technol. 7, 135-140 (2011)

[84] E. Buckley, “Computer generated holograms for real-time image display and sensor applications,” Ph.D. dissertation, Dept. Elect. Eng., Cambridge Univ., Cambridge, United Kingdom (2006).

[85] Hecht, E. (1998). Optics, 3rd edition, Addison-Wesley, San Francisco, California.

[86] P. W. M. Tsang， T-C Poon， J-P Liu, “Three-dimensional displays: a review and applications analysis,” Appl. Sci. 2018, 8(5), 830

[87] N. S. Holliman, N. A. Dodgson, G. E. Favalora, and L. Pockett, “Three-dimensional displays: a review and applications analysis,” IEEE Trans. on broadcasting, Vol. 57, No. 2, June, 362-371 (2011)

[88] K. V. Chellappan, E. Erden, and H. Urey, “Laser-based displays: a review,” Appl. Opt. Vol. 49, No. 25, F70-F98 (2010)

[89] R. A. Clark , Y. H. Pua, K. Fortin, C. Ritchie, K. E. Webster, L. Denehy, and A. L. Bryant, “Validity of the Microsoft Kinect for assessment of postural control,” Gait & Posture, 36, 372–377 (2012).

[90] Kress and J. Lee, “Optical gesture sensing and depth mapping technologies for head-mounted displays: an overview,” Proceeding of SPIE, Vol. 8720, paper 87200C-2 (2013).

[91] T. Shimobaba, T. Kakue, and T. Ito, “Real-time and low speckle holographic projection,” in Proc. IEEE 13th Int. Conf. Ind. Informat. (INDIN’15), 2015, pp. 732–741.

[92] D. Masiyano, N. Collings and J. Christmas, “Lasers for Phase Only Holographic Projection and Head Up Display Applications,” IEEE High Power Diode Laser and System Conference, 20-21, (2013).

[93] J. Rosen and R. Kelner, “Three-dimensional imaging by self-reference single-channel digital incoherent holography,” IEEE Trans. Ind. Informat., Volume: 12, pp. 1571-1583, 2016

[94] M. Y. Omel, M. V. Gladys, and L. Jesús, “Encoding complex fields by using a phase-only optical element,” Opt. Lett., vol. 39, pp. 1740–1743, 2014.

[95] J. P. Liu,W. Y. Hsieh, T. -C. Poon, and P.W.M. Tsang, “Complex Fresnel hologram display using a single SLM,” Appl. Opt., vol. 50, pp. H128–H135, 2011.

[96] P. W. M. Tsang, Y. Pan and T. C. Poon, “Binary hologram generation based on shape adaptive sampling,” Opt. Communication. Vol 319. 8–13, (2014).

[97] P. W. M. Tsang, T. C. Poon, A. S. M. Jiao, “Embedding intensity image in grid-cross down-sampling (GCD) binary holograms based on block truncation coding,” Opt. Commu. Vol 304. 62-70, (2013).

[98] S. N. Koreshev, O. V. Nikanorov, and A. D. Gromov, “Method of synthesizing hologram projectors based on breaking down the structure of an object into typical elements, and a software package for implementing it,” J. Opt. Technol. Vol. 79, No. 12, 769-774, (2012).

[99] R. Horisaki, and T. Tahara, “Phase-shift binary digital holography,” Opt. Letters, Vol. 39, No. 22, 6375-6378 (2014).

[100] T. Leportier, M. C. Park, Y. S. Kim and T, Kim, “Converting optical scanning holograms of real objects to binary Fourier holograms using an iterative direct binary search algorithm,” Opt. Express, Vol. 23, No. 3, 3403-3411 (2015).

[101] L. Schermelleh, P. M. Carlton, A. Haase, L. Shao, L. Winoto, P. Kner, B. Burke, M. C. Caroso, D. A. Agard, M. G. L. Gustafsson, H. Leonhardt and J. W. Sedat, “Subdiffraction multicolor imaging of the nuclear periphery with 3D structured illumination microscopy,” Science. Vol. 320 no. 5881, 6 June, 1332-1336 (2008) .

[102] A. Firsov, A. Firsov, B. Loechel, A. Erko, A. Svintsov and S. Zaitsev, “Fabrication of digital rainbow holograms and 3- D imaging using SEM based e-beam lithography,” Opt. Express, Vol. 22, No. 23 28756-28770 (2014).

[103] Kyoji Matsushima, and Sumio Nakahara, “Extremely high-definition full-parallax computer-generated hologram created by the polygon-based method,” Appl. Opt. Vol. 48, No. 34, H54-H63 (2009)

[104] C. Newswanger, “Holographic diffraction grating patterns and method for creating the same,” U.S. patent 5,291,317 (March 1994).

[105] M. Lucente, “Diffraction-Specific Fringe Computation for Electro- Holography,” Ph.D. thesis dissertation, Massachusetts Institute of Technology, (1994)

[106] J. Komrska, “Simple derivation of formulas for Fraunhofer diffraction at polygonal apertures,” J. Opt. Soc. Amer., vol. 72, pp. 1382–1384, 1982.

[107] A. Sommerfeld, Optics (Academic, New York, 1954), p. 233ff.

[108] D. Im, E. Moon, Y. Park, D. Lee, J. Hahn, and H. Kim, “Phase regularized polygon computer-generated Holograms,” Opt. Lett. 39, 3642–3645 (2014).

[109] H. Kim, J. Hahn, and B. Lee, “Mathematical modeling of triangle mesh-modeled three-dimensional surface objects for digital holography,” Appl. Opt. 47, D117–D127 (2008).

[110] S. N. Sivanandam and S. N. Deepa, Introduction to Genetic Algorithms (Springer-Verlag, 2008).

[111] J. E. Harvey, D. Bogunovic, and A. Krywonos, “Aberrations of Diffracted Wave Fields: Distortion,” Applied Optics, vol. 42, p. 1167, Mar. 2003.

[112] D. C. Cole, E. Barouch, U. Hollerbach, and S. A. Orszag, “Derivation and Simulation of Higher Numerical Aperture Scalar Aerial Images,” Japanese Journal of Applied Physics, vol. 31, pp. 4110-4119, Dec. 1992.

[113] G.-N. Nguyen, K. Heggarty, P. G_erard, B. Serio, and P. Meyrueis, “Computationally efficient scalar nonparaxial modeling of optical wave propagation in the far-field,” Applied Optics, vol. 53, p. 2196, Mar. 2014.

[114] Allen Taflove, “Computational Electrodynamics: The Finite-Difference Time-Domain Method,” 3rd edition, June 2005

Appendix A: MATLAB scripts

The MATLAB scripts shown below are written to implement the design algorithms (two-step optimization: hybrid genetic algorithm – genetic algorithm with local search, and direct search) and give the reconstructed results presented in Chapter 3. These scripts also give some hints on how to implement coarse-grained parallel genetic algorithm used in Chapter 4, because, for subpopulations on separate processors, genetic algorithm runs parallelly in the same manner as that in Chapter 3.

```
clc;
clear all;
close all;

target0=imread('E:\CGH\image.png');
target1=double(rgb2gray(target0));

Wa=2*size(target1,1); Wb=size(target1,2);

Mc=242;Nc=242;
W1=((Mc)/2)-((Wa)/2)+1; W2=((Mc)/2)+((Wa)/2); W3=((Nc)/2)-((Wb)/2)+1;
W4=((Nc)/2)+((Wb)/2);
target=zeros(Mc,Nc);
target(W1:Mc/2-1, W3:W4)=target1(2:end,:);

target=target+imrotate(target,180);
target=target/sqrt(sum(sum((abs(target)).^2)));  % the amplitude; the
total power of the target image is normalized to unit;
 figure; imagesc(target); colormap gray; axis image;

P=32; % the cell number and the pixel number in each cell;
Ran1=floor(P/2)+1; Ran2=floor(P/2)+1; % the centre point;

PP=1; % the scale of the size
m=-Mc*PP/2:Mc*PP/2-1; n=-Nc*PP/2:Nc*PP/2-1;

Pattern_tri=zeros(Mc*PP, Nc*PP); Pattern_t(:,:)=zeros(Mc*PP, Nc*PP);

Popul_num=80;


%%

Xr=Ran1; Yr=Ran2;
```

```
delta_Vector=[-1/(2*Nc), -1/(2*Mc)];

% the first triangle
tL_ab=[0, 1/Mc]; tL_bo=[(Yr/P)/Nc, (Xr/P-1)/Mc]; tL_oa=[(-Yr/P)/Nc, (-
Xr/P)/Mc];
nL_ab=[-1/Mc, 0]; nL_bo=[-(Xr/P-1)/Mc, (Yr/P)/Nc]; nL_oa=[-(-Xr/P)/Mc, (-
Yr/P)/Nc];
Mid_ab=[0, 1/Mc]./2+delta_Vector; Mid_bo=[(Yr/P)/Nc,
(1+Xr/P)/Mc]./2+delta_Vector; Mid_oa=[(Yr/P)/Nc,
(Xr/P)/Mc]./2+delta_Vector;

% the second triangle
tL_bc=[1/Nc, 0]; tL_co=[(Yr/P-1)/Nc, (Xr/P-1)/Mc]; tL_ob=[(-Yr/P)/Nc, (1-
Xr/P)/Mc];
nL_bc=[0, 1/Nc]; nL_co=[-(Xr/P-1)/Mc, (Yr/P-1)/Nc]; nL_ob=[-(1-Xr/P)/Mc,
(-Yr/P)/Nc];
Mid_bc=[1/Nc, 2/Mc]./2+delta_Vector; Mid_co=[(Yr/P+1)/Nc,
(Xr/P+1)/Mc]./2+delta_Vector; Mid_ob=Mid_bo;

% the third triangle
tL_cd=[0, -1/Mc]; tL_do=[(Yr/P-1)/Nc, (Xr/P)/Mc]; tL_oc=[(1-Yr/P)/Nc, (1-
Xr/P)/Mc];
nL_cd=[1/Mc, 0]; nL_do=[-(Xr/P)/Mc, (Yr/P-1)/Nc]; nL_oc=[-(1-Xr/P)/Mc,
(1-Yr/P)/Nc];
Mid_cd=[2/Nc, 1/Mc]./2+delta_Vector; Mid_do=[(1+Yr/P)/Nc,
(Xr/P)/Mc]./2+delta_Vector; Mid_oc=Mid_co;

% the forth triangle
tL_da=[-1/Nc, 0]; tL_ao=[(Yr/P)/Nc, (Xr/P)/Mc]; tL_od=[(1-Yr/P)/Nc, (-
Xr/P)/Mc];
nL_da=[0, -1/Nc]; nL_ao=[-(Xr/P)/Mc, (Yr/P)/Nc]; nL_od=[-(-Xr/P)/Mc, (1-
Yr/P)/Nc];
Mid_da=[1/Nc, 0]./2+delta_Vector; Mid_ao=Mid_oa; Mid_od=Mid_do;

tL(:,:,1)=[tL_ab; tL_bo; tL_oa];
tL(:,:,2)=[tL_bc; tL_co; tL_ob];
tL(:,:,3)=[tL_cd; tL_do; tL_oc];
tL(:,:,4)=[tL_da; tL_ao; tL_od];

nL(:,:,1)=[nL_ab; nL_bo; nL_oa];
nL(:,:,2)=[nL_bc; nL_co; nL_ob];
nL(:,:,3)=[nL_cd; nL_do; nL_oc];
nL(:,:,4)=[nL_da; nL_ao; nL_od];

Mid(:,:,1)=[Mid_ab; Mid_bo; Mid_oa];
Mid(:,:,2)=[Mid_bc; Mid_co; Mid_ob];
Mid(:,:,3)=[Mid_cd; Mid_do; Mid_oc];
Mid(:,:,4)=[Mid_da; Mid_ao; Mid_od];



   for S=1:4;
```

```
        MM_1=m*tL(1,1,S); NN_1=n*tL(1,2,S);
        DOT_1=MM_1'*ones(1,Nc*PP)+ones(Mc*PP,1)*NN_1;
        temp_1=DOT_1+(DOT_1==0);
        Sinc_1=sin(pi*DOT_1)./(pi*temp_1)+(temp_1==1);

        MM_2=m*tL(2,1,S); NN_2=n*tL(2,2,S);
        DOT_2=MM_2'*ones(1,Nc*PP)+ones(Mc*PP,1)*NN_2;
        temp_2=DOT_2+(DOT_2==0);
        Sinc_2=sin(pi*DOT_2)./(pi*temp_2)+(temp_2==1);

        MM_3=m*tL(3,1,S); NN_3=n*tL(3,2,S);
        DOT_3=MM_3'*ones(1,Nc*PP)+ones(Mc*PP,1)*NN_3;
        temp_3=DOT_3+(DOT_3==0);
        Sinc_3=sin(pi*DOT_3)./(pi*temp_3)+(temp_3==1);


        MM_4=m*nL(1,1,S); NN_4=n*nL(1,2,S);
        NormalD_1=MM_4'*ones(1,Nc*PP)+ones(Mc*PP,1)*NN_4;

        MM_5=m*nL(2,1,S); NN_5=n*nL(2,2,S);
        NormalD_2=MM_5'*ones(1,Nc*PP)+ones(Mc*PP,1)*NN_5;

        MM_6=m*nL(3,1,S); NN_6=n*nL(3,2,S);
        NormalD_3=MM_6'*ones(1,Nc*PP)+ones(Mc*PP,1)*NN_6;


        MM_7=m*Mid(1,1,S); NN_7=n*Mid(1,2,S);
        MidD_1=exp(-1i*(2*pi)*(MM_7'*ones(1,Nc*PP)+ones(Mc*PP,1)*NN_7));

        MM_8=m*Mid(2,1,S); NN_8=n*Mid(2,2,S);
        MidD_2=exp(-1i*(2*pi)*(MM_8'*ones(1,Nc*PP)+ones(Mc*PP,1)*NN_8));

        MM_9=m*Mid(3,1,S); NN_9=n*Mid(3,2,S);
        MidD_3=exp(-1i*(2*pi)*(MM_9'*ones(1,Nc*PP)+ones(Mc*PP,1)*NN_9));

        MMNN=2*pi*((m.^2)'*ones(1,Nc*PP)+ones(Mc*PP,1)*(n.^2));
        temp_4=MMNN+1i*(MMNN==0);
        Term=1i./temp_4;

     %   A=exp(-i*2*pi*(m'*ones(1,Nc*PP)*Mi/Mc+ones(Mc*PP,1)*n*Nj/Nc)); %
the factor related to the cell location;


Pattern_tri=Term.*(Sinc_1.*NormalD_1.*MidD_1+Sinc_2.*NormalD_2.*MidD_2+Si
nc_3.*NormalD_3.*MidD_3); % independent of location;

        Pattern_t(:,:,S)=Pattern_tri; % the pattern from the triangle #S;

   end
   A_pat=Pattern_t(:,:,1); %record the pattern of each triangle;
   B_pat=Pattern_t(:,:,2);
   C_pat=Pattern_t(:,:,3);
   D_pat=Pattern_t(:,:,4);
```

```
%%
Popul=zeros(Popul_num+2,4*Mc*Nc);
Popul_parental=zeros(Popul_num,4*Mc*Nc);
Popul_parental_c=zeros(Popul_num,4*Mc*Nc);
EA=zeros(Mc,Nc); EB=zeros(Mc,Nc); EC=zeros(Mc,Nc); ED=zeros(Mc,Nc);
tic;
for GA_k=1:300;

    if GA_k==1;
        Popul=randsrc(Popul_num+2,4*Mc*Nc,[1 0]); % random Popul.;
    else
        Popul=Popul_parental_c;
        Popul(Popul_num+1:Popul_num+2,:)=Elitism(1:2,:); % the last two
ranks replaced by the elite chromosomes;
    end

%%  Evaluation and Ranking

for Popul_k=1:Popul_num+2;



  for Mi=0:Mc-1;
    for Nj=0:Nc-1;

        EA(Mi+1,Nj+1)=Popul(Popul_k,Mi*4*Nc+4*Nj+1);
EB(Mi+1,Nj+1)=Popul(Popul_k,Mi*4*Nc+4*Nj+2);
        EC(Mi+1,Nj+1)=Popul(Popul_k,Mi*4*Nc+4*Nj+3);
ED(Mi+1,Nj+1)=Popul(Popul_k,Mi*4*Nc+4*Nj+4);

    end
  end


Chrom_Pattern_C=A_pat.*fftshift(fft2(EA))+B_pat.*fftshift(fft2(EB))+C_pat
.*fftshift(fft2(EC))+D_pat.*fftshift(fft2(ED));

  Power_total=sum(sum((abs(Chrom_Pattern_C)).^2));
  Pattern_Norm=Chrom_Pattern_C/sqrt(Power_total); % the amplitude; the
total power of the reconstructed image is normalized to unit;

  MSE(Popul_k)=sum(sum((abs((abs(Pattern_Norm(W1:(Mc)/2+1, W3:W4))).^2-
(abs(target(W1:(Mc)/2+1, W3:W4))).^2)).^2));

end

[Rank_MSE, index]=sort(MSE); % rank as the ascending of MSE;

MSE_eli_min(GA_k)=Rank_MSE(1); % the min MSE;
MSE_eli_mean(GA_k)=sum(Rank_MSE(1:Popul_num))/Popul_num; % the mean vaule
of MSE;
```

```
%% Elitism (the best two chromosomes are kept to the next G without
crossover and mutation);

Elitism(1,:)=Popul(index(1),:);
Elitism(2,:)=Popul(index(2),:);


%% Rank-based fitness function
if GA_k==1;
fitness = 1-cdf('Normal',1:Popul_num,Popul_num/2,Popul_num/6); % the
distribution of the fitness
cum_fitness = zeros(1,Popul_num);

for p_rank = 1:Popul_num
   if p_rank == 1
      cum_fitness(1) = fitness(1);
   else
      cum_fitness(p_rank) = fitness(p_rank) + cum_fitness(p_rank-1);
   end
end
fitness = fitness/cum_fitness(Popul_num); % normalized probability
cum_fitness = cum_fitness/cum_fitness(Popul_num); % the cumulative
probability;
end

%% Roulette game

pair_num = zeros(1,Popul_num);
for Popul_k1=1:Popul_num;
        R_temp=rand;
        pair_num(Popul_k1)=length(find(cum_fitness<R_temp))+1; % roulette
selection
      %  MSE_new(Popul_k1)=Rank_MSE(pair_num(Popul_k1));
        Popul_parental(Popul_k1,:)=Popul(index(pair_num(Popul_k1)),:); %
obtain the selected chromosomes;
end
% MSE_new_mean(GA_k)=sum(MSE_new)/Popul_num;
% MSE_new_min(GA_k)=min(MSE_new);

%% Double crossover;

cross_rate=0.9;
for Popul_k2=1:2:Popul_num;
    if rand<cross_rate;
        cross_p1=randi([2,4*Mc*Nc-1],1); % the first point;
        cross_p2=randi([cross_p1+1,4*Mc*Nc],1); % the second point;

    Popul_parental_c(Popul_k2,:)=Popul_parental(Popul_k2,:); % the
chromosome with the odd number;
    Popul_parental_c(Popul_k2,cross_p1:cross_p2)=
Popul_parental(Popul_k2+1,cross_p1:cross_p2); % obtain the genes between
the two points

    Popul_parental_c(Popul_k2+1,:)=Popul_parental(Popul_k2+1,:); % the
chromosome with the odd+1 number;
```

```
    Popul_parental_c(Popul_k2+1,cross_p1:cross_p2)=
Popul_parental(Popul_k2,cross_p1:cross_p2); % obtain the genes between
the two points

    else
        Popul_parental_c(Popul_k2,:)=Popul_parental(Popul_k2,:);
        Popul_parental_c(Popul_k2+1,:)=Popul_parental(Popul_k2+1,:);
    end

end

%% Mutation

mutation_rate=0.01*0.985^(GA_k-1);
for Popul_k3=1:Popul_num;
    mutation_p=find(rand(1,4*Mc*Nc)<mutation_rate);  % the positions of
all the genes for mutation
    Popul_parental_c(Popul_k3,mutation_p(1:end))=1-
Popul_parental_c(Popul_k3,mutation_p(1:end));
end

%% Local search

LS_num=1; % the number of Chromosomes selected randomly for LS
for LS_k=1:LS_num;
    LS_s=randi([1,Popul_num]);
    Chrom_LS=Popul_parental_c(LS_s,:); % the Chromosome selected;
    Chrom_LS_1=Chrom_LS; Chrom_LS_2=Chrom_LS;



        for Mi=0:Mc-1; % calculate the whole diffraction pattern from all
the cells (triangulars) for only one time;
            for Nj=0:Nc-1;
                EA(Mi+1,Nj+1)=Chrom_LS_1(1,Mi*4*Nc+4*Nj+1);
EB(Mi+1,Nj+1)=Chrom_LS_1(1,Mi*4*Nc+4*Nj+2);
                EC(Mi+1,Nj+1)=Chrom_LS_1(1,Mi*4*Nc+4*Nj+3);
ED(Mi+1,Nj+1)=Chrom_LS_1(1,Mi*4*Nc+4*Nj+4);
            end
        end

Pattern_LS_1=A_pat.*fftshift(fft2(EA))+B_pat.*fftshift(fft2(EB))+C_pat.*f
ftshift(fft2(EC))+D_pat.*fftshift(fft2(ED));

    for LS_p=1:4*Mc*Nc; % value would change one by one;

        Chrom_LS_2(1,LS_p)=1-Chrom_LS_1(1,LS_p); % inversion;
        M_index=ceil(LS_p/(4*Nc))-1; N_index=ceil((LS_p-M_index*4*Nc)/4)-
1; % the cell location of the value changed;

                    A_fourier=exp(-
1i*2*pi*(m'*ones(1,Nc*PP)*M_index/Mc+ones(Mc*PP,1)*n*N_index/Nc)); % the
factor related with the cell location;

                    if mod(LS_p,4)==1;
```

```
                        T_1=A_fourier.*Chrom_LS_1(1,LS_p).*A_pat; % the
diffraction of the triangular;
                        T_2=A_fourier.*Chrom_LS_2(1,LS_p).*A_pat; % the
diffraction of the triangular after the inversion;
                    elseif mod(LS_p,4)==2;
                        T_1=A_fourier.*Chrom_LS_1(1,LS_p).*B_pat;
                        T_2=A_fourier.*Chrom_LS_2(1,LS_p).*B_pat;
                    elseif mod(LS_p,4)==3;
                        T_1=A_fourier.*Chrom_LS_1(1,LS_p).*C_pat;
                        T_2=A_fourier.*Chrom_LS_2(1,LS_p).*C_pat;
                    else
                        T_1=A_fourier.*Chrom_LS_1(1,LS_p).*D_pat;
                        T_2=A_fourier.*Chrom_LS_2(1,LS_p).*D_pat;
                    end



        Pattern_LS_2=Pattern_LS_1-T_1+T_2;

        Power_total_1=sum(sum((abs(Pattern_LS_1)).^2));
        Pattern_Norm_1=Pattern_LS_1/sqrt(Power_total_1);
        MSE_LS_1=sum(sum((abs((abs(Pattern_Norm_1(W1:(Mc)/2+1,
W3:W4))).^2-(abs(target(W1:(Mc)/2+1, W3:W4))).^2)).^2));

        Power_total_2=sum(sum((abs(Pattern_LS_2)).^2));
        Pattern_Norm_2=Pattern_LS_2/sqrt(Power_total_2);
        MSE_LS_2=sum(sum((abs((abs(Pattern_Norm_2(W1:(Mc)/2+1,
W3:W4))).^2-(abs(target(W1:(Mc)/2+1, W3:W4))).^2)).^2));

        if MSE_LS_2<MSE_LS_1;
            Chrom_LS_1(1,LS_p)=Chrom_LS_2(1,LS_p); % accept the inversion
and transfer
            Pattern_LS_1=Pattern_LS_2;
        else
            Chrom_LS_2(1,LS_p)=Chrom_LS_1(1,LS_p); % reject the inversion
and keep the value before inversion
            Pattern_LS_2=Pattern_LS_1;
        end

    end
    Popul_parental_c(LS_s,:)=Chrom_LS_1;
end


fprintf('GA_k=%d\n', GA_k);

end
toc;
%%

  for Mi=0:Mc-1;
    for Nj=0:Nc-1;
        EA(Mi+1,Nj+1)=Popul(index(1),Mi*4*Nc+4*Nj+1);
EB(Mi+1,Nj+1)=Popul(index(1),Mi*4*Nc+4*Nj+2);
```

```
        EC(Mi+1,Nj+1)=Popul(index(1),Mi*4*Nc+4*Nj+3);
ED(Mi+1,Nj+1)=Popul(index(1),Mi*4*Nc+4*Nj+4);
    end
  end

Chrom_Pattern_C=A_pat.*fftshift(fft2(EA))+B_pat.*fftshift(fft2(EB))+C_pat
.*fftshift(fft2(EC))+D_pat.*fftshift(fft2(ED));
Power_total=sum(sum((abs(Chrom_Pattern_C)).^2));
Pattern_Norm=Chrom_Pattern_C/sqrt(Power_total);
MSE_ga=sum(sum((abs((abs(Pattern_Norm(W1:(Mc)/2+1, W3:W4))).^2-
(abs(target(W1:(Mc)/2+1, W3:W4))).^2)).^2));
figure; imagesc((abs(Pattern_Norm)));colormap gray; axis image;

figure; plot(1:GA_k,MSE_eli_mean, 1:GA_k, MSE_eli_min);









%%
   A_pat_2=A_pat;
   B_pat_2=B_pat;
   C_pat_2=C_pat;
   D_pat_2=D_pat;

   Chrom_Pattern_C_1=Chrom_Pattern_C; % the total pattern;
   MSE_1=MSE_ga;
   Ran3=Ran1.*ones(Mc,Nc); Ran4=Ran2.*ones(Mc,Nc);
   Counter=0;
 for N_round=1:10;  % the round number of the scan;
 for Mi=0:Mc-1; % scan cell by cell;
    for Nj=0:Nc-1;



for Xr=3*P/8:P/8:5*P/8; % scan pixel by pixel in each cell;
    for Yr=3*P/8:P/8:5*P/8;
    Counter=Counter+1;


delta_Vector=[-1/(2*Nc), -1/(2*Mc)];

% the first triangle
tL_ab=[0, 1/Mc]; tL_bo=[(Yr/P)/Nc, (Xr/P-1)/Mc]; tL_oa=[(-Yr/P)/Nc, (-
Xr/P)/Mc];
nL_ab=[-1/Mc, 0]; nL_bo=[-(Xr/P-1)/Mc, (Yr/P)/Nc]; nL_oa=[-(-Xr/P)/Mc, (-
Yr/P)/Nc];
Mid_ab=[0, 1/Mc]./2+delta_Vector; Mid_bo=[(Yr/P)/Nc,
(1+Xr/P)/Mc]./2+delta_Vector; Mid_oa=[(Yr/P)/Nc,
(Xr/P)/Mc]./2+delta_Vector;

% the second triangle
```

```
tL_bc=[1/Nc, 0]; tL_co=[(Yr/P-1)/Nc, (Xr/P-1)/Mc]; tL_ob=[(-Yr/P)/Nc, (1-
Xr/P)/Mc];
nL_bc=[0, 1/Nc]; nL_co=[-(Xr/P-1)/Mc, (Yr/P-1)/Nc]; nL_ob=[-(1-Xr/P)/Mc,
(-Yr/P)/Nc];
Mid_bc=[1/Nc, 2/Mc]./2+delta_Vector; Mid_co=[(Yr/P+1)/Nc,
(Xr/P+1)/Mc]./2+delta_Vector; Mid_ob=Mid_bo;

% the third triangle
tL_cd=[0, -1/Mc]; tL_do=[(Yr/P-1)/Nc, (Xr/P)/Mc]; tL_oc=[(1-Yr/P)/Nc, (1-
Xr/P)/Mc];
nL_cd=[1/Mc, 0]; nL_do=[-(Xr/P)/Mc, (Yr/P-1)/Nc]; nL_oc=[-(1-Xr/P)/Mc,
(1-Yr/P)/Nc];
Mid_cd=[2/Nc, 1/Mc]./2+delta_Vector; Mid_do=[(1+Yr/P)/Nc,
(Xr/P)/Mc]./2+delta_Vector; Mid_oc=Mid_co;

% the forth triangle
tL_da=[-1/Nc, 0]; tL_ao=[(Yr/P)/Nc, (Xr/P)/Mc]; tL_od=[(1-Yr/P)/Nc, (-
Xr/P)/Mc];
nL_da=[0, -1/Nc]; nL_ao=[-(Xr/P)/Mc, (Yr/P)/Nc]; nL_od=[-(-Xr/P)/Mc, (1-
Yr/P)/Nc];
Mid_da=[1/Nc, 0]./2+delta_Vector; Mid_ao=Mid_oa; Mid_od=Mid_do;

tL(:,:,1)=[tL_ab; tL_bo; tL_oa];
tL(:,:,2)=[tL_bc; tL_co; tL_ob];
tL(:,:,3)=[tL_cd; tL_do; tL_oc];
tL(:,:,4)=[tL_da; tL_ao; tL_od];

nL(:,:,1)=[nL_ab; nL_bo; nL_oa];
nL(:,:,2)=[nL_bc; nL_co; nL_ob];
nL(:,:,3)=[nL_cd; nL_do; nL_oc];
nL(:,:,4)=[nL_da; nL_ao; nL_od];

Mid(:,:,1)=[Mid_ab; Mid_bo; Mid_oa];
Mid(:,:,2)=[Mid_bc; Mid_co; Mid_ob];
Mid(:,:,3)=[Mid_cd; Mid_do; Mid_oc];
Mid(:,:,4)=[Mid_da; Mid_ao; Mid_od];



   for S=1:4;

        MM_1=m*tL(1,1,S); NN_1=n*tL(1,2,S);
        DOT_1=MM_1'*ones(1,Nc*PP)+ones(Mc*PP,1)*NN_1;
        temp_1=DOT_1+(DOT_1==0);
        Sinc_1=sin(pi*DOT_1)./(pi*temp_1)+(temp_1==1);

        MM_2=m*tL(2,1,S); NN_2=n*tL(2,2,S);
        DOT_2=MM_2'*ones(1,Nc*PP)+ones(Mc*PP,1)*NN_2;
        temp_2=DOT_2+(DOT_2==0);
        Sinc_2=sin(pi*DOT_2)./(pi*temp_2)+(temp_2==1);

        MM_3=m*tL(3,1,S); NN_3=n*tL(3,2,S);
        DOT_3=MM_3'*ones(1,Nc*PP)+ones(Mc*PP,1)*NN_3;
        temp_3=DOT_3+(DOT_3==0);
        Sinc_3=sin(pi*DOT_3)./(pi*temp_3)+(temp_3==1);
```

```
        MM_4=m*nL(1,1,S); NN_4=n*nL(1,2,S);
        NormalD_1=MM_4'*ones(1,Nc*PP)+ones(Mc*PP,1)*NN_4;

        MM_5=m*nL(2,1,S); NN_5=n*nL(2,2,S);
        NormalD_2=MM_5'*ones(1,Nc*PP)+ones(Mc*PP,1)*NN_5;

        MM_6=m*nL(3,1,S); NN_6=n*nL(3,2,S);
        NormalD_3=MM_6'*ones(1,Nc*PP)+ones(Mc*PP,1)*NN_6;


        MM_7=m*Mid(1,1,S); NN_7=n*Mid(1,2,S);
        MidD_1=exp(-1i*(2*pi)*(MM_7'*ones(1,Nc*PP)+ones(Mc*PP,1)*NN_7));

        MM_8=m*Mid(2,1,S); NN_8=n*Mid(2,2,S);
        MidD_2=exp(-1i*(2*pi)*(MM_8'*ones(1,Nc*PP)+ones(Mc*PP,1)*NN_8));

        MM_9=m*Mid(3,1,S); NN_9=n*Mid(3,2,S);
        MidD_3=exp(-1i*(2*pi)*(MM_9'*ones(1,Nc*PP)+ones(Mc*PP,1)*NN_9));

        MMNN=2*pi*((m.^2)'*ones(1,Nc*PP)+ones(Mc*PP,1)*(n.^2));
        temp_4=MMNN+1i*(MMNN==0);
        Term=1i./temp_4;



Pattern_tri=Term.*(Sinc_1.*NormalD_1.*MidD_1+Sinc_2.*NormalD_2.*MidD_2+Si
nc_3.*NormalD_3.*MidD_3);

        Pattern_t(:,:,S)=Pattern_tri; % the pattern from the triangle #S;

   end
   A_pat_2=Pattern_t(:,:,1);
   B_pat_2=Pattern_t(:,:,2);
   C_pat_2=Pattern_t(:,:,3);
   D_pat_2=Pattern_t(:,:,4);

   A_fourier=exp(-
1i*2*pi*(m'*ones(1,Nc*PP)*Mi/Mc+ones(Mc*PP,1)*n*Nj/Nc));



Chrom_Pattern_oneC_2=A_fourier.*Popul(index(1),Mi*4*Nc+4*Nj+1).*A_pat_2+A
_fourier.*Popul(index(1),Mi*4*Nc+4*Nj+2).*B_pat_2+A_fourier.*Popul(index(
1),Mi*4*Nc+4*Nj+3).*C_pat_2+A_fourier.*Popul(index(1),Mi*4*Nc+4*Nj+4).*D_
pat_2; % the pattern from one cell (four triangles);
  % the changed cell


Chrom_Pattern_oneC_1=A_fourier.*Popul(index(1),Mi*4*Nc+4*Nj+1).*A_pat+A_f
ourier.*Popul(index(1),Mi*4*Nc+4*Nj+2).*B_pat+A_fourier.*Popul(index(1),M
```

```
i*4*Nc+4*Nj+3).*C_pat+A_fourier.*Popul(index(1),Mi*4*Nc+4*Nj+4).*D_pat; %
the pattern from one cell (four triangles);
  % the original cell

 Chrom_Pattern_C_2=Chrom_Pattern_C_1-
Chrom_Pattern_oneC_1+Chrom_Pattern_oneC_2;

  Power_total=sum(sum((abs(Chrom_Pattern_C_2)).^2));
  Pattern_Norm=Chrom_Pattern_C_2/sqrt(Power_total);
  MSE_2=sum(sum((abs((abs(Pattern_Norm(W1:(Mc)/2+1, W3:W4))).^2-
(abs(target(W1:(Mc)/2+1, W3:W4))).^2)).^2));

  if MSE_2<MSE_1;
      Ran3(Mi+1,Nj+1)=Xr; Ran4(Mi+1,Nj+1)=Yr; % position accepted;

      Chrom_Pattern_C_1=Chrom_Pattern_C_2;
      MSE_1=MSE_2;
  else
      Xr=Ran3(Mi+1,Nj+1); Yr=Ran4(Mi+1,Nj+1);

      Chrom_Pattern_C_2=Chrom_Pattern_C_1;
  end

  MSE_3(Counter)=MSE_1;
    end
end



    end
    fprintf('%d\n', Mi);
 end
 end
figure;plot(MSE_3);

Power_total=sum(sum((abs(Chrom_Pattern_C_1)).^2));
Pattern_Norm_1=Chrom_Pattern_C_1/sqrt(Power_total);

figure; imagesc((abs(Pattern_Norm_1)));colormap gray; axis image;
toc;
```

Appendix B: Thin metal superlens imaging in nano-lithography[1]

Abstract

Superlens imaging system in nano-lithography can be regarded as a cascade of two F-P cavities, i.e. a superlens cavity, and a dielectric cavity between superlens and introduced mask of high loss, and the transfer function of system are obtained by considering multiple reflections inside the two cavities. For the range of wavevector of interest, the typical high peak of transmission coefficient of superlens coincides with a local minimum of transmission coefficient of dielectric cavity. The peak of transfer function of system corresponds to the peak of transmission coefficient of dielectric cavity. Thin superlens imaging system in nano-lithography is analysed based on transfer function, which can be flattened by simply tuning transmission coefficient of dielectric cavity and superlens cavity. The results are further validated by Finite Element Method (FEM) simulations.

Introduction

The metal planar superlens, with a negative permittivity at optical frequency and a positive permeability, was proposed by Pendry in 2000 [1] as an alternative to negative index media (NIM) by Veselago in 1967 [2] to break the diffraction limit. The principle of superlens is to compensate the exponential decay of the evanescent field away from the object by amplifying the evanescent waves through surface plasmons (SP) resonances. Since then, many research efforts had been devoted into superlens imaging [3-15], including theoretical models, numerical simulations, experimental demonstrations and applications.

The silver superlens was demonstrated experimentally with the resolution of one-sixth of the illumination wavelength by Zhang et al in 2005 [3]. In their experiment, 35 nm thick silver slab was used because it gives the optimum transfer function. Thinner sliver slabs show

[1] A side project irrelevant to CGH at the beginning of my phd program

higher but narrow enhancement bands, and thicker slabs show smaller enhancements in reference to zero-order transmission [4]. In order to eliminate the sharp peak of transfer function, Sheng et al proposed to design the metallic superlens close to the cutoff condition of the long-range SP mode to balance the amplification by the SP resonance and the flatness of the transfer function [5, 6]. Moore et al suggested a performance window for superlens with total thickness range from 120 nm to 140 nm in order to get a flatter transfer function [7]. Those were reported by considering transfer functions of the imaging system with three layers structure - a superlens sandwiched by two semi-infinite dielectrics [4-7].

Normally, when dealing with the imaging system, a perfectly absorbing, thin screen with slit(s) has been widely adopted. The transmission of the idealized screen will be 1 in the slit or 0 otherwise. While in real nano-lithography applications, there exists a metallic object mask. The metallic mask itself exists mainly for constructing a perfect object function in the object plane just behind the mask for the imaging system. To obtain a good enough object function, which means as close as the ideal case, i.e. 1 in the slit and 0 otherwise, a thick mask with intrinsic high loss, such as Chrome (Cr), is usually applied. Due to the introduction of mask, a dielectric cavity is naturally formed between the mask and superlens. Blaikie et al approximately considered the neglected recursive reflections in the dielectric cavity between the superlens and the mask, and modified transfer function by an improved transfer-matrix model [8]. Sheng et al further take possible SPR by mask itself in some cases into account and give a more general model with its transfer function, which is optimized by genetic algorithm (GA) [9].

For the superlens imaging system in nano-lithography, we regarded it as two cascaded F-P cavities – a superlens cavity, and a dielectric cavity between superlens and introduced mask of high loss, and the transfer function of system are obtained by considering multiple reflections inside the two cavities [9, 10]. We studied the transfer function of system and the transmission coefficient of two cavities, then revealed some relations among them. It is found that the peak shown in transmission coefficient of superlens always corresponds to the local minimum of transmission coefficient of dielectric cavity. Moreover, the peak of transfer function of system coincides with transmission coefficient of dielectric cavity instead of that of suprelens. Based on these, we propose to simply tune the transmission coefficient of

superlens cavity and dielectric cavity, a very thin metal superlens imaging system shows well-balanced transfer function and thus produce improved image, which is validated by FEM simulations.

Metal superlens imaging system

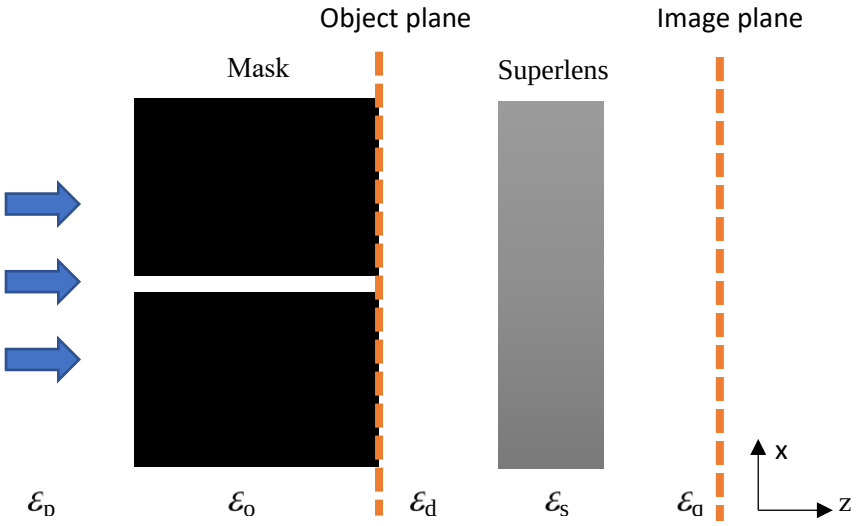


Fig. 1 Metal superlens imaging system for nano-lithography

Consider an imaging system for nano-lithography with a metal superlens of permittivity ε_s and a thick metallic mask of permittivity ε_o with high intrinsic loss to construct an object function, as shown in Fig. 1. For the purpose of the image analysis, herein a slit of nano-size is perforated through the object mask. A dielectric cavity, with permittivity ε_d , is formed between the superlens and the mask. The illuminating light of TM polarization is incident normally to the object mask from the semi-infinite dielectric medium ε_p . The image plane is placed in a semi-infinite dielectric medium ε_q at a distance d_q from the superlens. The incident plane wave of TM polarization was launched in dielectric medium ε_p, and then encounter the object mask ε_o. In the case where the lithographic metal mask layer is thick and of high intrinsic loss, the incident waves will be dissipated in the mask medium and can only pass through the slit. At the exit of the slit, the waves, labelled as source A, will be

scattered in the medium ε_d launching the Surface Plasmon Polaritons (SPPs) along interfaces of the superlens. Both the homogeneous propagating waves and inhomogeneous SPP waves will reach the image plane and contribute to the image in dielectric medium ε_q. The evanescent waves will simply undergo an exponential delay in the semi-infinite dielectric medium ε_q. Therefore, the imaging system for the source A shown in Fig. 1 can be regarded as two cascaded Fabry-Perot (F-P) cavities. The first cavity is the dielectric layer ε_d with two metal/dielectric interfaces, $\varepsilon_d / \varepsilon_o$ and $\varepsilon_d / \varepsilon_s$, respectively. The second cavity is the metal superlens layer with two interfaces, $\varepsilon_s / \varepsilon_d$ and $\varepsilon_s / \varepsilon_q$, respectively. The transfer function of the imaging system can be computed as the product of transmission coefficients of the two cavities, which are obtained by considering multiple reflections inside the two cavities. We analyze the mathematical expressions of the transmission coefficients corresponding to the two cavities in order to get a flat transfer function of the imaging system.

a. transmission coefficient of superlens cavity

The superlens cavity of permittivity ε_s is between two metal/dielectric interfaces of $\varepsilon_s / \varepsilon_d$ and $\varepsilon_s / \varepsilon_q$. Its transmission coefficient can be calculated by considering resonance of multiple reflected fields in the superlens cavity as [1, 5]

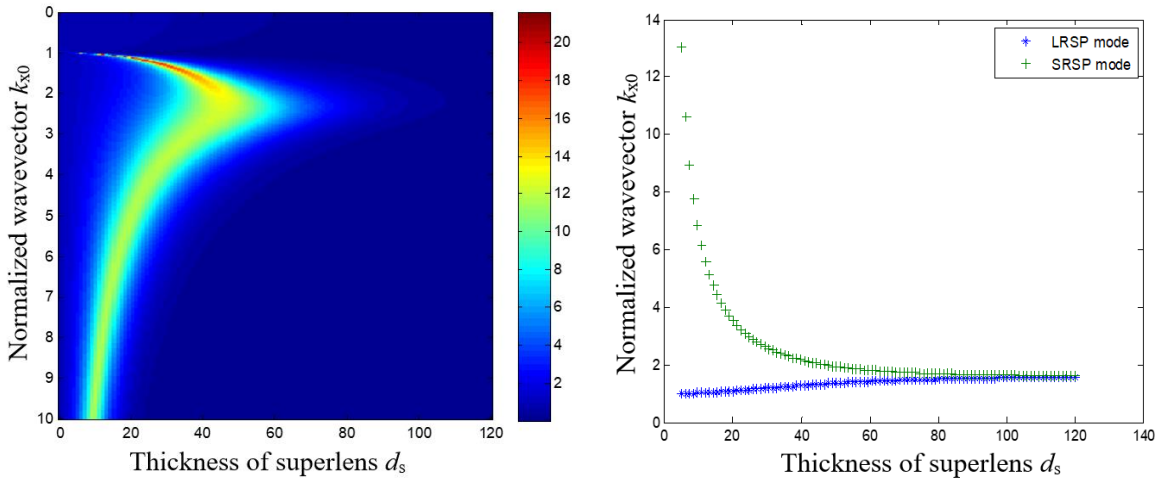
$$\tau_s = \frac{e_s t_{ds} t_{sq}}{1 - e_s^2 r_{sd} r_{sq}}, \tag{B. 1}$$

where the Fresnel reflection and transmission coefficients from medium *i* to medium *j*, with sub-indices *i, j* = *o, d, s* and *q*, are $r_{ij} = \left(\varepsilon_j k_{zi} - \varepsilon_i k_{zj}\right) \Big/ \left(\varepsilon_j k_{zi} + \varepsilon_i k_{zj}\right)$, $t_{ij} = 2\varepsilon_j k_{zi} \Big/ \left(\varepsilon_j k_{zi} + \varepsilon_i k_{zj}\right)$ with $k_{zi} = \sqrt{\varepsilon_i k_0^2 - k_x^2}$, the propagation factor $e_i = \exp\left(ik_{zi} d_i\right)$, describing the phase change of the propagating waves with $k_x^2 < \varepsilon_i k_0^2$ along the distance d_i and the exponent decay in amplitude of the evanescent waves $k_x^2 > \varepsilon_i k_0^2$ over d_i, respectively.

According to Eq. (B. 1) and for a silver superlens with variable thickness, the amplitude of transmission coefficient $|\tau_s|$ as a function of normalized wavevector $k_{x0} = k_x \Big/ \left(\sqrt{\varepsilon_d} k_0\right)$ is depicted in Fig. 2(a). Typically for a thin superlens of less than 30 nm thickness a sharp and high peak is shown at a low spatial frequency k_{x01} slightly larger than $k_{x0} = 1$ and a relative

broad and low peak is at a high spatial frequency k_{x02}. When increasing the thickness only one peak will appear at a spatial frequency between spatial frequency k_{x01} and k_{x02}. For large thickness of superlens, the amplitude of transmission coefficient become low and flat due to high loss in the thick metal superlens, as shown in Fig. 3(a).

According to the Maxwell's equations and the boundary conditions in the Dielectric – Metal – Dielectric (DMD) waveguide structure, the dispersion relation can be obtained and solved numerically [5]. The effective indices of waveguide modes are plotted for different thickness of Ag superlens in Fig. 2(b). Two modes, long-range surface plasmon (LRSP) mode and short-range surface plasmon (SRSP) mode, are supported for DMD waveguide of a thin superlens. The two peaks correspond to such two excited SP modes: a LRSP mode associated to a narrow peak located at lower spatial frequencies k_{x01} and a SRSP mode associated to a broad peak at higher spatial frequencies k_{x02}. When the thickness of the metal superlens increases, location of the peak at low frequency (LRSP mode) tends to shift to higher frequency, and location of the peak at high frequency (SRSP mode) tends to lower frequency. At a certain thickness of the superlens, only one mode will be supported which accounts for only one peak of the transmission coefficient.



(a) (b)

Fig. 2. For Ag superlens of thickness ranging from 5 nm to 120 nm sandwished by two dielectric mediums ε_d = ε_q =2.34, (a) The amplitude of transmission coefficient, $|\tau_s| < 1$ for propagating waves with $k_{x0} < 1$, while for evanescent waves with $k_{x0} > 1$, it can be amplified. (b) mode effective index of DMD waveguide.

If the imaging system is considered as a superlens between two semi-infinite dielectrics without presence of the metal lithographic mask, as that in ref. [4-7], then the very thin superlens is usually inappropriate for imaging due to the existence of sharp peak of $|\tau_s|$ resulting from excitation of LRSP mode. The LRSP mode can be cut off [5] if $r_{ds}+e_s^2 r_{sq}\approx 0$ is met, which means the reflection of metal layer is around 0.

b. transmission coefficient of dielectric cavity

The dielectric slab of ε_d is sandwiched by the metal superlens and the metal mask, constituting a dielectric F-P cavity. Its transmission coefficient can be calculated [9, 10] by taking into account the recursive reflections in the cavity as

$$\tau_d=\frac{e_d}{1-e_d^2\left(\frac{r_{ds}+e_s^2 r_{sq}}{1-e_s^2 r_{sd}r_{sq}}\right)\left(\frac{r_{do}+e_o^2 r_{op}}{1-e_o^2 r_{op}r_{od}}\right)}, \tag{B. 2}$$

For a thick mask with high loss used in the superlens image system for nano-lithography, $r_{op}=0$, then Eq. (B. 2) can be further written as

$$\tau_d=\frac{1}{\frac{1}{e_d}-e_d\left(\frac{r_{ds}+e_s^2 r_{sq}}{1-e_s^2 r_{sd}r_{sq}}\right)r_{do}}=\frac{1}{\frac{1}{e_d}-\frac{\left(r_{ds}+e_s^2 r_{sq}\right)r_{do}e_d}{e_s t_{ds}t_{sq}}\tau_s}. \tag{B. 3}$$

For evanescent waves with $k_x^2>\varepsilon_d k_0^2$, $e_d=\exp\left(i\sqrt{\varepsilon_d k_0^2-k_x^2}\,d_d\right)=\exp\left(-\sqrt{k_x^2-\varepsilon_d k_0^2}\,d_d\right)$ is real. The denominator in Eq. (B. 3) can be evaluated, based on $\big||a|-|b|\big|\le|a\pm b|\le|a|+|b|$ and $|ab|=|a||b|$, as

$$\left|\frac{\left(r_{ds}+e_s^2 r_{sq}\right)r_{do}e_d}{e_s t_{ds}t_{sq}}\right|\left|\tau_s\right|-\frac{1}{e_d}\le\left|\frac{1}{e_d}-\frac{\left(r_{ds}+e_s^2 r_{sq}\right)r_{do}e_d}{e_s t_{ds}t_{sq}}\tau_s\right|\le\left|\frac{\left(r_{ds}+e_s^2 r_{sq}\right)r_{do}e_d}{e_s t_{ds}t_{sq}}\right|\left|\tau_s\right|+\frac{1}{e_d}.$$

Thus, when the amplitude of transmission coefficient of the superlens $|\tau_s|$ shows a very high peak associated with the excitation of LRSP mode at a low spatial frequency k_{x01} slightly

larger than $k_{x0}=1$, the amplitude of transmission coefficient of the dielectric cavity $|\tau_d|$ will fall to a local minimum. While, $|\tau_d|$ will reach to a maximum when the absolute value of denominator of Eq. (B. 3) approaches to 0, i. e.

$$\left|\frac{1}{e_d}-\frac{\left(r_{ds}+e_s^2 r_{sq}\right)r_{do}e_d}{e_s t_{ds}t_{sq}}\tau_s\right|\to 0. \tag{B. 4}$$

Besides, it is interesting to notice that, if $r_{ds}+e_s^2 r_{sq}\approx 0$ in Eq. (B. 3), then the transmission coefficient $\tau_d \approx e_d$, implying an exponential decay of the evanescent waves in the dielectric cavity, just as that in the free space without any resonances, and is independent of the cavity thickness.

c. transfer function of imaging system

As the imaging system for the source A is a cascade of two F-P cavities, the total transfer function τ_A is calculated by

$$\tau_A=\tau_d\cdot\tau_s=\frac{1}{\dfrac{1}{e_d\tau_s}-\dfrac{\left(r_{ds}+e_s^2 r_{sq}\right)r_{do}e_d}{e_s t_{ds}t_{sq}}}, \tag{B. 5}$$

When the condition (B. 4) is satisfied, $|\tau_A|$ and $|\tau_d|$ will both reach to their maximums, according Eq. (B. 3) and Eq. (B. 5), at the same spatial frequency k_{x03}. This reveals that the peak of the total transfer function $|\tau_A|$ for source A in the imaging system corresponds to the peak of transmission coefficient of dielectric cavity $|\tau_d|$ at k_{x03} instead of that of superlens $|\tau_s|$ at k_{x01}. Actually, the value of $|\tau_A|$ at k_{x01} is always balanced as the transmission coefficient of the superlens $|\tau_s|$ can be supressed by the local minimum of $|\tau_d|$ at k_{x01}. The transfer function for source A is defined as the transmission coefficients of the system described in Eq. (B. 5) as a function of the spatial spectral frequency. According to SPP waveguide theory, the LRSP mode is cutoff when the propagation constant β is purely real,

which occurs when the nature of the mode changes from attenuating ($\mathrm{Im}[\beta] > 0$) to growing ($\mathrm{Im}[\beta] < 0$) with the propagation along the metal layer [5,6]. If the cut-off condition for LRSP mode, i.e. $r_{ds} + e_s^2 r_{sq} \approx 0$, is fulfilled, the transfer function $\tau_A = \tau_d \cdot \tau_s \approx e_d \cdot \tau_s$. In this case, a relative flat transfer function can be expected when consider together the well-controlled, proper amplification of SPP by superlens and an exponential decay of amplitude of evanescent waves in dielectric cavity.

Imaging by thin superlens imaging system

We first consider a Ag superlens imaging system with typical parameters: the metallic mask is Chrome with permittivity $\varepsilon_o = -8.55 + i8.96$ and thickness of d_o = 50nm, the silver slab with permittivity $\varepsilon_s = -2.6 + i0.25$ with incident wavelength 365 nm and thickness d_s = 30 nm, the dielectric layer between superlens and mask has the permittivity ε_d = 2.34 and thickness of d_d = 40 nm, the dielectric ε_q = 2.34, as shown in Fig. 1. From Fig 3 (a), the sharp peak of $|\tau_s|$ corresponds the local minimum of $|\tau_d|$ at $k_{x01} \approx 1.3$, which is the effective index of LRSP mode excited in DMD waveguide. The broad peak of $|\tau_s|$ at $k_{x02} \approx 3.6$ is shown because of the excitation of SRSP mode. The value of transfer function $|\tau_A|$ at $k_{x01} \approx 1.3$ is balanced by the multiplication of $|\tau_s|$ and $|\tau_d|$, which indicates that we cannot judge the imaging system only by transmission coefficient of superlens. The peak of transfer function $|\tau_A|$ locates at $k_{x03} \approx 1.6$, which is the same as that of transmission coefficient of dielectric cavity $|\tau_d|$. We change the thickness of Ag to d_s = 15 nm and thickness of dielectric d_d =10 nm, which form a thinner sueprlens imaging system. The transmission coefficient $|\tau_s|$ of 15 nm thick superlens has a shaper and higher peak, as shown in Fig. 3 (b), than that of superlens of 30 nm thickness in Fig. 3 (a). Still, the peak of $|\tau_s|$ corresponds the local minimum of $|\tau_d|$ at $k_{x01} \approx 1.1$, and thus the value of $|\tau_A|$ at $k_{x01} \approx 1.1$ is balanced in the same

manner. The peak of transfer function $|\tau_A|$ and transmission coefficient of the dielectric cavity $|\tau_d|$ is obtained around $k_{x03} \approx 1.4$, as depicted in Fig. 3 (b). The very thin superlens imaging system can show a flatter transfer function over a broader range of spatial frequency by simply tuning the transmission coefficient of superlens cavity and dielectric cavity.

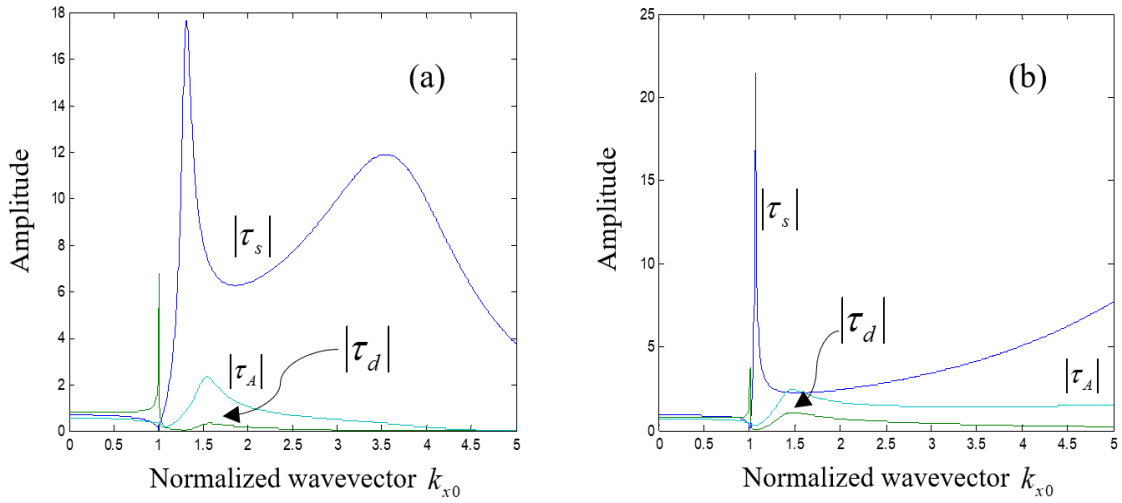


Fig. 3. The Amplitude of transmission efficient of superlens, dielectric and transfer function of the system as a function of normalized wavevector. (a) d_d =40nm, d_s =30 nm, (b) d_d =10nm, d_s =15nm.

For further validations, 2D Finite Element Method (FEM) by commercial software package COMSOL was employed to simulate the imaging system in (x, y) plane, as shown in Fig. 4. A TM (E_x, E_y, H_z) plane wave with wavelength of 365 nm was launched along y direction from the input port on the top boundary in the FEM simulations. A perfectly matched layer (PML) was applied at the bottom boundary and Periodical boundary conditions (PBC) were imposed on the left and right sides. A chrome mask with two slits of 40 nm was used. The centre-to-centre distance of two slits was 120 nm. The lateral length along x direction of each layer was 1 um. To clearly show the fields after the mask of high loss, the incident waves before and in the mask, and the mask itself are removed in Fig. 4 (a) and (c). The image of the two spaced slits is recorded at the imaging plane placed at a distance of 10 nm away the superlens, as shown in Fig. 4 (b) and (d). Obviously, the image in Fig. 4 (d) by thin superlens system with transfer function in Fig. 3(b) is better than the one in Fig. 4 (b) with the transfer function shown in Fig. 3 (a), because the sidelobes of the image in Fig.

4 (d) was largely suppressed compared with that in Fig. 4 (b). However, the amplitude of the unwanted centre peak is still high, as shown in Fig. 4 (b) and Fig. 4 (d).

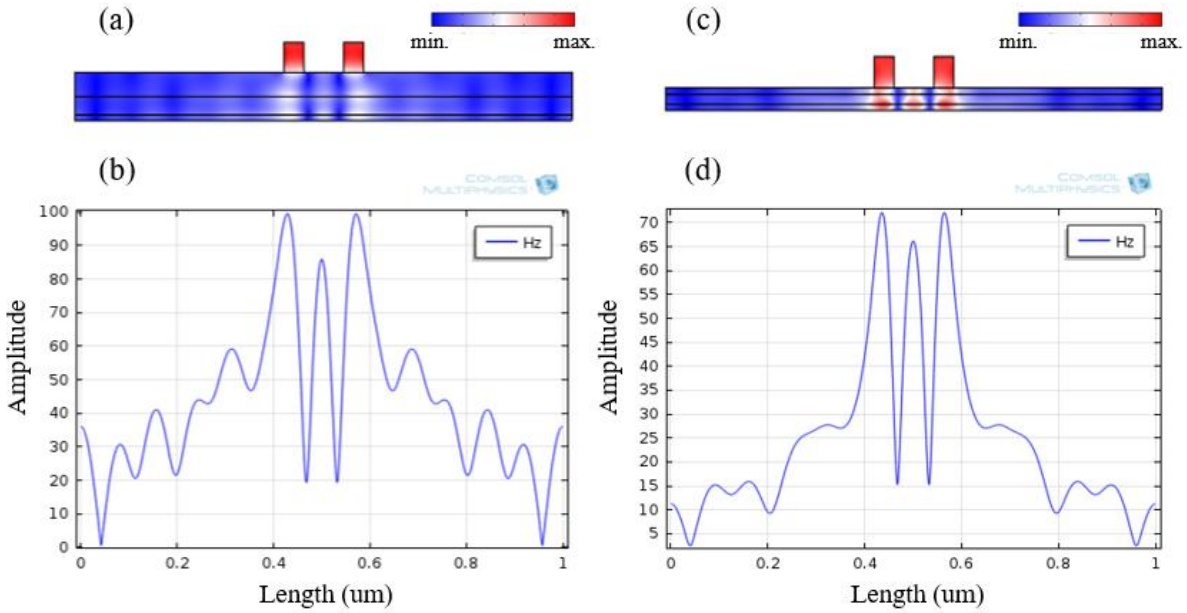


Fig. 4. Simulation results of the superlens imaging system by FEM. (a) Field distributions of system with d_d =40nm, d_s =30 nm; (b) image of two-slit object in the imaging plane in (a); (c) Field distributions of system with d_d =10nm, d_s =15 nm; (d) image of two-slit object in the imaging plane in (c).

The imaging performance of thin superlens imaging system can be further improved by approaching the LRSP mode cut-off condition as we demonstrated. Change the dielectric ε_d = 2.34 to ε_d = 4.08, and keep all others unchanged as those in Fig. 4(c). The transfer function $|\tau_A|$ and transmission coefficient $|\tau_d|$ and $|\tau_s|$ are plotted in Fig. 5 (a), from which, we can see that the high peak of $|\tau_s|$ associated with LRSP mode is mainly removed, and $|\tau_d|$ approximately follows an exponential decay for evanescent waves, and thus the transfer function $|\tau_A|$ is balanced to show flatness. The field by a FEM simulation in this superlens system is shown in Fig. 5 (b) and the image is notably improved as shown in Fig. 5 (c), in which, both the sidelobes and the centre peak are greatly suppressed compared with those in Fig. 4 (b) and Fig. 4 (d).

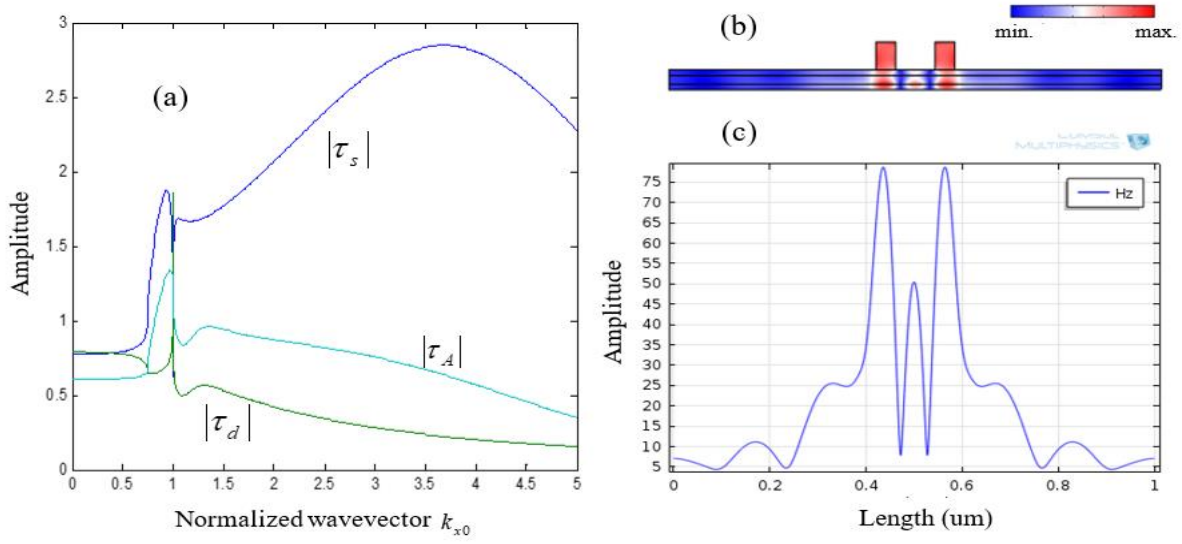


Fig. 5. (a) Amplitude of transmission coefficient of superlens, dielectric and transfer function of the system as a function of normalized wavevector. (a) LRSP mode cut-off condition is approached with d_d =10nm, d_s =15nm, ε_d =4.08, (b) field distributions in superlens imaging system, (c) image of two-slit object in the imaging plane in (b).

Conclusions

We investigated the imaging performance of the metal planar superlens for Nano-lithography, considering the superlens imaging system as a cascade of two F-P cavity: a superlens cavity, and a dielectric cavity between superlens and introduced mask of high loss. By analysing of the transmission coefficient of superlens and dielectric cavity, and the transfer function of whole system, we found that the peak of the transmission coefficient of the superlens always coincides with the local minimum of that of the dielectric cavity, and the peak of the whole transfer function of system actually corresponds to that of dielectric cavity instead of superlens. Then, we show that very thin superlens imaging system, which was usually believed improper for superlens imaging, has an improved imaging performance by simply tuning the transmission coefficient of superlens cavity and dielectric cavity. All the results are analyzed based on transfer functions and further confirmed by FEM simulations.

References

[1] J. B. Pendry, “Negative refraction makes a perfect lens,” Phys. Rev. Lett. 85, 3966-3969 (2000).

[2] V. G. Veselago, “The electromagnetics of substances with simultaneously negative values of ε and μ”, Soviet Phy. Uspekhi, vol. 10 No. 4, 509-514 (1968)

[3] N. Fang, H. Lee, C. Sun, et al., Sub-diffraction-limited optical imaging with a silver superlens, Science 308(5721) (2005), 534–537

[4] Hyesog Lee, Yi Xiong, Nicholas Fang et al, Realization of optical superlens imaging below the diffraction limit, 2005 New J. Phys. 7 255

[5] G. Tremblay and Y. Sheng, ‘Improving imaging performance of a metallic superlens using the long-range surface Plasmon polariton mode cutoff technique’, Applied Optics, Vol. 49, No. 7 (2010)

[6] G. Tremblay and Y. Sheng, ‘Designing the metallic superlens close to the cutoff of the long-range mode’, Optics Express, Vol. 18, No. 2, 740-745 (2010)

[7] Ciaran P. Moore and Richard J. Blaikie, "Robust design of a silver-dielectric near-field superlens for photolithography," J. Opt. Soc. Am. B 30, 3272-3277 (2013)

[8] Ciaran P. Moore, Richard J. Blaikie, and Matthew D. Arnold, "An improved transfer-matrix model for optical superlenses," Opt. Express 17, 14260-14269 (2009)

[9] G. Tremblay and Y. Sheng, ‘Modeling and designing metallic superlens with metallic objects’, Optics Express, Vol. 19, No. 21, 20634-20641 (2011)

[10] Fuyang Xu, Genhua Chen, Chinhua Wang et al, "Superlens imaging with a surface plasmon polariton cavity in imaging space," Opt. Lett. 38, 3819-3822 (2013)

[11] Beibei Zhang and Jacob B. Khurgin, Eigen mode approach to the sub-wavelength imaging with surface plasmon polaritons, Appl. Phys. Lett. 98, 263102 (2011)

[12] X. Zhang and Z. Liu, “Superlenses to overcome the diffraction limit,” Nature Materials volume 7, pages 435–441 (2008)

[13] S. Kawata, Y. Inouye, P. Verma, Plasmonics for near-field nano-imaging and superlensing. Nat. Photon., 3:388—394, 2009

[14] Changtao Wang, Wei Zhang, Zeyu Zhao et al, Plasmonic Structures, Materials and Lenses for Optical Lithography beyond the Diffraction Limit: A Review, Micromachines. 2016 Jul; 7(7): 118.

[15] Katherine A. Willets, Andrew J. Wilson, Vignesh Sundaresan et al, Super-Resolution Imaging and Plasmonics, Chem. Rev., 2017, 117 (11), pp 7538–7582

Appendix C: List of publications

The papers published in peer-reviewed scientific journals and international conferences are listed as below:

Jing Wang, Yunlong Sheng, “Design quadrilateral apertures in binary computer-generated holograms of large space bandwidth product,” Applied Optics Vol. 55, Issue 27, pp. 7636-7644 (2016)

Jing Wang, Yunlong Sheng, “Computer-Generated Very Large Space-Bandwidth Product Binary Hologram for Laser Projector,” IEEE Transactions on Industrial Informatics Vol. 12, Issue 1, pp. 179-186 (2016)

Jing Wang, Yunlong Sheng, "Thin metal superlens imaging in nanolithography," International Journal of Optics, Vol. 2019, Article ID 6513836, 6 pages, (2019).

Jing Wang, Yunlong Sheng, "Design of computer-generated hologram apertures with the Abbe transform", in Holography, Diffractive Optics, and Applications VII, Proceedings of SPIE Vol. 10022, 100221F. (2016)

Jing Wang, Yunlong Sheng, “Direct Design of Quadrilateral Apertures in Binary CGH by Parallel Genetic Algorithm and Direct Search,” Digital Holography & 3-D Imaging Meeting, Imaging and Applied Optics 2016, OSA Technical Digest (2016), paper DW5I.1.

Jing Wang, Yunlong Sheng, “Computer generated very large space bandwidth product binary hologram for laser projector”, INDIN 2015 IEEE International Conference on Industrial Informatics, pp. 691-695 (2015)

Jing Wang, Yunlong Sheng, “Binary hologram of very large space bandwidth product designed by the Genetic Algorithm,” Digital Holography & 3-D Imaging Meeting, OSA Technical Digest (2015), paper DM4A.3.